Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy

Chapter 10 Analysing and Presenting Results

Petra Macaskill, Constantine Gatsonis, Jonathan Deeks,

Roger Harbord, Yemisi Takwoingi.

Version 1.0 Released December 23rd 2010.

Please cite this version as: Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis C (editors), *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* Version 1.0. The Cochrane Collaboration, 2010. Available from: http://srdta.cochrane.org/.

Saved date and time 23/12/2010 15:08 Jon Deeks

Contents

10 Analysing and Presenting Results

10.1 Introduction

The statistical aspects of a systematic review of diagnostic test accuracy are more challenging than for reviews of interventions, and it is recommended that review teams include an individual with the statistical expertise needed to understand and implement the hierarchical models required for meta-analysis. This chapter has been written with this recommendation in mind. It first aims to both provide guidance to the key researchers in the review team on the purpose, possibilities and interpretation of methods of meta-analysis, and secondly provides the technical detail to assist a statistical expert in applying the methods recommended for Cochrane Reviews. Sections 10.1 to 10.4 and 10.6 outline the conceptual approach to meta-analysis, how analysis is undertaken for a single test accuracy study, and the graphical presentations and meta-analysis methods that are recommended. Section 10.5 is the more technical guide to the meta-analytical models to assist an informed statistician apply them in commercial statistical software programs, and is necessarily written presuming a level of familiarity with statistical hierarchical modelling. It includes examples with data sets, computer code and resulting output. Section 10.5 is therefore unlikely to be understood by all readers.

10.1.1 Aims of meta-analysis for DTA reviews

Health professionals (mainly physicians) use diagnostic tests to ascertain whether an individual (usually a patient) does or does not have a particular disease or condition. Cochrane diagnostic test accuracy reviews provide information on how well tests distinguish patients with the disease from those without. Most tests are imperfect, and errors will occur. Hence, the statistical methods focus on two statistical measures of diagnostic accuracy, the sensitivity of the test (the proportion of those with the disease who have an abnormal test result) and the specificity of the test (the proportion of those without the disease who have a normal test result). A Cochrane DTA review aims to quantify and compare these statistics for one or more diagnostic tests to describe how well each test classifies individuals, and estimate and compare the likely error rates (false positive and false negative diagnoses) that may be encountered. Publishing such reviews in the Cochrane Library aims to assist decision makers in rationally choosing and using tests by providing good evidence about their likely error rates.

Meta-analysis is a set of statistical techniques for combining results from two or more separate studies. Meta-analysis of diagnostic test accuracy studies provides summaries of the results of relevant included studies: providing an estimate of the average diagnostic accuracy of a test or tests, the uncertainty of this average, and the variability of study findings around the estimates. Meta-analytical regression models can statistically compare the accuracy of two or more different diagnostic tests and describe how test accuracy varies with test thresholds and other study characteristics.

Meta-analysis helps to make sense of apparently conflicting study results, as it identifies which differences are likely to be real, which are explicable by chance, and which can be explained by known differences in study characteristics. As the precision of estimates typically increases with the quantity of data, meta-analysis may have more power to detect real differences in test accuracy between tests than single studies, and may yield more precise estimates of expected sensitivity and

specificity. Also, by quantifying the variability of test accuracy across many settings, meta-analysis may provide insights into the consistency of test results. Meta-analysis models also provide a framework for comparing the accuracy of tests which have not directly been compared in individual studies.

10.1.2 When not to use a meta-analysis in a review

Meta-analysis is a powerful tool to use to summarise study findings, providing the estimates of test accuracy in the individual studies are both relevant and unlikely to be biased.

A common criticism of meta-analyses of studies of interventions is that 'they combine apples with oranges', implying that they may mix together estimates from studies which differ in important ways. This is one of the reasons why Cochrane reviews emphasise the importance of carefully defining inclusion criteria to identify studies which directly address the review question. In any analysis it is important to ensure that there are no differences between the studies in terms of the participants they recruit and the tests which they evaluate which would make the results of the meta-analysis uninformative. This is particularly important in reviews of test accuracy, as changes to patient selection criteria will alter the spectrum of disease and non-disease in the population, which can strongly impact on test accuracy as discussed in Chapter 9.

In addition it is important that the studies that are being combined in an analysis are methodologically rigorous. Meta-analysis of studies at risk of bias may be seriously misleading. If bias is present in individual studies meta-analysis may compound the errors and produce an erroneous result which may be inappropriately interpreted as having credibility. Meta-analysis involving regression modelling (see 10.5.3) may be useful to investigate how poor methodological quality can lead to bias in results.

10.1.3 How does meta-analysis of diagnostic test accuracy differ from meta-analysis of interventions?

The format of Cochrane DTA reviews allows for greater flexibility for structuring and reporting meta-analysis than is available in Cochrane Intervention reviews, and requires use of external statistical software. These differences arise for five main reasons:

1) Diagnostic test accuracy reviews can have diverse aims and address different types of question (as outlined in 10.1.4 below). Different comparisons and multiple aims may be addressed in a single review, often using data from the same studies in several analyses. To provide the flexibility needed RevMan requires separate steps of organising data entry and specifying analyses, unlike in Cochrane reviews of interventions where the two stages are combined. Thus there is a need to develop both an appropriate data structure and a clear analysis plan.
2) Evaluating test accuracy requires knowledge of two quantities, the test sensitivity and specificity. Meta-analysis methods for diagnostic test accuracy thus have to deal with two summary statistics simultaneously rather than one (as is the case for reviews of interventions).
3) A meta-analysis of diagnostic test accuracy has to allow for the trade-off between sensitivity and specificity that occurs between studies that vary in the threshold value used to define test positives and test negatives (see 10.2.4). Meta-analysis methods have been devised to enable studies to be combined that have used a test(s) at different thresholds, a common occurrence in many diagnostic test systematic reviews.

4) Heterogeneity is to be expected in results of test accuracy studies, thus random effects models are required to describe the variability in test accuracy across studies (see 10.4.3).
5) Methods for undertaking analyses which account for both sensitivity and specificity, the relationship between them, and the heterogeneity in test accuracy, require fitting hierarchical random effects models, which is beyond the analytical abilities of RevMan. Although exploratory analyses can be undertaken in RevMan, the definitive analyses needs to be undertaken in commercial software packages and sophisticated statistical programming environments such as SAS, Stata, S-Plus, R, MLwiN or winBUGS/OpenBUGS, for which collaboration with a statistical expert is highly recommended .

10.1.4 Questions which can be addressed in DTA analyses

There are three main types of question that can be addressed in a Cochrane DTA analysis concerning the accuracy of a test. The question types are mirrored as different options in the DTA module in RevMan when creating analysis definitions.

10.1.4.1 What is the accuracy of a test?

Such an analysis is restricted to characterising the accuracy of a single test, and aims either to estimate an average summary value of sensitivity and specificity or to describe how sensitivity and specificity vary with changing threshold by estimating a summary ROC curve. Which approach is used will depend on the nature of the test, and the variability in thresholds across the studies, which is discussed in more detail in 10.4.1.

10.1.4.2 How does the accuracy vary with clinical and methodological characteristics?

Planned investigations of heterogeneity investigate whether the observed test accuracy varies between studies according to characteristics associated with the tests, settings, participants or methodology of the studies. For purposes of graphical presentation it is best for the characteristic variable to group studies in categories. However, meta-regression models allow investigation of the relationship of accuracy to both categorical and continuous covariates, such as disease prevalence or test threshold. Both differences in key parameters of summary ROC curves and in summary sensitivity-specificity points can be investigated.

10.1.4.3 How does the accuracy of two or more tests compare?

Comparison of the accuracy of tests is an important part of a Cochrane DTA review, as it identifies which test (or tests) yields superior test accuracy. It is possible to compare multiple tests in a single analysis – there is no general restriction to comparing only pairs of tests, although it is often helpful to structure comparisons of multiple tests as a series of pairwise comparisons (bearing in mind problems caused by making excessive numbers of multiple comparisons). Methodologically, comparing two tests can be considered as a form of subgroup analysis, with studies evaluating each test each in a separate subgroup, so the same statistical modelling techniques are used as for investigating sources of heterogeneity. However, there is an important consideration to be made about the studies to be included in each pairwise comparison of two tests; whether all studies should be included, or whether the comparison should be restricted to only those which make direct comparisons themselves, either by testing all patients using all tests or by randomizing patients to different tests.

10.1.5 Planning the analysis

Undertaking meta-analyses for a Cochrane DTA reviews involves first developing an analysis plan and creating a series of analysis definitions in RevMan. Some of these decisions can be made at protocol stage (see 10.1.6), others only after the data has been extracted from the papers. The planning stages can be organised as follows:

- Clearly specifying the main questions which need answering, concerning which tests require estimates of test accuracy, and which tests should be compared with each other.
- Detailed planning of the way in which comparisons will be made, identifying the different tests or groups of tests which can be compared, the multiple and pairwise comparisons that will be made, and the studies and data that will be included in each analysis. A decision to consider here is whether comparative analyses should include all studies, or be restricted to those studies that evaluate both tests. Covariates for any heterogeneity analyses similarly need to be specified and coded.
- From these a list of the planned main analyses, test comparisons and heterogeneity analyses will be produced. The quantity of data that are available for each analysis should be determined to guide the choice of analysis method, and to assess whether adequate data are available for planned heterogeneity analyses. An analysis definition can be created in RevMan for each, and outlines of major results tables created.
- Plotting the results on forest plots and ROC plots using the functions in RevMan will familiarise the review author with the location and variability of the study results.
- A strategy needs to be specified to deal with the mixed reporting of thresholds that may occur across studies. A key issue is deciding whether an analysis should be restricted to studies that share a common threshold value (which allows estimation of the summary sensitivity and specificity of a test at that threshold) or to include all studies regardless of threshold value (which allows estimation of summary ROC curves but compromises the interpretation of sensitivity and specificity points). This will be informed by information about the thresholds at which the tests were evaluated in the primary studies, and knowledge of how the tests are applied in clinical practice.
- Once this analysis plan has been determined data must be exported from RevMan to the chosen statistics package, and appropriate models fitted. Results must be collated and tabulated as required, and parameter estimates for average sensitivity and specificity points and summary ROC curves copied back into the RevMan graphics to produce final graphical output.

10.1.6 Writing the analysis section of the protocol

As the analysis will to some extent depend on the type and quantity of data that are located through the literature search, it is often not possible to fully specify the analysis at the protocol stage. However, certain aspects must be predefined, and analysis strategies included where full details cannot be provided. Developing a protocol prior to reviewing the studies adds scientific credibility to the review, aiming to reduce the possibility that decisions made during the analysis are not data driven, in that analytical options are not selected in order to manipulate the findings. It also ensures that there is a clear plan for the collection and processing of data, which will inform the data extraction process and ensure that the analyses done will address the aims of the review.

A level of familiarity with key statistical summary measures should be presumed when writing a protocol. For example, it is not necessary to define summary measures such as sensitivity and

specificity, likelihood ratios, etc. Similarly, it is not necessary to include explanations of the meta-analytical methods used if they are those described in the Handbook. This chapter of the Handbook should be cited if any definitions and explanations are thought necessary. Where non-standard methods are required, these should be described and their use justified.

Key issues which need to be stated are:

- Definitions of key criteria, such as disease (specifying any binary classifications required) and categorisations of positive and negative test results. Where there are several possible options a strategy needs to be provided as to how a definition will be made, and plans for sensitivity analyses (10.6.1) included in order to investigate the robustness of the decisions made. Rules for handling known categories of indeterminate test results should be pre-stated where possible.

- A strategy needs to be included for handling multiple thresholds for test positivity, pre-specifying, if possible, any common thresholds at which summary estimates of sensitivity and specificity will be obtained (see 10.4.1).
- Approaches to modelling need to be outlined. In some cases, it may not be possible to specify in advance whether the modelling will focus on summary points and/or curves as this will be determined by how studies report their results. In this situation, reviewers should make it clear how they will make this decision once the data are available (see 10.4.1). The software that will be used for analysis should be stated (see 10.5.5).
- It needs to be stated clearly whether all studies will be included in test comparisons, whether comparisons will be based on paired data only, or whether both will be presented. If both, it needs to be clear which will be the primary analysis. Again, numbers of studies may affect the original intent (see 10.5.4**Error! Reference source not found.Error! Reference source not found.**).
- Planned investigations of heterogeneity should be outlined, stating covariate codings if known, and the approaches used for building models (see 10.5.3).
- Plans, if any, for investigating reporting biases should be outlined (see 10.6.3).

Any deviations from the protocol should be documented in the 'Differences between protocol and review' section at the end of the review.

10.2 Key concepts

10.2.1 Disease status

For the purposes of this Handbook, the accuracy of a diagnostic or screening test will be assessed by measures of the test's ability to detect the presence of disease. The true disease status of each individual will be considered as *binary (dichotomous),* diseased and not diseased. Although this represents a simplification of the reality of diagnosis, the vast majority of available methodology for the assessment of diagnostic and screening tests is predicated on the assumption of a dichotomous true disease status. Where there are alternatives for dichotomisation of disease status, binary categorisations which relate to decision-making options used in clinical practice should be chosen to ensure that the review will inform clinical decision-making. Where no consensus exists, consideration of alternative categorisations may be investigated in sensitivity analyses (10.6.1).

Statistical methodology is currently being developed for modelling test accuracy for multiple disease categories, but this is currently at a developmental stage and not ready for inclusion in Cochrane reviews of diagnostic test accuracy (see 10.6.4).

10.2.2 Types of test data

Systematic reviews of diagnostic and screening test accuracy involve test results of one or more of the following three data types:

- *Binary (dichotomous),* in which the test result is reported as a yes or no, positive or negative.
- *Ordinal*, in which the test result is reported on a set of ordered categories, often with verbal descriptors, such as 1=definitely normal, 2=presumably normal, 3=equivocal, 4=presumably abnormal, 5= definitely abnormal.
- *Continuous or Count*, in which the test result is reported on a continuous scale or as a count, such as the concentration of a substance or the number of features observed.

Many ordinal and binary categorizations arise, or can be conceptualized as arising from underlying continuous measurements by application of one or more thresholds. For example, laboratory tests that report results as positive or negative typically involve a numerical measurement which is categorized according to a pre-stated threshold, whereas imaging tests may report an ordinal grade for the certainty of the presence of a feature or the stage of disease progression.

To be included in a meta-analysis, ordinal, count or continuous test results need be re-categorized as binary by selecting a threshold and presenting the data as a 2x2 table. The issue of choice of such *positivity thresholds* and examination of accuracy at several thresholds is discussed in 10.2.4 and 10.4.1.

10.2.3 Analysis of a primary test accuracy study

This section defines summary statistics for test accuracy commonly used in reports of primary studies.

Having chosen a particular threshold for test positivity, the data from a primary study can be presented in a 2x2 table showing the cross classification of disease status (result of the reference standard) and test outcome (result of the index test) as in Table 10.1. For simplicity, throughout this chapter we refer to those with and without the target condition as defined by the reference standard as *diseased* and *non-diseased*, accepting that those without the target condition may well have other diseases.

Table 10.1 2x2 cross classification of test results and disease status

Test outcome (index test)	*Disease status (reference standard result)*		
	Diseased (D+)	*Non-diseased (D-)*	*Total*
Index test positive (T+)	True positives (a)	False positives (b)	Test positives (a+b)
Index test negative (T-)	False negatives (c)	True negatives (d)	Test negatives (c+d)
Total	Disease positives (a+c)	Disease negatives (b+d)	N (a+b+c+d)

Study specific as well as summary measures of test accuracy are then computed either as proportions of those disease positive or negative (in statistical terms, these are statistics that are *conditional* on the disease status) or test positive or negative (these are statistics that are *conditional* on the index test result) as described below.

10.2.3.1 Sensitivity and Specificity

Sensitivity and specificity are measures defined conditional on the disease status as they are computed as proportions of the number diseased and the number non-diseased respectively.

The sensitivity of a test is defined as the probability that the index test result will be positive in a diseased case. Formally, sensitivity=P(T+|D+) and is estimated using the numbers from the table as a/(a+c). Sensitivity is sometimes referred to as Detection Rate (DR), True Positive Rate (TPR) or True Positive Fraction (TPF). It is expressed either as a proportion or a percentage.

The specificity of a test is defined as the probability that the index test result will be negative in a non-diseased case. Formally, specificity=P(T-|D-) and is estimated using the numbers from the table as d/(b+d). Specificity is occasionally referred to as the True Negative Rate (TNR) or True Negative Fraction (TNF). More often, the terms False Positive Rate (FPR) and False Positive Fraction (FPF) are used for the complement of specificity (computed as 1−specificity or b/(b+d)). Again, both proportions and percentages are used.

Although the terms true positive fraction and false positive fraction are both technically more correct because sensitivity and specificity are fractions and not rates, true positive rate and false positive rate are the terms in most common usage and will be used in this Handbook.

The values of sensitivity and specificity are occasionally combined in a measure known as Youden's Index computed as sensitivity+specificity–1. Youden's Index has no direct probabilistic interpretation but provides a general index of test accuracy which gives equal weight to test errors (false negatives and false positives). Values close to 1 indicate high accuracy; a value of zero is equivalent to uninformed guessing and indicates that a test has no diagnostic value.

10.2.3.2 Predictive values

Predictive values are measures defined conditional on the index test results as they are computed as proportions of the total with positive and negative index test results.

The positive predictive value of a test is defined as the probability that a case with a positive index test result is diseased. Formally, positive predictive value=P(D+|T+) and is estimated using the numbers from the table as a/(a+b). Again, positive predictive values are reported either as proportions or percentages.

The negative predictive value of a test is defined as the probability that a case with a negative index test result is non-diseased. Formally, negative predictive value=P(D-/T-) and is estimated using the numbers from the table as d/(c+d). Again, negative predictive values are reported either as proportions or percentages.

10.2.3.3 Likelihood ratios

Likelihood ratios can be used to update the pre-test probability of disease using Bayes' theorem, once the test result is known. The updated probability is referred to as the post-test probability. For a test that is informative, the post-test probability should be higher than the pre-test probability if the test result is positive, whereas the post-test probability should be lower than the pre-test probability if the test result is negative. Considerations about the use of likelihood ratios in systematic reviews of test accuracy are explained in the Chapter 11.

The positive likelihood ratio describes how many times more likely positive index test results were in the diseased group compared to the non-diseased group. The positive likelihood ratio, which should be greater than 1 if the test informative, is defined as:

$$LR+ = P(T+|D+)/P(T+|D-) = sens/(1-spec),$$ and is estimated as $(a/(a+c)) / (b/(b+d)).$

The negative likelihood ratio describes how many times less likely negative index test results were in the diseased group compared to the non-diseased group. The negative likelihood ratio, which should be less than 1 if the test is informative, is defined as:

$$LR- = P(T-|D+)/P(T-|D-) = (1-sens)/spec,$$ and is estimated as $(c/(a+c)) / (d/(b+d)).$

10.2.3.4 Diagnostic odds ratios

The diagnostic odds ratio (DOR) summarizes the diagnostic accuracy of the index test as a single number that describes how many times higher the odds are of obtaining a test positive result in a diseased rather than a non-diseased person. The fact that it summarises test accuracy in a single number makes it easy to use this measure for meta-analysis as described in 10.5.1, but expressing accuracy in terms of ratios of odds means the measure has little direct clinical relevance, and it is rarely used as a summary statistic in primary studies. In fact, the clinician is usually interested in the sum of the number of false negative and false positive results whereas the DOR reflects their product. The DOR does, however, remain an important element in meta-analytic model building (see 10.5). It is formally defined as:

$$DOR= LR+/LR- = (sens \times spec)/(1-sens)\times(1-spec),$$ and is estimated as $(ad)/(bc)$.

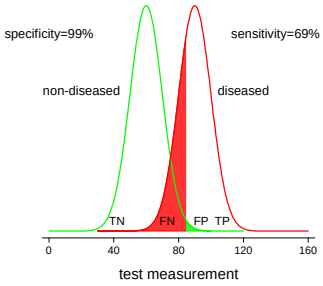
10.2.4 Positivity thresholds

Binary test outcomes are defined on the basis of a threshold for test positivity and change if the threshold is altered. This dependence on threshold is a fundamental aspect of diagnostic test evaluation. In the case of test sensitivity and specificity, the dependence induces a trade-off between the two quantities, one value increasing whilst the other decreases as the threshold for positivity is moved. This is illustrated in the panels in Figure 10.1, which each show the same hypothetical distributions of test results for diseased and non-diseased individuals on a continuous scale. The panels vary in the numerical value of the disease threshold used to define test positive. At each threshold, the sensitivity of the test is measured by the proportion of the area under the 'diseased' curve to the right of the threshold. Similarly, the specificity is measured by the proportion of the area under the 'non-diseased' curve to the left of the threshold. As the threshold decreases from panel (a) to panel (e), the proportion of those with disease who are above the threshold and hence have a positive test increases from 69% to 99%. These figures give the sensitivity of the test. At the same time the proportion of those without disease who are below the threshold and hence have a negative test result decreases from 99% to 69%. These figures give the specificity of the test.

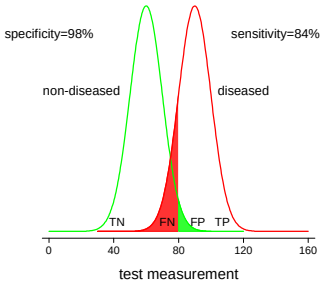
Throughout this chapter relationships of test performance are described presuming that higher test results are consistent with disease being present and lower tests results are consistent with disease being absent. If lower measures of the test quantity indicate disease, the relationships would be reversed.

Figure 10.1 Relationship between sensitivity, specificity and the positivity threshold

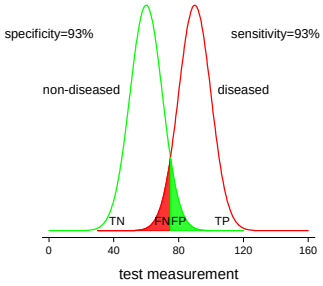
(a)



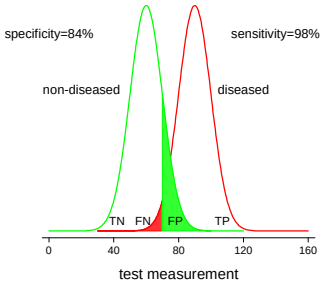
(b)



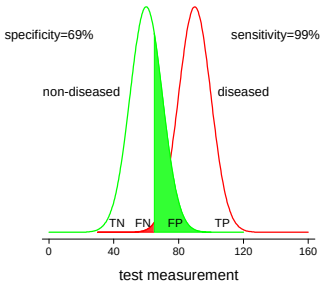
(c)



(d)



(e)



10.2.5 ROC curves

Primary studies that evaluate a test at several thresholds sometimes present results as ROC curves. The ROC curve of a test is the graph of the values of sensitivity and specificity that are obtained by varying the positivity threshold across all possible values. The graph plots sensitivity (true positive rate) against 1–specificity (false-positive rate). The curve for any test moves from the point where sensitivity and 1–specificity are both 1 (the upper right corner) which is achieved for a threshold at the lower end of its range (classifying all participants as test positive, so there are no false negatives but many false positives) to a point where sensitivity and 1-specificity are both zero (the lower left corner) which is achieved when the threshold moves to the upper end of its range (and all participants are classified as test negative, giving no false positives but many false negatives). The shape of the curve between these two fixed points depends on the discriminatory ability of the test.

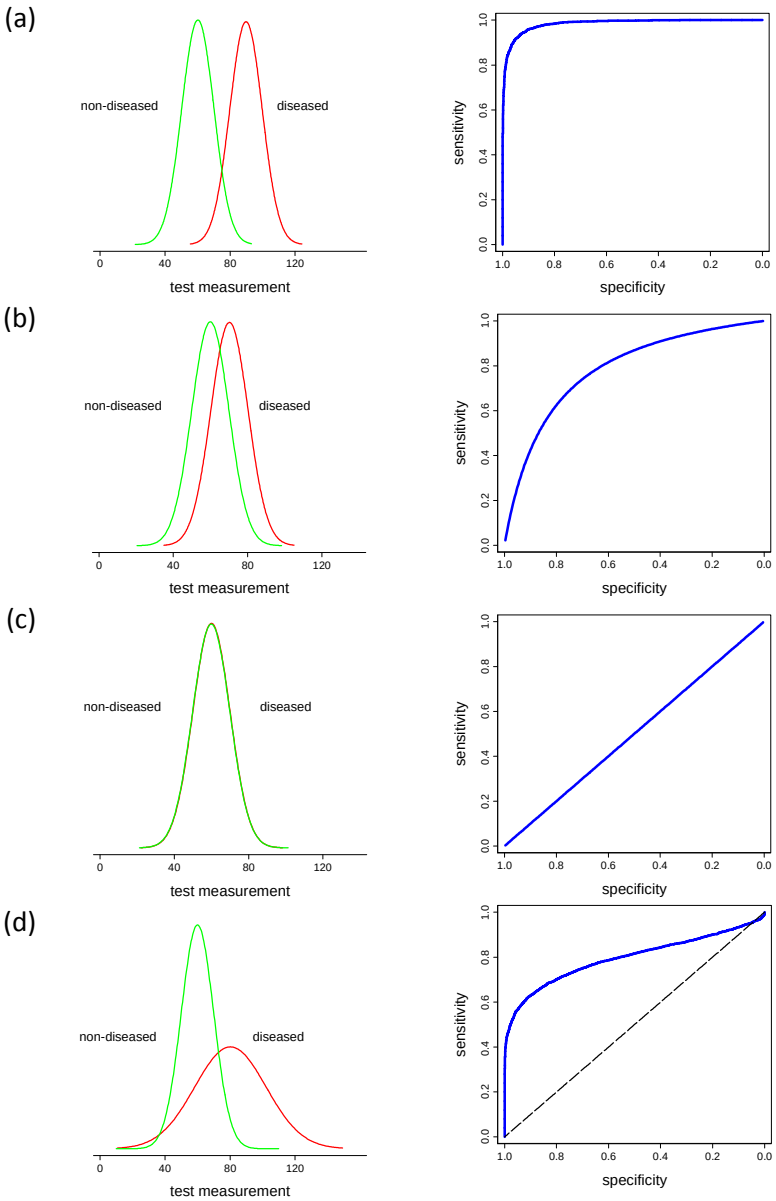
Figure 10.1 shows idealised distributions of test results for populations of diseased and non-diseased individuals, with shaded areas showing how the false negative rate (red) and the false positive rate (green) change as the positivity threshold varies. Figure 10.2(a) shows the resulting ROC curve. In practice, the ROC curve is estimated from a finite sample of test results and hence will not necessarily be a smooth curve as shown below. Note that the horizontal axis for each ROC plot in Figure 10.2 is labelled in terms of specificity decreasing from 1.0 to 0.0. This style of labelling is used in RevMan, and is equivalent to the usual labelling (1-specificity ranging from 0.0 to 1.0).

The position of the ROC curve depends on the degree of overlap of the distributions of the test measurement in diseased and non-diseased. Where a test clearly discriminates between diseased and non-diseased such that there is no or little overlap of distributions, the ROC curve will indicate that high sensitivity is achieved with a high specificity, that is the curve approaches the upper left hand corner of the graph where sensitivity is 1 and specificity is 1 (Figure 10.2(a)). If the distributions of test results in diseased and non-diseased coincide, the test would be completely uninformative and its ROC curve would be the upward diagonal of the square (Figure 10.2(c)).

The ROC curves shown in Figure 10.2(a)-(c) are all symmetrical about the sensitivity=specificity line (the downward diagonal of the square). It is also possible to get ROC curves which are not symmetrical as in Figure 10.2(d). Asymmetrical curves typically occur when the distribution of the test measurement in those with disease has more or less variability than the distribution in non-diseased people. Increased variability might occur, for example, where disease may cause a biomarker both to rise and become more erratic; reduced variability might occur where disease may lower biomarker values to a bounding level such as a lower level of detection.

The comparison of tests on the basis of their ROC curves takes into consideration their accuracy across a range of thresholds, and is aided by single summary statistics. Several such measures have been proposed in the literature. Most commonly used among them is the area under the curve (AUC), which equals 1 for a perfect test and 0.5 for a completely uninformative test. The AUC is equal to the probability that if a pair of diseased and non-diseased individuals is selected at random, the diseased individual will be have a higher test result than the non-diseased individual. The AUC can also be interpreted as an average sensitivity for the test, taken over all specificity values (or equally as the average specificity over all sensitivity values). Other summaries include partial areas under the curve, values of sensitivity corresponding to selected values of specificity (and vice versa), and optimal operating points, defined according to specified criteria.

Figure 10.2 Examples of ROC curves



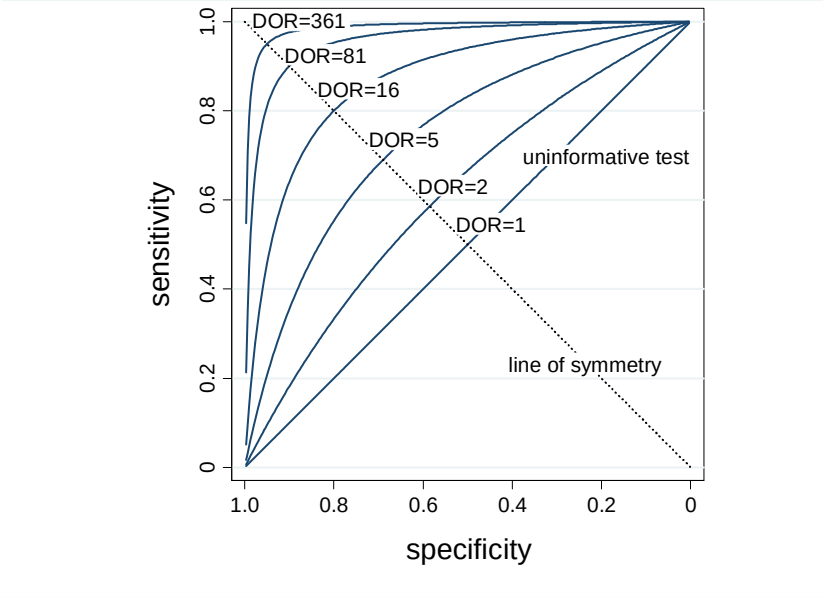
10.2.6 Relationships between ROC curves, diagnostic odds ratios and Q*

There is a useful link between ROC curves and diagnostic odds ratios which is important to appreciate to understand the way in which meta-analytical models are constructed. For the symmetric ROC curves displayed in Figure 10.3, all points on each curve have a common diagnostic odds ratio. This property arises when the test results in the diseased and non-diseased groups have a particular mathematical distribution known as a logistic distribution with equal variance in both groups. For example, a ROC curve with a diagnostic odds ratio of 21 would go through the (sensitivity, specificity) points of (0.70, 0.90), (0.82, 0.82) and (0.90, 0.70). Thus one way of summarising a symmetric ROC curve is by the value of the diagnostic odds ratio. Where ROC curves are asymmetric, the diagnostic odds ratio is not constant across the whole length of the curve but increases (or decreases) systematically with increasing threshold, and the curve can be mathematically described by noting how the diagnostic odds ratio changes with threshold, or a quantity related to threshold.

These relationships are not used in primary studies of tests, but form the basis of the ROC based meta-analytical models of test accuracy described in 10.4**Error! Reference source not found.Error! Reference source not found.** and 10.5 below.

ROC curves are sometimes described by quoting a point known as Q* where the ROC curve intersects the downward diagonal shown in Figure 10.3. By definition, at this point the sensitivity and specificity values are equal. The use of Q* values is discouraged in Cochrane reviews as they often give the wrong impression of the accuracy, particularly if SROC curves are asymmetric, or the study points lie away from the downward diagonal of the sensitivity=specificity line.

Figure 10. 3 Relationship between DOR and ROC curves



10.3 Graphical and tabular presentation

A Cochrane review of diagnostic test accuracy uses two main forms or graphical display, summary ROC plots and forest plots. Review authors create these figures within RevMan for each analysis that is specified.

10.3.1 Summary ROC plots

Summary ROC plots display the results of individual studies in ROC space, each study is plotted as a single sensitivity-specificity point. The size of points can be controlled to depict the precision of the estimate (typically scaled according to the inverse of the standard error of the logit(sensitivity) and logit(specificity)) or according to their sample sizes. In RevMan it is possible to mark studies as rectangles, with their height relating to the number of diseased (and hence precision of sensitivity estimate) and width relating to the number of non-diseased (and hence the precision of the specificity estimate).

Summary ROC plots depict the scatter of the study results. Occasionally 'cross-hairs' are added to each study point to indicate confidence limits for sensitivity and specificity, but this can make the plot very cluttered should there be many studies. This is not implemented in RevMan. Even if they depict the precision of the estimates from individual studies, it is difficult to gauge visually a sense of random variability versus heterogeneity.

Two types of meta-analytical summary can be added to the graph: summary ROC (SROC) curves and summary sensitivity and specificity points. Confidence regions for the summary sensitivity and specificity points can be included, as can prediction regions which give an indication of between study heterogeneity (see also 10.5.2.1).

Studies can also be plotted using different symbols or colours to indicate attribution to different subgroups for investigations of heterogeneity or for test comparisons.

10.3.2 Linked ROC plots

Linked ROC plots are used in analyses of pairs of tests, where both tests have been evaluated in each study. The points are plotted as in a normal summary ROC plot, but the two estimates (one for each test) from each study are joined by a line. It is thus possible to get a sense of the change in accuracy within study between the tests, and to note the degree of consistency in this change. Summary estimates of sensitivity and specificity for each tests, as well as summary ROC curves obtained from meta-analysis can be added to these plots (see 10.5.4.5 for an example plot).

10.3.3 Coupled forest plots

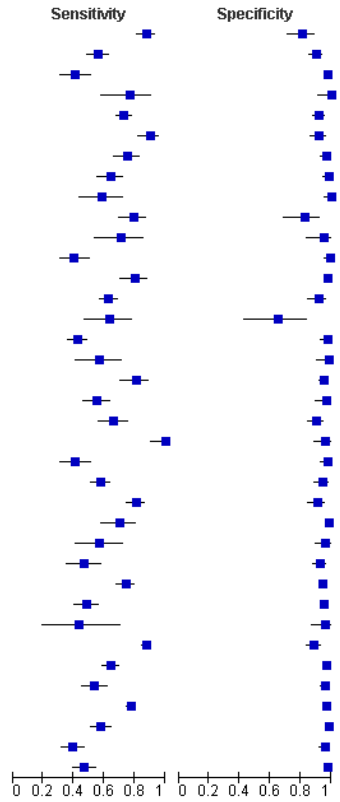
Forest plots for diagnostic test accuracy report the number of true positives and false negatives in diseased and true negatives and false positives in non-diseased participants in each study, and the estimated sensitivity and specificity, together with confidence intervals. The plots are known as coupled forest plots as they contain two graphical sections: one depicting sensitivity, and one specificity. The order of the studies can be sorted, often they are presented sorted by values of sensitivity, or grouped by test type or covariate values. Whilst it is possible to observe heterogeneity in sensitivity and specificity individually on such plots, it is not as easy to visualise whether there are threshold-like relationships. Summary statistics computed from meta-analyses are rarely added to coupled forest plots. In Cochrane DTA reviews an archive of coupled forest plots for all the tests for which data are entered into RevMan is published with the review to make the 2x2 tables widely accessible.

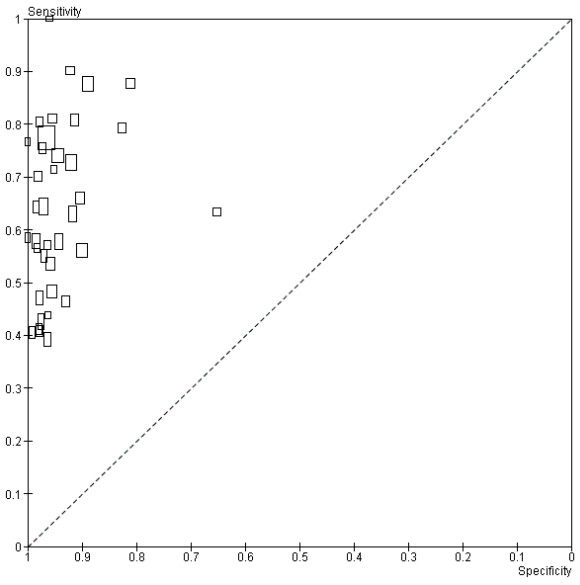
10.3.4 Example 1: Anti-CCP for the diagnosis of rheumatoid arthritis - Descriptive Plots.

These data are taken from a review (Nishimura 2007) of anti-cyclic citrullinated peptide antibody (anti-CCP). The reference standard was based on the 1987 revised American College of Rheumatology (ACR) criteria or clinical diagnosis. Thirty seven studies were included in the meta-analysis and their sensitivities and specificities are shown on the forest plot, and the study specific estimates are also shown in a scatterplot in ROC space below.

The forest plot below shows the studies in alphabetical order. The figure gives the numbers for the 2×2 table (TP, FP, FN, TN) for each study which will form the basis for statistical analyses. Study specific estimates of sensitivity and specificity are shown, with their 95% confidence intervals. These estimates (and confidence intervals) are also shown graphically. The most striking feature of this figure is the greater uncertainty (indicated by the confidence interval width) and variability (indicated by the scatter of point estimates) in sensitivity than specificity. The studies can be ordered in different ways (e.g. in increasing order of sensitivity) to provide a visual representation of any association between sensitivity and specificity. The figure also includes information on a covariate, the CCP generation, which may be associated with heterogeneity in test accuracy. (This will be explored in 10.5.3.1).

Study	TP	FP	FN	TN	Generation	Sensitivity	Specificity
Aotsuka 2005	115	17	16	73	CCP2	0.88 [0.81, 0.93]	0.81 [0.71, 0.89]
Bas 2003	110	24	86	215	CCP1	0.56 [0.49, 0.63]	0.90 [0.85, 0.93]
Bizzaro 2001	40	5	58	227	CCP1	0.41 [0.31, 0.51]	0.98 [0.95, 0.99]
Bombardieri 2004	23	0	7	39	CCP2	0.77 [0.58, 0.90]	1.00 [0.91, 1.00]
Choi 2005	236	20	88	231	CCP2	0.73 [0.68, 0.78]	0.92 [0.88, 0.95]
Correa 2004	74	11	8	130	CCP2	0.90 [0.82, 0.96]	0.92 [0.86, 0.96]
De Rycke 2004	89	4	29	142	CCP2	0.75 [0.67, 0.83]	0.97 [0.93, 0.99]
Dubucquoi 2004	90	2	50	129	CCP2	0.64 [0.56, 0.72]	0.98 [0.95, 1.00]
Fernandez-Suarez 2005	31	0	22	75	CCP2	0.58 [0.44, 0.72]	1.00 [0.95, 1.00]
Garcia-Berrocal 2005	69	8	18	38	CCP2	0.79 [0.69, 0.87]	0.83 [0.69, 0.92]
Girelli 2004	25	2	10	40	CCP2	0.71 [0.54, 0.85]	0.95 [0.84, 0.99]
Goldbach-Mansky 2000	43	1	63	120	CCP1	0.41 [0.31, 0.51]	0.99 [0.95, 1.00]
Greiner 2005	70	5	17	228	CCP2	0.80 [0.71, 0.88]	0.98 [0.95, 0.99]
Grootenboer-Mignot 2004	167	8	98	88	CCP2	0.63 [0.57, 0.69]	0.92 [0.84, 0.96]
Hitchon 2004	26	8	15	15	CCP2	0.63 [0.47, 0.78]	0.65 [0.43, 0.84]
Jansen 2003	110	3	148	118	CCP1	0.43 [0.37, 0.49]	0.98 [0.93, 0.99]
Kamali 2005	26	1	20	56	CCP2	0.57 [0.41, 0.71]	0.98 [0.91, 1.00]
Kumagai 2004	64	14	15	293	CCP2	0.81 [0.71, 0.89]	0.95 [0.92, 0.97]
Kwok 2005	71	2	58	66	CCP2	0.55 [0.46, 0.64]	0.97 [0.90, 1.00]
Lee 2003	68	14	35	132	CCP2	0.66 [0.56, 0.75]	0.90 [0.84, 0.95]
Lopez-Hoyos 2004	38	3	0	73	CCP2	1.00 [0.91, 1.00]	0.96 [0.89, 0.99]
Nell 2005	42	2	60	96	CCP2	0.41 [0.32, 0.51]	0.98 [0.93, 1.00]
Nielen 2005	149	7	109	114	CCP2	0.58 [0.51, 0.64]	0.94 [0.88, 0.98]
Quinn 2006	147	10	35	106	CCP2	0.81 [0.74, 0.86]	0.91 [0.85, 0.96]
Rantapaa-Dahlqvist 2003	47	7	20	375	CCP2	0.70 [0.58, 0.81]	0.98 [0.96, 0.99]
Raza 2005	24	3	18	79	CCP2	0.57 [0.41, 0.72]	0.96 [0.90, 0.99]
Saraux 2003	40	11	46	146	CCP1	0.47 [0.36, 0.58]	0.93 [0.88, 0.96]
Sauerland 2005	171	26	60	443	CCP2	0.74 [0.68, 0.80]	0.94 [0.92, 0.96]
Schellekens 2000	72	14	77	298	CCP1	0.48 [0.40, 0.57]	0.96 [0.93, 0.98]
Soderlin 2004	7	2	9	51	CCP2	0.44 [0.20, 0.70]	0.96 [0.87, 1.00]
Suzuki 2003	481	23	68	185	CCP2	0.88 [0.85, 0.90]	0.89 [0.84, 0.93]
Vallbracht 2004	190	12	105	408	CCP2	0.64 [0.59, 0.70]	0.97 [0.95, 0.99]
van Gaalen 2005	82	13	71	301	CCP2	0.54 [0.45, 0.62]	0.96 [0.93, 0.98]
van Venrooij 2004	865	79	252	2218	CCP2	0.77 [0.75, 0.80]	0.97 [0.96, 0.97]
Vincent 2002	139	7	101	464	CCP1	0.58 [0.51, 0.64]	0.99 [0.97, 0.99]
Vittecoq 2004	69	5	107	133	CCP2	0.39 [0.32, 0.47]	0.96 [0.92, 0.99]
Zeng 2003	90	7	101	313	CCP1	0.47 [0.40, 0.54]	0.98 [0.96, 0.99]





The ROC scatter plot shown below also shows the greater variability in estimated sensitivity than specificity across studies. Covariate information (e.g. generation of CCP) can be used to distinguish between studies in different subgroups (e.g. CCP1 vs CCP2), and ROC curves can be superimposed for a descriptive analysis. However, a more formal statistical analysis is required to provide summary estimates of test accuracy and to explore heterogeneity. These will be covered in 10.5.3.2 and 10.5.3.5. Before proceeding to these statistical analyses, the review author must decide whether it is appropriate to focus on a summary point(s) or a summary curve(s) in the statistical analyses that follow. This will be determined by the threshold(s) used by the studies to define a positive test result (see 10.5.2).

10.3.5 Tables of results

Review authors need to construct additional tables to report results from their meta-analytical models. Unlike for Cochrane intervention reviews, this output is not automatically included in the review document. Authors might consider creating tables for the following purposes:

- To report the numbers of studies and individuals available for each of the key analyses.
- To report the estimates of diagnostic accuracy for each of the tests
- To report statistics of comparative accuracy and tests of statistical significance for the pairwise comparisons between tests (a half-matrix display of all possible pairwise comparisons may be useful). Separate tables for direct and uncontrolled comparisons may be needed (see 10.5.4)
- Results of investigations of heterogeneity, including estimates of test accuracy in subgroups, summary statistics of comparative accuracy and tests of statistical significance (see 10.5.3)
- Results of sensitivity analyses (see 10.6.1)

This list is not exhaustive, and authors should use their inspiration to identify the best ways of communicating the results of their analyses.

Cochrane DTA reviews also include Summary of Results tables which are described in Chapter 11.

10.4 Meta-analytical summaries

Meta-analysis aims to compute and compare estimates of the expected diagnostic accuracy of a test and investigate the variability of results between studies. A choice needs to be made of which summary statistics are to be computed. In Cochrane reviews the choice is between estimating expected values of sensitivity and specificity for the test at a common threshold (referred to as the average operating point), or to estimate the expected ROC curve for a test across many thresholds (referred to as the summary ROC curve or SROC curve). Other summary statistics (such as likelihood ratios at the summary point and area(s) under the curve) can be computed from these summaries should they be required to assist interpretation and application of the results (see Chapter 11).

10.4.1 Should I estimate a SROC curve or a summary point?

In a systematic review it is likely that the collected data will be at a mixture of different positivity thresholds. Whilst for some tests there is consensus of what value the positivity threshold should take, more often tests are evaluated at different thresholds in different studies. Presentation of results at multiple thresholds within a single study is also encountered, with some studies presenting estimates of ROC curves (see 0) which depict the accuracy of the test at all possible thresholds. In addition, selective reporting of thresholds identified to optimise test accuracy can introduce bias if they are selected in a data driven manner (Leeflang 2008).

A key principle underlying the choice of statistical summary in meta-analysis of test accuracy is that the sensitivity and specificity of a test will vary as the positivity thresholds varies, as graphically depicted using a ROC curve (see 0). It is important to note that the hierarchical models recommended for meta-analysis for Cochrane DTA reviews account for correlation between sensitivity and specificity observed across studies which is due to the functional relationship between sensitivity and specificity as the threshold varies within each study. This occurs regardless of whether a summary ROC curve or a summary point is the output of choice.

A review author needs to decide whether they will use all the studies available to estimate the curve (in which case the meta-analysis will estimate the summary ROC curve) or to estimate a summary sensitivity and specificity point on this curve at a chosen threshold. Estimating summary sensitivity and specificity by pooling studies which mix thresholds will produce an estimate that relates to some notional unspecified average of the thresholds that occur in the included studies, which is clinically unhelpful and must be avoided.

Variation in threshold is highly likely where there is no explicit numerical cutpoint and definitions of a test positive are based on judgement rather than measurement. But even when it is possible to define a common cutpoint on the basis of a numerical value or a point on a rating scale, it must be acknowledged that there will still remain some variability in the actual threshold between studies through calibration differences between equipment, differences between raters or observers, as well as variation in the implementation of tests. The consequence of such variability will be additional heterogeneity in test results observed at the common cutpoint. The summary sensitivity and specificity point will reflect the average observed accuracy, whilst the prediction region will reflect the heterogeneity in how it is applied (see example 10.5.2.2).

Thus the two main strategies to handle mixed and variable thresholds in an analysis are:

- Estimating summary sensitivity and specificity of the test for a common threshold, or at each of several different common thresholds. Each study can contribute to one or more analyses depending on what thresholds it reports. Studies which do not report at any of the selected thresholds are excluded.
- Estimating the underlying ROC curve which describes how sensitivity and specificity trade-off with each other as thresholds vary. In this case one threshold per study is selected to be included in the analysis.

The choice of analytical approach will be influenced by the variation of thresholds in the available studies. For example, if there is little consistency in the thresholds used, meta-analyses which restrict to common thresholds will contain very little data, and estimating a summary ROC may be preferred. If there is little variation in threshold between studies attempting to fit a summary ROC curve will be difficult as the points are likely to be too tightly clustered in ROC space.

It is reasonable to estimate both SROC curves and average operating points in a review, as they may complement each other in providing clinically useful summaries, and powerful ways of detecting effects. For example, separate analyses of test data at different thresholds may be used to provide clinically informative estimates of sensitivity and specificity, whereas including all studies to estimate how summary ROC curves depend on covariates or test type will be the most powerful way to test hypotheses and investigate heterogeneity.

10.4.2 Meta-analytical methods not routinely used in Cochrane Reviews

Methods that are not routinely included in Cochrane reviews are commonly encountered in the literature for diagnostic meta-analysis. Separate pooling of sensitivity and specificity estimates fails to account for the trade-off between sensitivity and specificity, which may lead to underestimates of test accuracy (Deeks 2001). Similarly separate pooling of likelihood ratios ignores correlations

between positive and negative likelihood ratios, and theoretically can produce estimates which are impossible (Zwinderman 2008).

Pooling of predictive values is possible using the Bivariate method, but is not recommended as it is known that predictive values depend on prevalence which is likely to vary between studies. The consequences of this are two-fold: firstly that between study variation in prevalence may induce greater heterogeneity than is observed for sensitivity and specificity, and secondly that the average predictive values will relate to use of the test at some average, but unknown, prevalence.

10.4.3 Heterogeneity

Heterogeneity is to be expected in meta-analyses of diagnostic test accuracy. A consequence of this is that meta-analyses of test accuracy studies tend to focus on computing *average* rather than *common* effects. In systematic reviews of interventions it is sometime noted that the estimates of the effect of the intervention in the different studies are very similar, the differences between them being small enough to be explicable by chance. In such situations it is appropriate to use a *fixed effect* approach meta-analysis that estimates the underlying *common effect* (and is interpreted as the actual effect of the intervention). In test accuracy reviews large differences are commonly noted between studies, too big to be explained by chance, indicating that actual test accuracy varies between the included studies, or that there is heterogeneity in test accuracy. *Random effects* meta-analysis methods are recommended when data are heterogeneous, which focus on providing an estimate of the *average accuracy* of the test, and describing the variability in this effect. In Cochrane DTA reviews, heterogeneity is presumed to exist and random effects models are fitted by default, only simplified to fixed effect models where there are too few studies to estimate between study variability, or analysis demonstrates that fixed effects are appropriate.

Univariate tests for heterogeneity in sensitivity and specificity and the estimates of the I^2 statistic (Higgins 2003) are not routinely used in Cochrane DTA reviews as they do not account for heterogeneity explained by phenomena such as positivity threshold effects. If in a meta-analysis there is variation in threshold, what is of importance is the degree to which the observed study results lie close to the summary ROC curve, not how scattered they are in ROC space. The magnitude of observed heterogeneity is best depicted graphically where such relationships can be observed by the scatter of points and from the prediction ellipse. The numerical estimates of the random effect terms in the hierarchical models do quantify the amount of heterogeneity observed, but are not easily interpreted as they represent variation in parameters expressed on log odds scales.

10.5 Model fitting

10.5.1 Moses-Littenberg SROC curves (RevMan)

The Moses-Littenberg method (Moses 1993) (Littenberg 1993) provides a simple model for deriving a SROC. It was one of the earliest models to be proposed and has been used extensively in meta-analyses of diagnostic test accuracy. It is more akin to a fixed effect than a random effects model, as it does not provide estimates of the heterogeneity between studies. Even though it has been superseded by more complex hierarchical models that properly allow for random effects in diagnostic test accuracy, the Moses-Littenberg model is used in RevMan to provide reviewers with the facility to undertake purely exploratory analyses based on SROC curves without needing to export data out of RevMan. Because of the limitations of the Moses-Littenberg method, RevMan

does not provide parameter estimates or standard errors from this model as inferences should be based on hierarchical models that take separate account of within study sampling error and additional unexplained heterogeneity between studies.

A brief description of the Moses-Littenberg method is provided here to explain how the SROC curves produced by RevMan are derived. The method proceeds in three steps:

(i) the pairs of sensitivity and specificity estimates from each study are transformed onto the log odds (logit) scale to compute,

$$\mathrm{D} = \mathrm{logit}(sensitivity) - \mathrm{logit}(1 - specificity),\ \text{and}$$

$$S = \mathrm{logit}(sensitivity) + \mathrm{logit}(1 - specificity)$$

where D is the natural logarithm of the diagnostic odds ratio (lnDOR) and S is a quantity related to the overall proportion of positive test results. S can be considered as a proxy for test threshold since S will increase as the overall proportion of test positives, in the diseased and non-diseased groups, increases. The relationship between D and S is expected to be linear.

(ii) The simple linear regression model $\mathrm{D} = \alpha + \beta \mathrm{S} + error$ characterizes how test accuracy, as measured by the diagnostic log odds ratio (D), varies with S, a proxy of the positivity threshold across studies.

(iii) The estimates of α and β are then used to obtain the estimated sensitivity across a chosen range of possible values of specificity using

$$E(sensitivity) = 1/\left[1 + \exp(-\left[\alpha + (1+\beta)\mathrm{logit}(1 - specificity)\right]/(1-\beta))\right].$$

This will provide the estimated SROC curve in the original ROC coordinates. The range of specificities over which the curve is drawn is usually confined to the range observed in the data to avoid extrapolation.

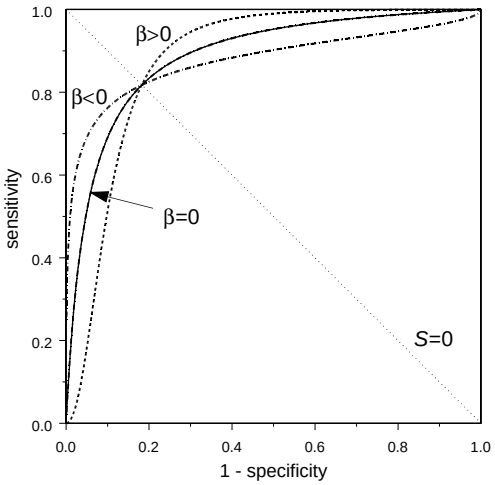
10.5.1.1 Properties of Moses-Littenberg SROC curve

Figure 10.4 illustrates three possible SROC curves that could arise from the Moses-Littenberg model. All share the same value of α (taken to be 3 for each curve), but with varying β (taken to be -0.35, 0 and 0.35). The point of intersection of all three curves lies on the diagonal where sensitivity=specificity (*S*=0). The sensitivity and specificity of the test at this point is also referred to as *Q** (see 10.2.6). When $\beta = 0$, the curve is symmetric about the diagonal line given by S=0. The lnDOR is the same (and equal to α) at every point on this symmetric curve since there is no association between accuracy (D) and threshold, (S) in the model. However, when $\beta \neq 0$, the curve is not symmetric and the expected accuracy (lnDOR) increases (or decreases) with threshold.

It is possible in some datasets for the estimated value of β to lead to improper SROC curves which do not go through the bottom left (sensitivity=0, specificity =1) and top right (sensitivity=1, specificity=0) corners of the SROC plot. If $\beta \geq 1$ or $\beta \leq -1$ the estimated SROC curve has the unintuitive property that sensitivity decreases as 1-specificity (false positive rate) increases. Such situations may arise if there are outlying studies that are influential in determining the slope of the

regression line. Excluding the outlier study allows assessment of its influence on the fitted SROC curve. Extreme values of β may also result if there is heterogeneity in test accuracy between subgroups of studies. Such heterogeneity can be explored through subgroup analyses when there are sufficient studies to allow for this.

Figure 10.4. SROC curves for alternative values of model parameters



10.5.1.2 Choice of weights

The regression line can be fitted using the method of weighted least squares (WLS) to account for differences in the sampling error in D between studies by weighting each study by the inverse variance of $\ln DOR$ for that study (estimated as $\text{var}(\ln DOR) = 1/a + 1/b + 1/c + 1/d$, where *a*, *b*, *c* and *d* represent the cells of the 2x2 table shown in Table 10.1). An alternative approach is to assign equal weight to all studies on the basis that the unexplained heterogeneity in test accuracy between studies is likely to be large compared with the variability due to sampling error (Moses 1993) (Irwig 1995). Both weighted and unweighted (equally weighted) least squares are implemented in RevMan. In practice, both weighting schemes often lead to similar curves.

Neither approach addresses the issue of sampling error in the explanatory variable (S) (violating a basic assumption of linear regression) and do not deal appropriately with additional unexplained heterogeneity in D. Consequently the Moses-Littenberg method for SROC analysis described above is used only for preliminary exploratory analyses and should not be used to compute confidence intervals for summary estimates of test accuracy, or to establish whether differences between subgroups are within the bounds of what we expect to see by chance alone.

10.5.2 Hierarchical models

More statistically rigorous approaches based on hierarchical models have been proposed that overcome the limitations of the Moses-Littenberg method. In this section, the Bivariate model (Reitsma 2005) and the hierarchical SROC (HSROC) model of Rutter and Gatsonis (Rutter 2001) are described and discussed.

Both hierarchical models involve statistical distributions at two levels. At the lower level, they model the cell counts in the 2×2 tables extracted from each study using binomial distributions and logistic (log-odds) transformations of proportions. At the higher level, random study effects are assumed to account for heterogeneity in diagnostic test accuracy between studies beyond that accounted for by sampling variability at the lower level. The Bivariate model and Rutter and Gatsonis HSROC model are mathematically equivalent when no covariates are fitted (Harbord 2007), (Arends 2008), but differ in their parametrizations. The Bivariate parametrization models sensitivity, specificity and the correlation between them directly, whereas the Rutter and Gatsonis HSROC parameterization models functions of sensitivity and specificity to define a summary ROC curve.

Parameter estimates from both the Bivariate model or Rutter and Gatsonis HSROC model can be input to RevMan to produce

- the summary ROC curve,
- the summary operating point, (i.e. summary values for sensitivity and specificity),
- a 95% confidence region around the summary operating point, and
- a 95% prediction region.

This prediction region is one way of illustrating the extent of statistical heterogeneity by depicting a region within which, assuming the model is correct, we have 95% confidence that the true sensitivity and specificity of a future study should lie (Harbord 2007).

From the summary ROC curve the expected sensitivity at a given value of specificity (or vice-versa) can be computed. In addition, summary values and confidence intervals can also be derived for the positive and negative likelihood ratios or the diagnostic odds ratio at the summary point.

Not all of these possible summary measures will be relevant or appropriate for a given analysis. The choice of summary measure(s) must be informed by the research question and also the variability in thresholds used across studies for defining test positivity.

The motivation for choosing one of these two alternative hierarchical models becomes clear when covariates are to be added to explore heterogeneity in test accuracy. Ultimately, the choice of method will be determined by the focus one wishes to adopt, and which of the two directly addresses the research question (see 10.4.1).

Both models require the use of external statistical software, as fitting them requires methods that are too complex to implement within RevMan. However, publication ready graphical output can be created in RevMan by estimating parameter estimates from either model to add model summaries to summary ROC plots.

Alternative specifications for summary curves based on functions of the Bivariate model parameters have recently been proposed (Arends 2008), (Chappell 2009). These require further evaluation and are not supported currently in RevMan. This chapter will focus on the Rutter and Gatsonis model as it is the most established of the HSROC specifications.

10.5.2.1 Bivariate model

The Bivariate method models the sensitivity and specificity directly. The model can be regarded as having two levels corresponding to variation *within* and *between* studies. At the first level, the within study variability for both sensitivity and specificity is assumed to follow a binomial distribution. For sensitivity (denoted by *A*), the number testing positive $y_{Ai} \sim B(n_{Ai}, \pi_{Ai})$ where n_{Ai} and π_{Ai} respectively represent the total number of diseased individuals tested and the probability of a positive test result in that group in study i. Similarly, for specificity (denoted by *B*), the number testing negative $y_{Bi} \sim B(n_{Bi}, \pi_{Bi})$ where n_{Bi} and π_{Bi} respectively represent the total number of non-diseased individuals tested and the probability of a negative test result in that group study i. The sensitivity-specificity pair for each study must be modelled jointly within study at level one of the analysis because they are linked by shared study characteristics including the positivity threshold. At the higher level, the logit-transformed sensitivities are assumed to have a normal distribution with mean μ_A and variance σ^2_A, while the logit-transformed specificities have a normal distribution with mean μ_B and variance σ^2_B. Their correlation is included by modelling both at once by a single *bivariate* normal distribution:

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

where σ^2_A and σ^2_B describe the between-study variability in true logit sensitivity and specificity respectively, and σ_{AB} is the covariance between logit sensitivity and specificity. The model may also be parameterized using the correlation $\rho_{AB} = \sigma_{AB} / (\sigma_A \sigma_B)$, which may be more interpretable than the covariance. The Bivariate model therefore has five parameters when no covariates are included: μ_A, μ_B, σ^2_A, σ^2_B and ρ_{AB}. (Note: we follow Harbord (Harbord 2007) in using μ where Reitsma (Reitsma 2005) used θ in order to avoid confusion with the notation from that of the HSROC model which follows).

The inclusion of a correlation parameter in the model allows for the expected trade-off in sensitivity and specificity as the test positivity threshold across studies varies. Where variation between studies arises through such a trade-off this correlation is expected to be negative, but the correlation may be positive if there are other sources of heterogeneity.

Reitsma (Reitsma 2005) originally proposed fitting these models by approximating the binomial within-study distributions by normal distributions. Although this allows the model to be fitted in a slightly larger range of software (e.g. the MIXED procedure in SAS), Chu (Chu 2006) later demonstrated that the approximation can perform poorly and recommended that software be used that can explicitly model the binomial within-study distributions.

10.5.2.2 Example 1 continued: Anti-CCP for the diagnosis of rheumatoid arthritis.

We now undertake the first stage of a formal statistical analysis of the data from a review (Nishimura 2007) of anti-cyclic citrullinated peptide antibody (anti-CCP). If it can be presumed that the anti-CCP test is deemed positive if any anti-CCP antibody is detected and that detection can be considered a common threshold, it makes sense to focus on summary estimates for sensitivity and specificity.

As noted in the descriptive analyses of these data, there appears to be greater variability in estimated sensitivity than specificity across studies, which could arise either through heterogeneity or through estimates of sensitivity being based on smaller samples than estimates of specificity. The parameter estimates from the Bivariate model are shown below.

Fit Statistics

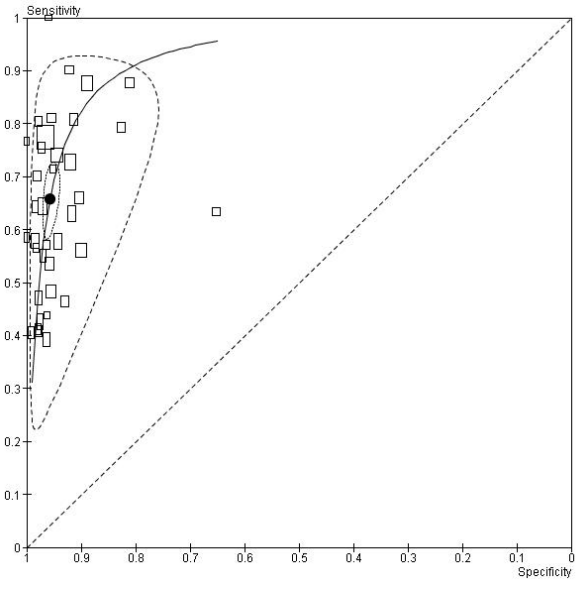
-2 Log Likelihood	545.6
AIC (smaller is better)	555.6
AICC (smaller is better)	556.4
BIC (smaller is better)	563.6

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
msens	0.6534	0.1275	35	5.13	<.0001	0.05	0.3946	0.9122	3.959E-6
mspec	3.1090	0.1459	35	21.31	<.0001	0.05	2.8128	3.4052	3.473E-8
s2usens	0.5426	0.1463	35	3.71	0.0007	0.05	0.2455	0.8397	-6.62E-6
s2uspec	0.5717	0.1873	35	3.05	0.0043	0.05	0.1914	0.9520	1.36E-6
covsesp	-0.2704	0.1199	35	-2.26	0.0304	0.05	-0.5137	0.02710	-1.59E-6

Covariance Matrix of Parameter Estimates

Row	Parameter	msens	mspec	s2usens	s2uspec	covsesp
1	msens	0.01625	-0.00741	0.000890	-0.00004	-0.00004
2	mspec	-0.00741	0.02128	-0.00006	0.004286	-0.00116
3	s2usens	0.000890	-0.00006	0.02142	0.003997	-0.00874
4	s2uspec	-0.00004	0.004286	0.003997	0.03509	-0.01184
5	covsesp	-0.00004	-0.00116	-0.00874	-0.01184	0.01436



The parameter estimates in the boxes above can be input to RevMan to produce the summary point, 95% confidence region, and 95% prediction region shown in the Figure. The Bivariate output box in RevMan requires: the summary estimate for logit(sensitivity) which is 0.6534, the summary estimate for logit(specificity) which is 3.1090; and the variances of the random effects for logit(sensitivity), logit(specificity) and their covariance which are 0.5426, 0.5717 and -0.2704 respectively (all of these estimates appear in the red box). Computation of confidence and prediction regions also requires the standard error of the summary estimates for logit(sensitivity), logit(specificity) and their covariance which are 0.1275, 0.1459 and -0.00741 respectively (shown in the blue boxes).

The variance coefficients indicate similar heterogeneity in sensitivities and specificities. The magnitude of the heterogeneity is also evident in the size of the prediction region on the SROC plot. The summary estimate of sensitivity and specificity is shown by the solid black dot. The sensitivity

and specificity at this point can be computed by inverse transformation of the logit estimates to give a sensitivity and specificity of 0.66 and 0.96 respectively. Confidence intervals can be computed by inverse transformation of intervals computed on the logit scale.

The plot shows a potential outlier, Hitchon 2004 with a sensitivity of 0.63 and specificity of 0.65. A sensitivity analysis can be performed to assess the influence of this study on the summary estimates.

10.5.2.3 The Rutter and Gatsonis HSROC model

The HSROC model proposed by Rutter and Gatsonis (Rutter 1995), (Rutter 2001) is based on a latent scale logistic regression model (McCullagh 1980), (Tosteson 1988). The HSROC model assumes that there is an underlying ROC curve in each study with parameters α and β that characterize the accuracy and asymmetry of the curve, in a similar (though technically distinct) way to the α and β parameters in the linear regression method of Moses and Littenberg. Unlike the Moses-Littenberg model, the Rutter and Gatsonis model is constrained to provide a ROC curve where sensitivity cannot decrease as specificity increases.

Accuracy, defined in terms of the lnDOR, determines the position of the summary curve relative to the top left corner of the ROC axes. As with the SROC regression method, each study contributes data at a single threshold to the analysis. The 2×2 table for each study then arises from dichotomizing at a positivity threshold denoted by θ. The parameters α and θ are assumed to vary between studies: both are assumed to have normal distributions as in conventional random-effects meta-analysis.

The HSROC model can also be regarded as having two levels corresponding to variation *within* and *between* studies. At the first level, the number of diseased individuals who test positive is denoted by y_{i1} for the i^{th} study, and the corresponding number of non-diseased who test positive is denoted by y_{i2}. For each study (i), the number testing positive in each disease group (j) is assumed to follow a binomial distribution such that $y_{ij} \sim B(n_{ij}, \pi_{ij}),\ j = 1,2$ where n_{ij} and π_{ij} respectively represent the total number tested and the probability of a positive test result. The number testing positive in each diseased and non-diseased pair is analysed jointly within each study at level one of the analysis.

The model takes the form

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

where dis_{ij} represents the 'true' disease status (coded as -0.5 for the non-diseased and 0.5 for the diseased) thereby taking into account the within study variability at level one. Using the usual terminology for this model, we generally refer to θ_i represents the proxy for positivity threshold calculated as the mean of the log odds of a positive test result for the diseased and the log odds of a positive test result for the non-diseased groups in study i (equivalent to $S_i/2$ in the Moses-Littenberg model). α_i (the lnDOR for study i) represents a measure of diagnostic accuracy in the i^{th} study that incorporates both sensitivity and specificity for that study. The scale parameter (β) provides for asymmetry in the SROC by allowing accuracy to vary with threshold. Since each study

contributes only one estimate of sensitivity and specificity at a single threshold, it is necessary to assume that the shape of the true underlying ROC curve in each study is the same, and hence β is fitted as a fixed effect.

The threshold and diagnostic accuracy for each study are specified as random effects and are assumed to be independent (uncorrelated) and normally distributed. The accuracy parameter has mean Λ (capital lambda) and variance σ_α^2, while the positivity (threshold) parameter has mean Θ (capital theta) and variance σ_β^2. The shape parameter (β) is estimated using data from the studies considered jointly, assuming normally distributed random effects for test accuracy. When no covariates are included, the HRSOC model also has five parameters: Λ, Θ, β, σ_α^2 and σ_β^2.

A summary ROC curve can be constructed from the HSROC model by choosing a range of values of 1-specificity and using the estimated average location parameter (Λ) and scale parameter (β) to compute the corresponding values for sensitivity. The expected sensitivity at a chosen false positive fraction (1–specificity) is given by

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

When $\beta = 0$, test accuracy can be summarized by Λ which represents the expected accuracy (log DOR), and the resulting summary curve will be symmetric.

10.5.2.4 Example 2: Rheumatoid Factor as a marker for Rheumatoid Arthritis.

In this example we will investigate the diagnostic performance of Rheumatoid factor (RF) as a marker for rheumatoid arthritis (RA). The 50 studies included in the analysis are taken from the same review as Example 1 (Nishimura 2007). The reference standard was again based on the 1987 revised American College of Rheumatology (ACR) criteria or clinical diagnosis.

The cut-off for test positivity for RF varied between studies and ranged from 3 to 100 U/ml. The variability in threshold used to define test positivity between studies is reflected in the variability in study specific estimates of sensitivity and specificity shown in the SROC plot shown in the Figure. Because of the variation in threshold across studies, a summary ROC curve is appropriate to summarise these data. The HSROC model was used to estimate a summary curve using Proc NLMIXED in SAS.

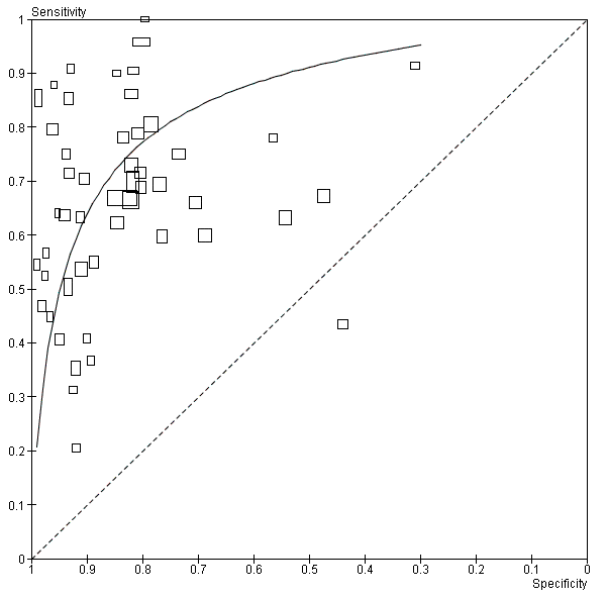
Proc NLMIXED Output:

Fit Statistics

-2 Log Likelihood	806.9
AIC (smaller is better)	816.9
AICC (smaller is better)	817.6
BIC (smaller is better)	826.5

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
alpha	2.6016	0.1862	48	13.97	<.0001	0.05	2.2273	2.9759	2.227E-6
theta	-0.4370	0.1469	48	-2.98	0.0046	0.05	-0.7323	-0.1417	4.573E-6
beta	0.2267	0.1624	48	1.40	0.1691	0.05	-0.09978	0.5532	-1.16E-6
s2ua	1.3014	0.3046	48	4.27	<.0001	0.05	0.6890	1.9137	-6.42E-7
s2ut	0.5423	0.1237	48	4.39	<.0001	0.05	0.2937	0.7909	-6.99E-6



The parameter estimates highlighted above can be input to RevMan to draw the summary curve as shown in Figure; 2.6016 estimates the mean of the random effects for accuracy (i.e. Λ, lambda), -0.4370 estimates the mean of the random effects for threshold (theta), 0.2267 estimates the shape parameter (beta), 1.3014 estimates the variance of the random effects for accuracy, and 0.5423 estimates the variance of the random effects for threshold. The resulting curve shows the expected trade-off between sensitivity and specificity across thresholds.

When interpreting the results of the analysis, it is important to note that RF constitutes part of the ACR criteria. Hence, there is risk of bias in the estimated curve since the index test is incorporated in the reference standard. This could result in an overestimation of the diagnostic accuracy of RF, and could result in giving a distorted picture of the expediency of using RF as a first test for resolving uncertainty in a suspected case of rheumatoid arthritis.

10.5.3 Investigating heterogeneity

In diagnostic reviews it is usual to observe variability in test accuracy between studies that is considerably greater than would be expected from within study sampling error alone. This is reflected in the model specifications for the Bivariate and HSROC models which both allow for random study effects. For the Bivariate model, the summary estimates of sensitivity and specificity represent an *average* operating point across studies. Similarly, the estimated summary ROC curve represents an *average* ROC curve across studies on the assumption that the true underlying ROC curve in each study has the same shape.

Some of this heterogeneity in test accuracy between studies is likely to arise due to differences in patient characteristics, test methods, study design and other factors. Exploratory analyses can be conducted in RevMan to investigate whether such study characteristics appear to be associated with test accuracy using the Moses-Littenberg SROC method, but this method cannot be used to provide valid statistical evidence of such associations. A separate SROC curve is fitted for each subgroup, and the results can be compared graphically across subgroups. The feasibility of such analyses will obviously be influenced by the number of available studies in each subgroup.

Statistically, it is generally more efficient to make use of all of the data available across studies when investigating heterogeneity by adding study level covariates to a hierarchical model to identify factors associated with diagnostic test accuracy. This meta-regression approach also allows statistical inferences to be made. It is usually assumed that each covariate has a fixed effect when added to the model. This approach is also applicable to test comparisons, as discussed in 10.5.4.

The Bivariate and HSROC models differ in how study level covariates are included. Published accounts of the Bivariate method focus on the estimation of a summary estimate of sensitivity and specificity, and how the expected values of these may vary with study level covariates. Published accounts of the HSROC approach, by contrast, focus on the estimation of the summary ROC curve as the basis for assessing test accuracy, and how the position and shape of the curve may vary with study level covariates.

Both models allow the use of categorical and continuous covariates. In practice, covariates relating to study characteristics are usually categorical and indicator variables are created as is done in standard regression modelling. For continuous covariates, particular care should be taken to check that the assumption of linear associations are valid. For the Bivariate model, this refers to association with logit(sensitivity) and/or logit(specificity). For the HSROC model, this refers to association with the accuracy parameter (lnDOR) and/or the threshold parameter.

The uses and limitations of investigating heterogeneity using sub-group analysis and meta-regression in Section 9.6 of the Cochrane Handbook for Systematic Reviews of Interventions (Deeks 2008) applies equally to diagnostic studies.

10.5.3.1 Heterogeneity and Regression Analysis using the Bivariate model

The Bivariate model allows covariates to affect summary sensitivity or summary specificity, or both. Using the notation of Harbord (Harbord 2007), and assuming that we have a single study level covariate Z that may affect both sensitivity and specificity, then the model can be extended as follows:

$$\begin{pmatrix} \mu_{Ai} \\ \mu_{Bi} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_A + v_A Z_i \\ \mu_B + v_B Z_i \end{pmatrix}, \Sigma \right)$$

As before, Σ represents the covariance matrix for the random effects for logit sensitivity and logit specificity. If the covariate does explain some of the heterogeneity in sensitivity and/or specificity then we would expect that the estimated variance for one or both random effects to be reduced. The estimated covariance (correlation) parameter may also change.

Assuming that we have a binary study level covariate (Z) coded as 0 or 1 to represent the two groups of studies, then μ_A estimates the logit sensitivity at the expected summary operating point for the referent group (Z =0), and $\mu_A + v_A$ estimates the logit sensitivity at the expected summary operating point for the other group (Z =1). Hence, $\exp(v_A)$ estimates the odds ratio for sensitivity in group 1 relative to the referent group. The expected sensitivity is estimated as $\exp(\mu_A)/(1+\exp(\mu_A))$ for the referent group of studies, and as $\exp(\mu_A + v_A)/(1+\exp(\mu_A + v_A))$ for the other group. Comparisons of specificity between the two groups of studies follow the same approach as described above based on μ_B and v_B. The fit of the model, with and without the additional parameters v_A and v_B, can be used to test whether the covariate is associated with sensitivity and or specificity. This joint test will have 2 degrees of freedom if Z is binary. Separate tests of statistical significance of the covariate with sensitivity and specificity can also be conducted, first to assess whether v_A differs from 0 (a significant result indicates that there is evidence that sensitivity differs between the two groups of studies) and secondly whether v_B differs from 0 (a significant result indicates that there is evidence that specificity differs between the two groups of studies). See also 10.5.3.4 relating to criteria for model selection.

The standard error of a new estimate based on a function of the model parameter estimates can be obtained using the delta method on the assumption that the error distribution of the new estimate is approximately normal. The delta method is implemented in standard statistical software such as SAS and Stata.

The model is easily extended to allow for more than one covariate. However, this may not be feasible in practice if the number of studies is not large. Also, it is important to note that a covariate may only be associated with sensitivity and not specificity, or vice versa. It is not required that the same covariates are fitted for both sensitivity and specificity, although this may commonly be the case. Where a covariate (or covariates) is allowed to affect both the sensitivity and the specificity, the Bivariate model is equivalent to an HSROC model in which the covariate or covariates are allowed to affect both the accuracy and the positivity threshold but not the shape parameter. However, using the estimates from the Bivariate model to test for the effect of covariates on the shape and position of the summary ROC curve is not straightforward. Using the HSROC model parameterization allows this to be done in a more direct and straightforward manner.

It is usually assumed that the variance of the random effects (and their correlation in the case of the Bivariate model) are not associated with the covariate. This is probably a reasonable assumption in most analyses investigating heterogeneity in test accuracy for a single index test. However, for analyses that compare different index tests, this assumption is less likely to hold. See 10.5.4.

10.5.3.2 Example 1 (cont).: Investigation of heterogeneity in diagnostic performance of anti-CCP

The studies included in the review to assess the diagnostic performance of anti-CCP used two different generations of the assay: first generation (CCP1, 8 studies) and second generation (CCP2, 29 studies). A binary covariate for generation of CCP with coefficient se2 for sensitivity, and and coefficient sp2 for specificity were added to the model. The covariate was coded as 0 for CCP1 and 1 for CCP2. Allowing test performance to vary by generation of CCP in the model resulted in a –2Log Likelihood of 533.4, a reduction of 12.2 compared with the model that contained no covariates. Hence, there is statistical evidence (chi-square=12.2, 2df, P=0.002) that test performance is associated with generation of CCP but further investigation is required to ascertain whether this association is for sensitivity, specificity, or both.

The parameter estimates required to draw the summary points and regions in RevMan can again be extracted from the Proc NLMIXED output (see Appendix for SAS program). The variances of the random effects for logit(sensitivity) and logit(specificity), and their covariance are common for both generations of CCP (see blue box in output). For the referent group (CCP1 in this case) the summary estimates for logit(sensitivity), logit(specificity), the corresponding standard errors and covariance are shown in the red boxes in the output. The logit(sensitivity) for CCP2 is estimated by msens+se2, and the logit(specificity) is estimated by mspec+sp2. The standard errors and covariance of these additional estimates can be obtained using the ESTIMATE command in Proc NLMIXED as shown in the program in the Appendix. Alternatively, a simple way of getting these results is to refit the model using CCP2 as the referent group (coded as 0) and CCP1 as the other group (coded as 1). The fit of the model and results for the random effects will be the same, but the estimates for 'msens' and 'mspec' will now be for CCP2 and hence the required estimates can then be extracted from the standard output (revised output not shown). The resulting plot is shown in the Figure below.

Fit Statistics

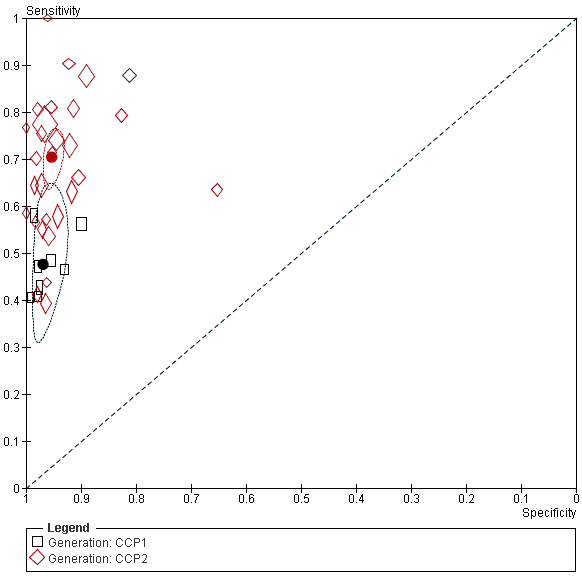
-2 Log Likelihood	533.4
AIC (smaller is better)	547.4
AICC (smaller is better)	549.1
BIC (smaller is better)	558.6

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
msens	-0.09653	0.2203	35	-0.44	0.6640	0.05	-0.5438	0.3507	0.000317
mspec	3.4467	0.2982	35	11.56	<.0001	0.05	2.8412	4.0522	-0.00005
s2usens	0.3598	0.1022	35	3.52	0.0012	0.05	0.1524	0.5673	8.325E-6
s2uspec	0.5399	0.1802	35	3.00	0.0050	0.05	0.1742	0.9057	0.000159
covsesp	-0.1969	0.09836	35	-2.00	0.0531	0.05	-0.3965	0.002824	-0.00004
se2	0.9626	0.2513	35	3.83	0.0005	0.05	0.4523	1.4728	0.000319
sp2	-0.4302	0.3377	35	-1.27	0.2111	0.05	-1.1158	0.2554	-0.00004

Covariance Matrix of Parameter Estimates

Row	Parameter	msens	mspec	s2usens	s2uspec	covsesp	se2	sp2
1	msens	0.04854	-0.02464	-0.00012	-0.00001	-0.00003	-0.04855	0.02465
2	mspec	-0.02464	0.08895	-0.00002	0.004771	-0.00065	0.02463	-0.08834
3	s2usens	-0.00012	-0.00002	0.01044	0.002118	-0.00440	0.000693	-0.00005
4	s2uspec	-0.00001	0.004771	0.002118	0.03246	-0.00860	-0.00006	-0.00039
5	covsesp	-0.00003	-0.00065	-0.00440	-0.00860	0.009674	0.000100	-0.00091
6	se2	-0.04855	0.02463	0.000693	-0.00006	0.000100	0.06317	-0.03160
7	sp2	0.02465	-0.08834	-0.00005	-0.00039	-0.00091	-0.03160	0.1141



Based on the confidence regions in the figure it is clear that the sensitivity varies by generation, but not specificity. The summary estimates of specificities were: 0.97 (95%CI 0.95, 0.98) for CCP1 and 0.95 (95%CI 0.94, 0.97). The summary estimates of sensitivity were 0.48 (95%CI 0.37, 0.58) for CCP1 and 0.70 (95% CI 0.65, 0.75) for CCP2. These results indicate an improvement in sensitivity, without loss of specificity for generation 2 compared with generation 1 CCP. Further models may be fitted to formally test the effect of removing the covariate for specificity from the model.

Comparing the output from this model with that of the model with no covariates (see 10.5.2.1), it is clear that the variances of the random effects have reduced, particularly for sensitivity. Also, checks of the distributions of the random effects (not shown here) show that adjusting for generation of anti-CCP results in distributions that more closely follow a normal distribution.

10.5.3.3 Heterogeneity and Regression Analysis using the Rutter and Gatsonis HSROC model

The HSROC model allows covariates to be added to explore heterogeneity in test positivity (threshold), position of the curve (accuracy) and shape of the curve. A covariate may be associated with some, but not all three model parameters.

Assuming that we have a binary study level covariate (Z) coded as 0 or 1 to represent the two groups of studies, then the HSROC model can be extended to estimate the log odds of a positive test for study *i* and disease group *j* as follows:

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

where γ, λ and δ are all assumed to be a fixed effect. Hence, the distribution of the random effects for threshold and accuracy are now given by $\theta_i \sim N\left(\Theta + \gamma Z_i, \sigma_\theta^2\right)$, and $\alpha_i \sim N\left(\Lambda + \lambda Z_i, \sigma_\alpha^2\right)$ respectively. The shape parameter for the summary curves for the two groups is estimated as β for the referent group of studies (Z =0) and $\beta + \delta$ for the other group (Z =1). If the covariate does explain some of the heterogeneity in threshold and/or accuracy then we would expect that the estimated variance for one or both random effects to be reduced.

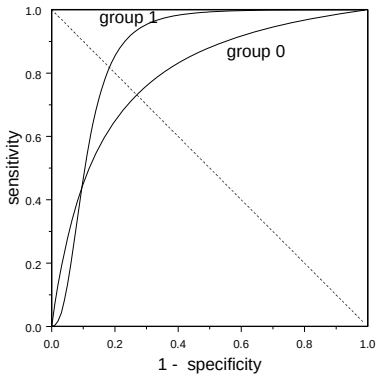
The first step would be to investigate the shape of the summary curve. If $\delta \neq 0$, then the shape of the summary curve differs for the two groups of studies which means that the relative accuracy of the test will vary with threshold. (Figure 10.5(a)) This represents the most complex scenario, and the model would not generally be simplified any further. In practice, it is difficult to detect a statistically significant difference in the shape of the curve across groups because the number of studies in each

group is usually limited. Also, it is important when investigating shape to consider the effect of outlying and potentially influential studies. When there is good evidence that the curves differ in shape, a plot of the estimated curves for the two groups will aid in interpretation. Focusing on the region of the plot that covers the observed data, it is then possible compare the estimated curves. Where one curve consistently lies above another, there is evidence of superior accuracy even though the differential between the curves will vary across thresholds. If the curves cross, then the interpretation of which curve shows superior accuracy will depend on threshold.

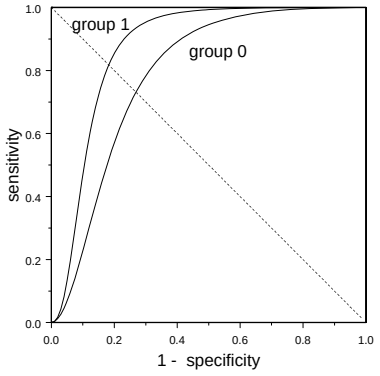
If, based on statistical evidence, similarity of curve shapes and investigation of potentially influential studies, it can be assumed that $\delta = 0$, then the covariate can be removed for shape. The estimated SROC curves for the two groups will then have the same shape, even though they are not symmetric (Figure 10.5(b)) , and the relative diagnostic accuracy of the two curves can be summarized using the relative diagnostic odds ratio ($RDOR = \exp(\lambda)$). The RDOR will be constant across all possible values of θ. If the model can be simplified further and both curves can be assumed to be symmetric, i.e. $\beta = 0$, the RDOR again provides a measure of relative accuracy as described above.

Figure 10.5 Summary ROC curves with and without a difference in shape

(a) relative accuracy depends on particular specificity values, the curves crossing



(b) group 1 dominates across all specificity values, the curves do not cross



If the curves can be assumed to have the same shape (either both asymmetric or both symmetric), then the question is whether the covariate is associated with accuracy. If there is evidence that $\lambda \neq 0$, then the RDOR gives an estimate of the overall relative diagnostic accuracy. This would correspond to a clear separation between the SROC curves for the two groups. Alternatively, $\lambda = 0$ implies that there is no separation between the curves and no association between the covariate and accuracy.

If λ can be assumed to be 0, then the model can be further simplified by removing the covariate for accuracy which will result in a single summary curve (assuming that the shape of the curve is the same for the two groups of studies). An association between the covariate and the threshold parameter (i.e. $\gamma \neq 0$) would indicate that the underlying test positivity rate for the two groups of studies differs. Such an association is often difficult to interpret unless the curves can be assumed to have the same shape and accuracy.

When the actual cut-point to define a positive test is available for each study, this can be fitted as a covariate to the threshold parameter to allow estimation of the expected sensitivity and specificity on the summary curve at a selected cut-point. However, this presumes a particular functional relationship between threshold and sensitivity and specificity.

10.5.3.4 Criteria for model selection

Irrespective of which model is used, reviewer authors must specify what modelling strategy will be used for adding or removing covariates and what criterion will be used to decide whether or not a covariate should be included in a model.

The decision as to whether a covariate should be retained in the model may be based in part on statistical tests. Commonly used software for fitting these models, such as SAS for instance, will provide Wald statistics and corresponding p-values for each variable in the model. A p-value based on the likelihood ratio chi-squared statistic is generally more reliable. The chi-squared statistic is computed as the change in the -2Log likelihood when a covariate is added (or removed) from a model. The degrees of freedom is equal to the difference in the number of parameters fitted in these models. The effect of adding (or removing) covariates on measures of model fit such as Akaike's information criterion (AIC) or Bayesian information criterion (BIC) can also be used. The deviance information criterion (DIC) is commonly used for models fitted by *Markov chain Monte Carlo* (MCMC) simulation

Likelihood ratio tests can also be used to assess the significance of the variance terms for the two random effects in either model, or whether allowing for variance to relate to test accuracy provides a better fitting model.

10.5.3.5 Example 2 (cont.): Investigating heterogeneity in diagnostic accuracy of Rheumatoid Factor

We will now investigate whether the laboratory technique used to measure RF is associated with diagnostic performance. Of the 50 studies, 15 used nephelometry (N), 16 latex agglutination (LA), 16 ELISA, one study used RA hemagglutination, and 2 did not report the method used. The analysis is restricted to studies that used N, LA or ELISA. The HSROC model was again used because of the variation in threshold used for test positivity across studies. Covariates (indicator variables for technique, using LA as the referent category) were included in the model to assess whether accuracy, threshold, or the shape of the SROC curve varied with technique.

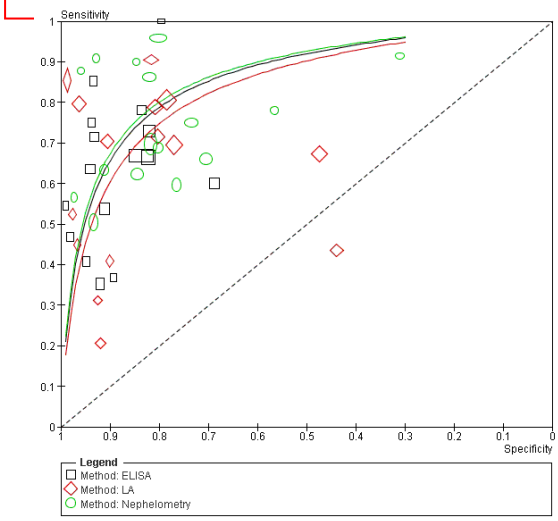
The -2Log likelihood for the most complex model that included covariates for shape, accuracy and threshold parameters was 752.9. The increase in the -2Log likelihood was negligible (an increase to 753.1) when the covariate for shape was removed from the model (chi-square = 753,1-752.9 = 0.2, 2 df, P=0.90). Parameter estimates for the model that assumes a common shape are given below, and the corresponding HSROC curves shown in the Figure. The estimates of alpha, theta and beta can be input to RevMan to obtain the summary curve for the referent group (LA). The variances of the random effects for threshold and accuracy are common to all three techniques, as is the shape parameter beta. However, the threshold and accuracy parameter estimates for ELISA are given by theta+t1 and alpha+a1 respectively, and for N are given by theta+t2 and alpha+a3.

Fit Statistics

-2 Log Likelihood	753.1
AIC (smaller is better)	771.1
AICC (smaller is better)	773.2
BIC (smaller is better)	787.7

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
alpha	2.4552	0.3245	45	7.57	<.0001	0.05	1.8017	3.1087	-0.0004
theta	-0.5490	0.2137	45	-2.57	0.0136	0.05	-0.9794	-0.1186	0.000139
beta	0.1995	0.1702	45	1.17	0.2472	0.05	-0.1432	0.5423	-0.00018
s2ua	1.2865	0.3109	45	4.14	0.0002	0.05	0.6603	1.9128	-0.00038
s2ut	0.4786	0.1139	45	4.20	0.0001	0.05	0.2492	0.7080	0.00062
a1	0.2483	0.4408	45	0.56	0.5760	0.05	-0.6395	1.1361	-0.00038
a2	0.3328	0.4439	45	0.75	0.4573	0.05	-0.5612	1.2269	0.000093
t1	-0.1962	0.2614	45	-0.75	0.4568	0.05	-0.7227	0.3303	-0.00017
t2	0.4960	0.2627	45	1.89	0.0654	0.05	-0.03301	1.0250	0.000366



From the figure, it appears that LA may be less accurate than the other 2 methods, however, removal of the covariate for accuracy (coefficients a1 and a2) from the model has negligible effect on the fit of the model (χ^2= 753.7-753.1=0.6 on 2 d.f., p=0.74) indicating no statistical evidence of a difference in diagnostic accuracy of RF according to technique. This indicates that it is reasonable to fit a single summary ROC for RF.

10.5.4 Comparing Index Tests

For many diagnostic reviews, a key objective is to compare the diagnostic accuracy of two alternative index tests that may be used to diagnose the same condition. In this section, the focus will be on the comparison of two index tests, but the approach can be extended to allow for more than two tests.

Two approaches are generally adopted for test comparisons. The first approach utilises test accuracy data from all eligible studies that have evaluated one or both tests. The second approach restricts the analysis to studies that have evaluated both tests either in the same individuals, or have randomized individuals to undergo one or other of the two tests. The second approach has advantages because the comparison is less likely to be biased due to confounding and hence these results should be relied upon where possible. However, the number of studies that report such direct comparisons is often very limited, which means that such an analysis may not be feasible or might only be considered as a sensitivity analysis (see 10.6.1)

10.5.4.1 Test comparisons based on all available studies

Often, many of the available studies evaluate only one of the tests of interest. By using all studies that have evaluated at least one of the tests, we maximize the number of studies in the analysis. However, the studies are likely to be heterogeneous in terms of design and patient characteristics that are associated with test accuracy and hence confounding may be an issue. In preliminary exploratory analyses in RevMan this can be dealt with by comparing the tests within subgroups of studies that are homogeneous with respect to important potential confounders such as study design or spectrum of disease. The value and feasibility of such exploratory analyses will be affected by the number of available studies, and missing or inconsistent reporting across studies of information on potential confounders.

The statistical methods described in this section follow directly from the earlier description of hierarchical models and how they can be used to investigate heterogeneity in test accuracy. For the comparison of two index tests, the type of test is represented by a binary covariate that is used to identify the test that gave rise to each 2×2 table included in the analysis. Confounders can potentially be adjusted for, however this is often difficult to do in practice because the number of studies is small and/or data on important confounders may be poorly recorded or incomplete.

Both the Bivariate model and the Rutter and Gatsonis HSROC model can be used to investigate the relative accuracy of two index tests. However, as noted previously, the choice of approach will be influenced by the nature of the available data, and the interpretation of the results will also depend on which approach is used.

10.5.4.2 Test comparisons using the Bivariate model

If, for each index test, the available studies have used a consistent cut-point on a continuous or ordinal scale to define test positivity then the Bivariate model provides an appropriate framework for test comparisons. It may also be reasonable to assume a consistent cut-point when a test comprises a ‘test kit’ that produces positive and negative results (such as a coloured line appearing on a device). By adopting the same strategy described earlier (10.5.3.1), a binary covariate for test type can be included in the model to investigate whether the expected sensitivity and/or specificity differs between the tests.

Care must be taken with the interpretation of the results of such a model, particularly if the common cut-point for test positivity for either test is applied to a continuous or ordinal scale. Any inferences made about the relative diagnostic accuracy of the two tests is only valid at the chosen cut-point for each of the two tests and cannot be extrapolated to other possible cut-points. Where other cut-points are reported, the analysis can be repeated using the available data to investigate the relative diagnostic accuracy of the tests at those alternative cut-points.

Because we are analyzing test accuracy data for two alternative index tests, it may not be reasonable to assume that the variances of the random effects for logit(sensitivity) and logit(specificity) are the same for the two tests. The Bivariate model can be extended to allow the variance of the random effects for both to depend on the covariate for test type. This will also affect the estimated correlation between them. Statistically, estimation of the variances of the random effects for logit(sensitivity) and logit(specificity) and correlation between them is subject to a higher level of uncertainty than for the main parameters of interest. However, if based on preliminary plots of the study level estimates of sensitivity and specificity in ROC space there are marked differences in heterogeneity between studies for the two tests, it is advisable to assess whether the assumption of equal variances of random effects for the two tests is reasonable. This is usually done by comparing the fit of the alternative models (variances do or do not depend on the covariate for test type) using a likelihood ratio test. A comparison of the main estimates of interest between the alternative models is also useful to assess whether conclusions about the relative sensitivity and/or specificity of the tests are robust to assumptions about the variances of the random effects. Again, such an investigation will not be feasible if the number of studies is small.

It is usual for most of the studies in this approach to the analysis of test comparisons to have evaluated only one of the tests, but some studies will have evaluated both. If the proportion of studies that have evaluated both is very small, then treating the results of the two tests in a study as if they were obtained from different studies is unlikely to affect the results. Although this is often done in practice, such an approach is not recommended if the proportion of studies evaluating both tests is not small because it is likely to result in inappropriate standard errors for the test comparison parameters for sensitivity and specificity. In that case the paired sensitivity/specificity data for both tests from each study should be at level one of the analysis, and a binary covariate for test type included to identify which 2×2 table corresponds to each test.

10.5.4.3 Example 3: CT versus MRI for the diagnosis of coronary artery disease

Schuetz et al (Schuetz 2010) evaluated the diagnostic performance of multislice computed tomography (CT) and magnetic resonance imaging (MRI) for the diagnosis of coronary artery disease (CAD). Prospective studies that evaluated either CT or MRI (or both), used conventional coronary angiography (CAG) as the reference standard, and used the same threshold for clinically significant coronary artery stenosis (a diameter reduction of 50% or greater) were included in the review. A total of 103 studies provided a 2x2 table for one or both tests and were included in the meta-analysis: 84 studies evaluated only CT, 14 evaluated only MRI, and 5 studies evaluated both CT and MRI. (See Appendix for data and SAS programs).

Because the studies were selected based on a common threshold for clinically significant coronary artery stenosis, the Bivariate model was used for data synthesis and test comparison. In the first stage of the analysis, we base our test comparison on all studies that evaluated at least one test. The approach follows closely the method illustrated in Example 1 for exploring heterogeneity using the Bivariate model.

A binary covariate (testtype) is added to the model which is coded as 0 if the 2x2 table is for MRI (the referent category), and coded as 1 if the 2x2 table is for CT. The five studies that evaluated both tests contribute a 2x2 table for each test, hence there are 19 studies included for MRI and 89 studies included for CT. Allowing test performance to vary by type of test resulted in a -2Log likelihood of 953.0, a reduction of 42.5 compared with the model that contained no covariates. Hence, there is statistical evidence (chi-square=42.5, 2df, P<0.001) that sensitivity and/or specificity are associated with test type. Removing the covariate for sensitivity from the model (chi-square=976.7-953.0=23.7, 1df, P <0.001) shows strong statistical evidence of a difference in sensitivity between the two tests. Similarly, removing the covariate for specificity from the model (chi-square=976.2-953.0=23.2, 1df, P <0.001) shows strong statistical evidence of a difference in specificity between the two tests.

The SAS output for the model that allows both sensitivity and specificity to vary by test is:

Fit Statistics

-2 Log Likelihood	953.0
AIC (smaller is better)	967.0
AICC (smaller is better)	967.5
BIC (smaller is better)	985.5

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
msens	2.1771	0.2457	101	8.86	<.0001	0.05	1.6896	2.6645	0.000046
mspec	0.8754	0.2111	101	4.15	<.0001	0.05	0.4566	1.2942	-0.00008
s2usens	0.8749	0.2293	101	3.82	0.0002	0.05	0.4201	1.3297	0.000033
s2uspec	0.8447	0.1696	101	4.98	<.0001	0.05	0.5084	1.1810	-4.31E-6
covsesp	0.1803	0.1384	101	1.30	0.1956	0.05	-0.09424	0.4548	-0.00002
se_CT	1.3033	0.2625	101	4.97	<.0001	0.05	0.7827	1.8240	0.000053
sp_CT	1.0415	0.2154	101	4.84	<.0001	0.05	0.6143	1.4687	-0.00005

Covariance Matrix of Parameter Estimates

Row	Parameter	msens	mspec	s2usens	s2uspec	covsesp	se_CT	sp_CT
1	msens	0.06038	0.005241	0.01100	0.000342	0.002242	-0.05262	-0.00404
2	mspec	0.005241	0.04457	0.000518	0.003360	0.001651	-0.00376	-0.03861
3	s2usens	0.01100	0.000518	0.05257	0.000694	0.007608	0.003839	-0.00060
4	s2uspec	0.000342	0.003360	0.000694	0.02875	0.005537	-0.00023	0.000438
5	covsesp	0.002242	0.001651	0.007608	0.005537	0.01915	0.001155	-0.00110
6	se_CT	-0.05262	-0.00376	0.003839	-0.00023	0.001155	0.06889	0.004479
7	sp_CT	-0.00404	-0.03861	-0.00060	0.000438	-0.00110	0.004479	0.04638

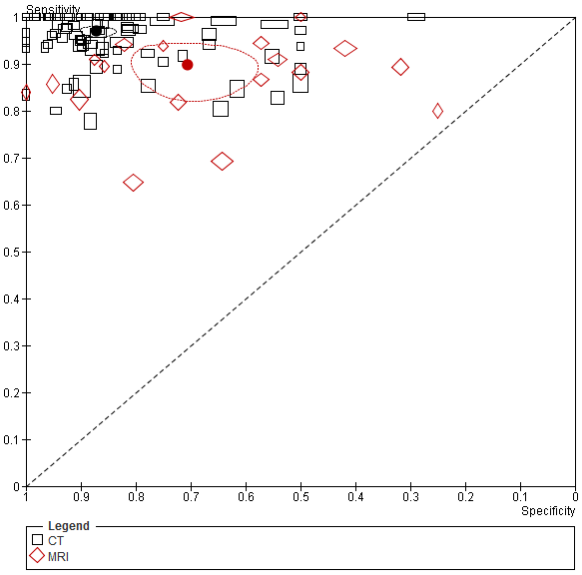
Additional Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
logitsens CT	3.4804	0.1550	101	22.45	<.0001	0.05	3.1729	3.7879
logitspec CT	1.9169	0.1172	101	16.36	<.0001	0.05	1.6844	2.1494

Covariance Matrix of Additional Estimates

Row	Label	Cov1	Cov2
1	logitsens CT	0.02403	0.001916
2	logitspec CT	0.001916	0.01373

The estimated logit(sensitivity) and logit(specificity) for the referent category (MRI) are given by msens and mspec respectively. These estimates, their standard errors and their covariance are shown in the red boxes. The ESTIMATE command in SAS has been used to obtain the corresponding estimates for CT (shown in the blue boxes). The variances of the random effects and their covariance are shown in the green box.



The above estimates can be input to RevMan to produce a ROC scatter plot with summary operating points for MRI and CT and their confidence regions superimposed as shown in the figure where the black symbols represent CT and the red symbol represent MRI. Because of the large number of studies for CT, the summary estimate and region are difficult to see. The figure could be redrawn with just the summary points and regions and shown separately from the ROC scatter plot.

From the output above, the 95% confidence limits for the summary estimates for logit(sensitivity) and logit(specificity) can be found in the columns headed “lower” and “upper”. Using inverse transformation, the summary estimates for sensitivity are 0.90 (95%CI 0.84, 0.93) for MRI and 0.97 (95% CI 0.96, 0.98) for CT . The summary estimates for specificity 0.71 (95%CI 0.61, 0.78) for MRI and 0.87 (95%CI 0.84, 0.90) for CT. Hence, based on this analysis, there is strong evidence that CT has higher sensitivity and specificity than MRI, for detecting clinically significant coronary artery stenosis defined as a diameter reduction of 50% or more.

10.5.4.4 Test comparisons using the Rutter and Gatsonis HSROC model

Simple separate comparisons of summary estimates of sensitivity (or specificity) of alternative tests can be misleading if the included studies have used different cut-points to define test positivity. In this situation, comparisons based on SROC curves provide a more informative approach.

The hierarchical modelling strategy used to investigate heterogeneity described earlier for the Rutter and Gatsonis HSROC methods (10.5.3.3) can be used for comparisons of test accuracy when there is variability in threshold between studies. The type of test is represented by a binary covariate that is used to identify the test that gave rise to each 2×2 table included in the analysis. This covariate then allows the reviewer to investigate whether test type is associated with the shape and position of the summary ROC curve. Interpretation of the results follows directly from the discussion of the interpretation of investigations of heterogeneity in 10.5.3.

Statistically, estimation of the variances of the random effects for threshold and accuracy is subject to a higher level of uncertainty than for the main model parameters of interest. If preliminary plots of the study level estimates of sensitivity and specificity in ROC space show marked differences in

heterogeneity between studies for the two tests, it is advisable to assess whether the assumption of equal variances of the random effects for the two tests is reasonable. This is usually done by comparing the fit of the alternative models (i.e. where variances do, or do not, depend on the covariate for test type). A comparison of the main estimates of interest between the alternative models is also useful to assess whether conclusions about the relative shape and accuracy of the summary curves for the two tests are robust to assumptions about the variances of the random effects. Again, such an investigation will not be feasible if the number of studies is small.

As noted for the Bivariate model, it is usual for most of the included studies to have evaluated only one of the tests, but some studies will have evaluated both. If the proportion of studies that have evaluated both is very small, then treating the results of the two tests in a study as if they were obtained from different studies is unlikely to affect the results. However, more accurate standard errors will be obtained for the test comparison parameters if the data for both tests are modelled within the study at level one of the analysis. A binary covariate for test type must be included to identify which 2×2 table corresponds to each test.

10.5.4.5 Test comparison based on studies that directly compare tests

As noted earlier, heterogeneity in the estimated accuracy of a diagnostic test across studies is likely to occur. This could confound the comparison of two tests if different studies are used to estimate the diagnostic accuracy of each test. Ideally, the comparison should be based on studies that have made a direct comparison of the tests of interest by either applying both tests to each individual, or by randomizing each individual to receive one of the tests. A common reference standard should be applied to both tests. If there are sufficient studies of this type on which to base a test comparison, the results are less prone to bias than an analysis based on all available studies that have evaluated one or both tests.

A preliminary graphical analysis can be conducted in RevMan by plotting the estimated sensitivity and specificity for both tests, for each study in ROC space. The two points contributed by each study (one for each test) are joined by a line to highlight the relative test accuracy within each study (see 10.3.2). This figure illustrates the pairing of test accuracy estimates at the study level.

The rationale described above for choosing between the Bivariate model and the HSROC model when making test comparisons is also applicable here, and the same points relating to interpretation apply. The only major difference is that the analysis does not include any studies that have evaluated only one of the tests.

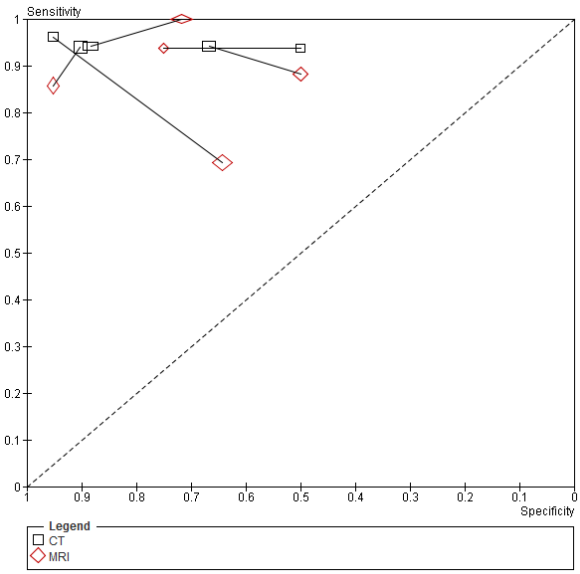
Because each study contributes a 2×2 table for each of the two tests to be compared, the data for the two tests must be analysed within study at level one of the analysis, and a binary covariate for test type included to identify which 2×2 table corresponds to each test. Entering a separate 2×2 table for each test (within each study) for analysis in a hierarchical model effectively assumes that the data arise from a randomized design. This represents a conservative approach that is often necessitated by the lack of information on paired results at the individual level for truly 'paired' studies that have applied both tests to the same individual.

Meta-analytical models that account for pairing of test results within an individual in studies which have used a paired study design are not commonly used, and require further development and testing before they are implemented in a Cochrane review. Such an extension would also require

that researchers publish a cross classification of test results within both the diseased and non-diseased groups. This is not common practice at present.

10.5.4.6 Example 3 (cont.): CT versus MRI for the diagnosis of coronary artery disease

The meta-analysis by Schuetz also included 5 studies that made a direct comparison of CT and MRI. Basing the analysis on these 5 studies (ten 2x2 tables) has the advantage that the results should be less prone to bias. However, the number of studies in the analysis is dramatically reduced which reduces the precision of the summary estimates. As we will see in this example, simplifying assumptions may also be required to fit complex Hierarchical models to these data. We will again apply the Bivariate model for these data.



The ROC scatter plot shows the data for the 5 paired studies, with black used to denote CT and red used to denote MRI. A line is used to join the results for CT and MRI within each study. Examining this plot, we can see that sensitivity for CT is lower than for MRI in one study, equivalent in one study, and higher in the other three studies. Specificity is higher for CT than for MRI in 3 studies and lower in the other 2.

Fitting a model to these data is difficult, particularly for the Bivariate model where convergence is more problematic than for the Rutter and Gatsonis model (see 10.5.6).

A preliminary series of models were fitted to assess whether random effects should be included for both sensitivity and specificity (this model did not include the covariate for test type). The model that included random effects only for specificity gave a -2Log likelihood of 106.4, a better fit than the model that assumed a fixed effect for both sensitivity and specificity (-2Log likelihood 114.8). The model that assumed random effects only for sensitivity provided no improvement to the fit compared with the fixed effect model. Hence, the covariate for test type was added to the model with random effects for specificity only. The SAS output for this model is shown below.

Fit Statistics

-2 Log Likelihood	89.6
AIC (smaller is better)	99.6
AICC (smaller is better)	103.9
BIC (smaller is better)	97.6

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
msens	1.8083	0.2412	4	7.50	0.0017	0.05	1.1385	2.4781	0.000011
mspec	0.8910	0.3606	4	2.47	0.0689	0.05	-0.1101	1.8921	1.197E-6
s2uspec	0.4239	0.3744	4	1.13	0.3208	0.05	-0.6156	1.4634	-1.35E-7
se_CT	1.0051	0.4195	4	2.40	0.0747	0.05	-0.1596	2.1698	-6.32E-6
sp_CT	0.9378	0.2955	4	3.17	0.0337	0.05	0.1175	1.7581	3.672E-6

Covariance Matrix of Parameter Estimates

Row	Parameter	msens	mspec	s2uspec	se_CT	sp_CT
1	msens	0.05820	-7E-12	-131E-13	-0.05820	-147E-14
2	mspec	-7E-12	0.1300	-0.01214	1.13E-11	-0.03387
3	s2uspec	-131E-13	-0.01214	0.1402	2.43E-11	0.003983
4	se_CT	-0.05820	1.13E-11	2.43E-11	0.1760	9.95E-12
5	sp_CT	-147E-14	-0.03387	0.003983	9.95E-12	0.08729

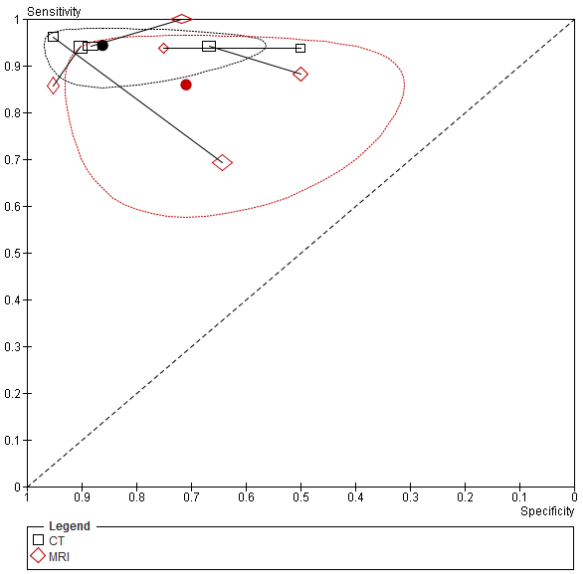
Additional Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
logitsens CT	2.8134	0.3432	4	8.20	0.0012	0.05	1.8606	3.7663
logitspec CT	1.8287	0.3867	4	4.73	0.0091	0.05	0.7550	2.9025

Covariance Matrix of Additional Estimates

Row	Label	Cov1	Cov2
1	logitsens CT	0.1178	1.28E-11
2	logitspec CT	1.28E-11	0.1496

The layout and interpretation of the output follow that for the indirect test comparison example discussed earlier, with the red indicating MRI results and blue CT. Note that the estimated covariance between msens and mspec is equivalent to zero. The green box shows the estimated variance of the random effects for specificity. These estimates were input to RevMan to superimpose summary estimates and their confidence regions on the ROC scatterplot. A zero value was entered for the variance of the random effects for sensitivity.



Inverse logit transformation of the estimates and their lower and upper 95% confidence limits gives estimated sensitivities of 0.86 (95%CI 0.76, 0.92) for MRI and 0.94 (95% CI 0.87, 0.98) for CT; and estimated specificities of 0.71 (95%CI 0.47, 0.87) for MRI and 0.86 (95%CI 0.68, 0.95) for CT. These estimates are consistent with the previous analysis which showed that CT had higher sensitivity and specificity than MRI.

The confidence regions shown on the figure are wider than would be indicated by the confidence intervals given above. The confidence regions computed by RevMan appear to be overly conservative when the number of studies is small. This issue will investigated and modifications made to later versions of the software if required.

The t-statistics in the output above provide only weak evidence of a difference in sensitivity (P=0.075 for parameter se_CT) and evidence of a difference in specificity (P=0.034 for parameter sp_CT). The P-values based on changes in the -2Log likelihood are lower (0.012 and 0.0011 respectively), indicating stronger evidence for these effects. Given the small number of studies in this analysis and the resulting difficulty in checking model assumptions regarding the distributions of the random effects, it may be advisable to take a conservative approach. The key inference here is that the results of this analysis are consistent with the conclusions of the earlier indirect comparison which was based on all available studies that evaluated at least one of the index tests.

10.5.5 Computer software

Both of the hierarchical models we have focused on can be fitted using a range of statistical packages. WinBUGS (or its recent open-source version OpenBUGS) provides a flexible Bayesian framework for model fitting. It can be used to fit both the Bivariate and HSROC models. To obtain the parameters needed to create confidence and prediction regions for the RevMan SROC plots requires estimates of the standard errors, which need to be generated from the posterior distributions.

Many analysts find fitting the models using standard commonly used software packages such as SAS, Stata and MLwiN more straightforward.

The Bivariate model can be fitted using software that can fit a generalized linear mixed model. Commonly used routines are Proc NLMIXED (or Proc GLIMMIX) in SAS, `xtmelogit` (or the user written package `glamm`) in Stata. All of these programs assume that the random effects are normally distributed. WinBUGS can allow alternative distributions for the random effects if that is deemed necessary. A user written command `metandi` is available in Stata to fit a model without covariates (Harbord 2009), and a macro METADAS written in SAS which includes models with and without covariates (Takwoingi 2008). Both macros neatly tabulate output required for populating the plotting functions in RevMan.

The parameterization for the Rutter and Gatsonis model represents a generalized *non*- linear mixed model. If covariates are to be included and tested in the model, then the range of available software is more limited because of the non-linear form of the model if the shape parameter is included in the model. This model is usually fitted using Proc NLMIXED in SAS (Macaskill 2004). The SAS METADAS macro also fits HSROC models with and without covariates. The Stata `metandi` command can be used to fit the Rutter and Gatsonis HSROC model without covariates. However, it should be noted that it does this by exploiting the mathematical equivalence between the Bivariate and HSROC models when there are no covariates in the model. Hence, it is applicable when an overall HSROC curve is to be fitted to a group of studies but does not allow inclusion of covariates.

RevMan can use the parameter estimates from one or other model to estimate: a summary curve; summary operating point; a confidence region and prediction region for the summary point. However, the review author needs to be clear which of these summary measures are appropriate for their analysis.

10.5.6 Approaches to analysis with small numbers of studies

When the number of studies is small it may be difficult to decide on which terms should be included in a model, and which is the 'best' model. For instance, when fitting a summary ROC curve, the uncertainty associated with the estimation of the shape parameter could be very high, and the estimate may also be strongly influenced by the inclusion/exclusion of individual studies. For both the Bivariate and HSROC models, estimates of the variances of the random effects will be subject to a high level of uncertainty.

It is important to keep in mind that estimation of a single summary point using the Bivariate model, or estimation of a single summary curve using the HSROC model, requires five parameters to be estimated in the full model specification. There is little information on which to base these estimates when the number of studies is small, so analysts must take this into account when interpreting the results. In some situations models may fail to converge.

It is not possible to give hard and fast rules about how to proceed when dealing with small numbers of studies. However, some strategies are outlined here which may help in some situations. Ultimately, judgement must be exercised regarding whether a model is sufficiently reliable to report.

Failure of a model to converge may be symptomatic of several problems:

- In some cases, it may be due to poor choices of starting values for the parameter estimates. If so, it may help to fit the model first assuming a fixed effect for the model parameters, and then use these as the starting values for the random effects model.

- For small data sets, convergence may also be affected by the inclusion/removal of individual studies. The effect of such influential studies should be investigated.

- Convergence problems can also arise when the variance of one of the random effects is close to zero. This is particularly an issue for the Bivariate model parameterisation, where an examination of the scatter plot may help to identify strong heterogeneity in sensitivity but homogeneity in specificity, or vice versa. Restricting the model to have random effects for one parameter, and a fixed effect for the other may then be warranted. This particular problem can also occur when the number of studies is relatively large.

- The standard error for the shape parameter in the HSROC model may be large. It would be advisable to check how much the shape is influenced by the removal of individual studies. When the shape is uncertain and also very dependent on individual studies, then some analysts may choose to assume symmetry for the summary curve to acknowledge that the shape cannot be estimated reliably. Again, this needs to be reported and discussed in the report of the analyses.

10.6 Special topics

10.6.1 Sensitivity analysis

The process of undertaking a systematic review involves a sequence of decisions. Whilst many of these decisions are clearly objective and non-contentious, some will be somewhat arbitrary or unclear. For instance, if inclusion criteria involve a numerical value, the choice of value is usually arbitrary: for example, defining groups of older people may reasonably have lower limits of 60, 65, 70 or 75 years, or any value in between. Other decisions may be unclear because a study fails to include the required information. Further decisions are unclear because there is no consensus of the best method to use for a particular problem, such as defining a reference standard or analysing missing data or intermediate test results.

It is desirable to demonstrate that the findings from a systematic review are not dependent on such arbitrary or unclear decisions. A sensitivity analysis is a repeat of the primary analysis or meta-analysis, substituting alternative decisions of ranges of values for decisions that were arbitrary or unclear. For example, if the eligibility of some studies in the meta-analysis is dubious because they do not contain full details, sensitivity analysis may involve undertaking the meta-analysis twice: first including all studies, and second, only including those that are definitely known to be eligible. A sensitivity analysis asks the question “Are the findings robust to the decisions made in the process of obtaining and analysing them?”. A sensitivity analysis is not the same as a subgroup analysis where the purpose is to investigate how study design and patient characteristics are associated with test accuracy. The aim in the subgroup analysis is to explore and explain heterogeneity in test accuracy.

There are many decision nodes within the systematic review process which can generate a need for sensitivity analysis. Examples include:

Searching for studies:

- Should abstracts whose results cannot be confirmed in subsequent publications be included in the review?

Eligibility criteria

- Characteristics of participants: where a majority but not all people in a study meet the required presentation or demographic, should the study be included?
- Characteristics of tests: what versions of a test technology should be included? What threshold definition constitutes a common threshold?
- Characteristics of the reference standard: where there are variations on information used in a clinical opinion based reference standard, should they all be included? Where the reference standard involves follow-up, what lengths of follow-up are considered adequate?
- Study methods: should only fully uniformly verified studies be included? Should unblinded studies be included? Should case-control studies be included? Or should inclusion be restricted by any other methodological criteria?

What data should be analysed?

- How should uninterpretable test results be handled in the analysis? Should they be classified as test negatives or excluded?
- How should missing data be handled in the analysis?

Analysis methods

- Should a common or symmetric shape for an ROC curve be presumed across subgroups or tests?
- Can equal variances be presumed for all tests in a comparison?

Reporting of sensitivity analyses in a systematic way may best be done by producing a summary table. Some sensitivity analyses can be pre-specified in the study protocol, but many issues suitable for sensitivity analysis are only identified during the review process where the individual peculiarities of the studies under investigation are identified. Where sensitivity analysis show the overall result and conclusions are not affected by the different decisions made during the review process, the results of the review can be regarded with a higher degree of certainty. Where sensitivity analyses identify particular decisions or missing information that greatly influence the findings of the review, greater resources can be deployed to try and resolve uncertainties and obtain extra information, possibly through contacting study authors. If this cannot be achieved, the results must be interpreted with an appropriate degree of caution. Such findings may generate proposals for further investigations and future research.

Sensitivity analysis may be sometimes confused with subgroup analysis. Although some sensitivity analysis may involve restricting the analysis to a subset of the totality of the studies, the two methods differ in two ways. First sensitivity analyses do not attempt to estimate the effect of the covariate in the group of studies removed from the analysis, whereas in subgroup analysis estimates are produced for all groups. Second, in sensitivity analysis informal comparisons are made between

different ways of estimating the same thing, whereas in subgroup analysis formal statistical comparisons are made across the subgroups.

10.6.2 Investigating and handling verification bias.

An examination of the potential for verification bias in a systematic review would ordinarily be part of the assessment of study quality. If verification bias is present, corrections would need to be implemented within individual studies, before proceeding to the meta-analysis. The literature on methods for correcting verification bias in individual studies is by now extensive. For example, the analyst may want to consult Chapter 10 in Zhou et al (Zhou 2002) and Chapter 7 in Pepe (Pepe 2003). It may also be useful to investigate the presence of verification bias as a source of heterogeneity among studies. Issues arise in the extraction of study data when adjustments have been made for verification bias, as described in the Chapter 8.

10.6.3 Investigating and handling publication bias

Systematic reviewers must undertake comprehensive searches to attempt to locate all relevant studies. If the studies included in the review have results that differ systematically from relevant studies that are missed, estimates derived from the meta-analysis will be affected by publication bias (Begg 1994a).

Although there is substantial literature relating to publication bias in systematic reviews of randomized controlled trials, little research has been done in the context of systematic reviews of diagnostic studies. However, it is clear that the determinants of publication bias for reviews of RCTs (Dickersin 1990), (Ioannidis 1998) are unlikely to be generalizable to reviews of diagnostic studies. For instance, when considering diagnostic test accuracy, statistical significance is not particularly relevant as few studies formulate and test hypotheses. Another difference is the likely relationship between study size and methodological quality. Whereas large RCTs require large-scale funding and are on average conducted and analysed with greater methodological rigour than small RCTs, large diagnostic studies may be no more than an analysis of a large laboratory database of routinely-collected data.

Statistical tests detect funnel plot asymmetry in general rather than publication bias specifically (see section 10.4 of the Cochrane Handbook for Systematic Reviews of Interventions). Tests for funnel plot asymmetry designed primarily for use in randomized trials, including the Egger (Egger 1997), Begg (Begg 1994b), Harbord (Harbord 2006) and Peters (Peters 2006) tests, should *not* be used with diagnostic studies. It is well established that the accuracy of such tests for funnel plot asymmetry is reasonable if the odds ratio is close to 1 (as occurs in many randomized trials), but deteriorates as the odds ratio moves away from 1 (Macaskill 2001), (Schwarzer 2002). For diagnostic studies, the odds ratio is expected to be large. Applying such tests for funnel plot asymmetry in systematic reviews of diagnostic test accuracy is likely to result in publication bias being incorrectly indicated by the test far too often (a Type I error rate that is too high) (Deeks 2005).

A more appropriate method for detecting funnel plot asymmetry in reviews of diagnostic studies has been developed (Deeks 2005). It tests for association between the lnDOR and the 'effective sample size', a simple function of the number of diseased and non-diseased individuals. A simulation study has shown that the test has modest power for detecting funnel plot asymmetry. However, when there is heterogeneity in the DOR, even this test has low power, as do all tests for funnel plot asymmetry.

Since heterogeneity in test accuracy is to be expected in many diagnostic reviews, review authors are warned against interpreting statistical evidence of funnel plot asymmetry as necessarily implying publication bias. Study size may be related to test accuracy for reasons other than publication bias. Exploration of heterogeneity in test accuracy should be undertaken, as patient and study characteristics may be associated with study size as well as test accuracy (Deeks 2005). Further research is required to improve our understanding of the determinants and extent of publication bias for diagnostic studies.

10.6.4 Developments in meta-analysis for DTA reviews

This chapter reflects the currently established methods for meta-analysis of diagnostic test accuracy. Methodological developments occur often in this field, and the methods used in Cochrane DTA reviews are sure to develop over time to extend the scope of the models and data structures which can be included. As new methods are shown to be robust and of importance, and software made available for their implementation, they will be included in updates of this chapter.

Of particular interest are analytical methods being developed to include data from multiple thresholds for each study, which allow both more accurate estimation of summary ROC curves and estimates of average sensitivity and specificity values at stated thresholds, but these require further evaluation before they will be incorporated in Cochrane reviews (Dukic 2003), (Hamza 2009).

Appendix

The programs listed below are in SAS, but the results can be reproduced using other software. The export facility in RevMan5 was used to create a .csv file which contained the 2x2 tables for each study included in each review. The .csv file can be read by Excel and can also be imported into statistical programs such as SAS for further analysis. (Note: additional columns of data that are not relevant to our analyses are not shown). Files are available from the Cochrane DTA website (srdta.cochrane.org).

Data and SAS file for Example 1- Anti-CCP for the diagnosis of rheumatoid arthritis

Data (nashimura CCP.csv)

Test	study_id	CCP generation	tp	fp	fn	tn	method of measurement
Anti-CCP	Aotsuka 2005	CCP2	115	17	16	73	
Anti-CCP	Bas 2003	CCP1	110	24	86	215	
Anti-CCP	Bizzaro 2001	CCP1	40	5	58	227	
Anti-CCP	Bombardieri 2004	CCP2	23	0	7	39	
Anti-CCP	Choi 2005	CCP2	236	20	88	231	
Anti-CCP	Correa 2004	CCP2	74	11	8	130	
Anti-CCP	De Rycke 2004	CCP2	89	4	29	142	
Anti-CCP	Dubucquoi 2004	CCP2	90	2	50	129	
Anti-CCP	Fernandez-Suarez 2005	CCP2	31	0	22	75	
Anti-CCP	Garcia-Berrocal 2005	CCP2	69	8	18	38	
Anti-CCP	Girelli 2004	CCP2	25	2	10	40	
Anti-CCP	Goldbach-Mansky 2000	CCP1	43	1	63	120	
Anti-CCP	Greiner 2005	CCP2	70	5	17	228	
Anti-CCP	Grootenboer-Mignot 2004	CCP2	167	8	98	88	
Anti-CCP	Hitchon 2004	CCP2	26	8	15	15	
Anti-CCP	Jansen 2003	CCP1	110	3	148	118	
Anti-CCP	Kamali 2005	CCP2	26	1	20	56	
Anti-CCP	Kumagai 2004	CCP2	64	14	15	293	
Anti-CCP	Kwok 2005	CCP2	71	2	58	66	
Anti-CCP	Lee and Schur 2003	CCP2	68	14	35	132	
Anti-CCP	Lopez-Hoyos 2004	CCP2	38	3	0	73	
Anti-CCP	Nell 2005	CCP2	42	2	60	96	
Anti-CCP	Nielen 2005	CCP2	149	7	109	114	
Anti-CCP	Quinn 2006	CCP2	147	10	35	106	
Anti-CCP	Rantapaa-Dahlqvist 2003	CCP2	47	7	20	375	
Anti-CCP	Raza 2005	CCP2	24	3	18	79	
Anti-CCP	Saraux 2003	CCP1	40	11	46	146	
Anti-CCP	Sauerland 2005	CCP2	171	26	60	443	
Anti-CCP	Schellekens 1998	CCP1	72	14	77	298	
Anti-CCP	Soderlin 2004	CCP2	7	2	9	51	
Anti-CCP	Suzuki 2003	CCP2	481	23	68	185	
Anti-CCP	Vallbracht 2004	CCP2	190	12	105	408	
Anti-CCP	van Gaalen 2005	CCP2	82	13	71	301	
Anti-CCP	van Venrooij 2004	CCP2	865	79	252	2218	
Anti-CCP	Vincent 2002	CCP1	139	7	101	464	
Anti-CCP	Vittecoq 2004	CCP2	69	5	107	133	
Anti-CCP	Zeng 2003	CCP1	90	7	101	313	

SAS Program (nishimura CCP.sas):

```
/* Import data */
proc import out=nishimura
   datafile='C:\chapter10\nishimura CCP.csv'
         dbms=csv replace;
   getnames=yes;
run;

data nishimura_accp;
         set  nishimura;
         where test='Anti-CCP';
run;


/* Create a two separate records for the true results in each study,
   the first for the diseased group, and the second for the non-diseased group.
   The variable sens is an indicator which takes the value 1 if true=true positives and 0 otherwise,
   the variable spec is also an indicator that takes the value 1 if true =true negatives and 0 otherwise */
data nishimura_accp;
         set  nishimura_accp;
         sens=1; spec=0; true=tp; n=tp+fn; output;
         sens=0; spec=1; true=tn; n=tn+fp; output;
run;


/* Ensure that both records for a study are clustered together */
proc sort data=nishimura_accp;
by study_id ;
run;

/* Run the Bivariate model with no covariates
   The "cov" option requests that a covariance matrix is printed for
   all model parameter estimates. The "ecov" option requests a covariance matrix
   for all additional estimates that are computed. */

proc nlmixed data=nishimura_accp cov ecov ;

/* specify starting values for all parameters to be estimated
   and ensure that the variances of the random effects cannot be negative */

parms msens=1 to 2 by 0.5 mspec=2 to 4 by 0.5 s2usens=0.2 s2uspec=0.6 covsesp=0;
         bounds s2usens>=0;
         bounds s2uspec>=0;
         logitp = (msens + usens)*sens + (mspec + uspec)*spec;
    p = exp(logitp)/(1+exp(logitp));
    model true ~ binomial(n,p);

   /* usens and uspec represent the random effects. The are both assumed to be
      normally distributed with mean zero. Their variances estimates are s2usens and s2uspec,
      and their covariance estimate is  covesp */

         random usens uspec ~ normal([0 , 0],[s2usens,covsesp,s2uspec])
                           subject=study_id out=randeffs;

/* Additional estimates that are functions of the model parameters can be estimted here:
         e.g the positive and negative likelihood ratios */
estimate 'logLR+' log((exp(msens)/(1+exp(msens)))/(1-(exp(mspec)/(1+exp(mspec)))));
estimate 'logLR-' log((1-(exp(msens)/(1+exp(msens))))/(exp(mspec)/(1+exp(mspec))));

run;


/* Check assumption of normality for the random effects */
proc univariate data=randeffs plot normal;
  class effect;
  var estimate;
run;

/* Create a dummy variable for CCP generation, coded as 0 for 'CCP1' (the referent generation)
   and coded as 1 for 'CCP2'. This new variable is added to the dataset set created above. */

data nishimura_accp;
         set nishimura_accp;
         ccpg=0;
         if ccp_generation ="CCP2"  then ccpg=1;
run;


/* add the covariate CCPG to the model to allow both sensitivity and specificity to be
   associated with generation of the test */

proc nlmixed data=nishimura_accp cov ecov;
parms msens=1 mspec=2 s2usens=0.2 s2uspec=0.6 covsesp=0 se2=0 sp2=0;
```

```
   bounds s2usens>=0;
   bounds s2uspec>=0;
   logitp=(msens+usens+se2*ccpg)*sens+(mspec+uspec+sp2*ccpg)*spec;
   p = exp(logitp)/(1+exp(logitp));
   model true ~ binomial(n,p);
   random usens uspec ~ normal([0 , 0],  [s2usens,covsesp,s2uspec])
   subject=study_id out=randeffs;

/* Estimate logit(sensitivity) and logit(specificity) for CCP2
   (their correlation will be output because of the "ecov" option for nlmixed),
   and also log likelihood ratios for CCP1 and CCP2 */

estimate 'logitsens CCP2'  msens + se2;
estimate 'logitspec CCP2'  mspec + sp2;
estimate 'logLR+ CCP1' log((exp(msens)/(1+exp(msens)))/(1-(exp(mspec)/(1+exp(mspec)))));
estimate 'logLR- CCP1' log((1-(exp(msens)/(1+exp(msens))))/(exp(mspec)/(1+exp(mspec))));
estimate 'logLR+ CCP2' log((exp(msens+se2)/(1+exp(msens+se2)))/(1-(exp(mspec+sp2)/(1+exp(mspec+sp2)))));
estimate 'logLR- CCP2' log((1-(exp(msens+se2)/(1+exp(msens+se2))))/(exp(mspec+sp2)/(1+exp(mspec+sp2))));
run;


/* Check assumption of normality for the random effects */
proc univariate data=randeffs plot normal;
  class effect;
  var estimate;
run;
```

Data and SAS file for Example 2 - Rheumatoid Factor as a marker for Rheumatoid Arthritis.

Data (nishimura RF.csv)

test	study_id	CCP generation	tp	fp	fn	tn	method of measurement
RF	Young 1991		25	1	14	20	Rheumatoid arthritis hemagglutination
RF	Nell 2005		56	11	46	87	Not reported
RF	Quinn 2006		115	53	67	63	Not reported
RF	Bizzaro 2001		61	36	37	196	Nephelometry
RF	Bombardieri 2004		27	6	3	33	Nephelometry
RF	Das 2004		42	46	14	127	Nephelometry
RF	Fernandez-Suarez 2005		30	2	23	73	Nephelometry
RF	Girelli 2004		32	29	3	13	Nephelometry
RF	Goldbach-Mansky 2000		70	39	36	93	Nephelometry
RF	Greiner 2005		75	42	12	191	Nephelometry
RF	Grootenboer-Mignot 2004		64	18	29	73	Nephelometry
RF	Hitchon 2004		32	10	9	13	Nephelometry
RF	Jansen 2003		130	8	128	113	Nephelometry
RF	Kwok 2005		77	16	52	52	Nephelometry
RF	Lopez-Hoyos 2004		36	3	5	70	Nephelometry
RF	Sauerland 2005		161	89	7	360	Nephelometry
RF	Spiritus 2004		57	9	33	93	Nephelometry
RF	Suzuki 2003		383	38	166	170	Nephelometry
RF	Swedler 1997		89	3	9	39	Nephelometry
RF	Aho 1999		64	16	27	153	LA
RF	Anuradha and Chopra 2005		482	2	82	153	LA
RF	Berthelot 1995		80	50	39	45	LA
RF	Choi 2005		261	54	63	197	LA
RF	Cordonnier 1996		20	2	29	18	LA
RF	De Rycke 2004		93	28	25	118	LA
RF	Despres 1994		143	39	63	130	LA
RF	Kamali 2005		20	32	26	25	LA
RF	Lee and Schur 2003		73	22	29	90	LA
RF	Raza 2005		22	2	20	80	LA
RF	Saraux 1995		8	8	31	91	LA
RF	Soderlin 2004		5	4	11	49	LA
RF	Thammanichanond 2005		57	25	6	111	LA
RF	Vittecoq 2001		26	1	32	29	LA
RF	Winkles 1989		113	19	29	481	LA
RF	Banchuin 1992		36	6	41	313	ELISA
RF	Bas 2003		143	43	53	196	ELISA
RF	Carpenter and Bartkowiak 1989		60	8	20	119	ELISA
RF	Davis and Stein 1989		18	3	31	25	ELISA
RF	de Bois 1996		8	8	0	31	ELISA
RF	Dubucquoi 2004		84	41	56	90	ELISA
RF	Gomes-Daudrix 1994		48	1	40	99	ELISA
RF	Jonsson 1998		50	14	20	191	ELISA
RF	Rantapaa-Dahlqvist 2003		49	23	28	359	ELISA
RF	Saraux 2003		35	8	51	149	ELISA
RF	Schellekens 2000		80	28	69	284	ELISA
RF	Vallbracht 2004		196	75	99	345	ELISA
RF	van Leeuwen 1988		163	10	28	140	ELISA
RF	Vasiliauskiene 2001		75	21	21	106	ELISA
RF	Visser 1996		157	287	78	1466	ELISA
RF	Vittecoq 2004		62	11	114	127	ELISA

SAS Program (nishimura RF.sas):

```
/* Import data */
proc import out=nishimura
   datafile='C:\chapter10\nishimura RF.csv'
   dbms=csv
         replace;
   getnames=yes;
run;

/* select only studies that have evaluated RF */
data nishimura_RF;
         set nishimura;
         where test='RF';
run;

proc print;
run;


data nishimura_RF;
         set  nishimura_RF;

         /* Create separate records for the diseased and non-diseased groups in each study
         The variable dis is the disease indicator which takes the value 0.5 if diseased
         and -0.5 if not diseased. */
         dis=0.5; pos=tp; n=tp+fn; output;
         dis=-0.5; pos=fp; n=tn+fp; output;
run;

/* Ensure that both records for a study are clustered together  */
proc sort data=nishimura_RF;
         by study_id dis;
run;

/* Run the Rutter and Gatsonis HSROC model with no covariates.
   request covariance matrices for model parameters ("cov") and
   also for additional estimates that are computed ("ecov") */

proc nlmixed data=nishimura_RF ecov cov ;

   /* set starting values for all model parameters to be estimated */

         parms alpha=2 theta=0 beta=0 s2ua=0 s2ut=0 ;

         logitp = (theta + ut + (alpha + ua)*dis) * exp(-(beta)*dis);

         p = exp(logitp)/(1+exp(logitp));

         model pos ~ binomial(n,p);

         /* the random effects for accuracy (ua) and threshold (ut) are assumed to be
            approximately normally distributed, both with mean zero and with variances
            s2ua and s2ut respectively. The covariance of the random effects is set to 0. */

         random ut ua ~ normal([0,0],[s2ut,0,s2ua]) subject=study_id out=randeffs;

run;

         /* Create two dummy variables for to allow for the three RF measurement methods.
       LA is the referent method
       Delete the 2 studies that did not report the method, and the study that used
       a different method.  */

data nishimura_RF;
         set  nishimura_RF;
         if method_of_measurement ne "ELISA" and method_of_measurement ne "Nephelometry" and
            method_of_measurement ne "LA" then delete;
         rfm1=0;  rfm2=0; ;
         if method_of_measurement ="ELISA"  then rfm1=1;
         if method_of_measurement ="Nephelometry"  then rfm2=1;
run;

/* Ensure that both records for a study are clustered together */
proc sort data=nishimura_RF;
         by study_id dis;
run;

/* include covariates to allow accuracy, threshold and shape to vary by method */

proc nlmixed data=nishimura_RF ecov cov ;

         parms alpha=2 theta=0 beta=0 s2ua=1 s2ut=1 a1=0 a2=0 t1=0 t2=0 b1=0 b2=0 ;
```

```
          logitp = (theta + ut + t1*rfm1 + t2*rfm2 + (alpha + ua + a1*rfm1 + a2*rfm2)*dis)*
             exp(-(beta + b1*rfm1 + b2*rfm2)*dis);

          p = exp(logitp)/(1+exp(logitp));

          model pos ~ binomial(n,p);

          random ut ua ~ normal([0,0],[s2ut,0,s2ua]) subject=study_id out=randeffs;

          /* parameter estimates for the methods of RF measurement; */
          estimate 'alpha ELISA'  alpha + a1;
          estimate 'theta ELISA'  theta + t1;
          estimate 'beta ELISA'       beta + b1;
          estimate 'alpha Nephelometry' alpha + a2;
          estimate 'theta Nephelometry' theta + t2;
          estimate 'beta Nephelometry' beta + b2;
run;


/* simplify the model to assume that all three curves have the same shape */
proc nlmixed data=nishimura_RF ecov cov ;

          parms alpha=2 theta=0 beta=0 s2ua=1 s2ut=1 a1=0 a2=0 t1=0 t2=0 ;

          logitp = (theta + ut + t1*rfm1 + t2*rfm2 + (alpha + ua + a1*rfm1 + a2*rfm2)*dis)*
             exp(-(beta)*dis);

          p = exp(logitp)/(1+exp(logitp));

          model pos ~ binomial(n,p);

          random ut ua ~ normal([0,0],[s2ut,0,s2ua]) subject=study_id out=randeffs;

          /* parameter estimates for the methods of RF measurement; */
          estimate 'alpha ELISA'  alpha + a1;
          estimate 'theta ELISA'  theta + t1;
          estimate 'alpha Nephelometry' alpha + a2;
          estimate 'theta Nephelometry' theta + t2;
run;

/* check assumption of normality for random effects */
proc univariate data=randeffs plot normal;
  class effect;
  var estimate;

run;


/* this model assumes that all three curves have the same shape and position.
   The position is the same because there are no covariates included for accuracy.
   Comparison with the previous model allows us to test whether accuracy varies by method. */

proc nlmixed data=nishimura_RF ecov cov ;

          parms alpha=2 theta=0 beta=0 s2ua=1 s2ut=1 t1=0 t2=0 ;

          logitp = (theta + ut + t1*rfm1 + t2*rfm2 + (alpha + ua)*dis)*
             exp(-(beta)*dis);

          p = exp(logitp)/(1+exp(logitp));

          model pos ~ binomial(n,p);

          random ut ua ~ normal([0,0],[s2ut,0,s2ua]) subject=study_id out=randeffs;
run;


/* check assumption of normality for random effects */
proc univariate data=randeffs plot normal;
  class effect;
  var estimate;

run;
```

Data and SAS file for Example 3 - CT versus MRI for the diagnosis of coronary artery disease

Data (schuetz.csv)

Test	Study_ID	tp	fp	fn	tn	Indirect
CT	Achenbach 2005	25	4	0	19	1
CT	Alkadhi 2008	57	12	2	79	1
CT	Andreini 2007	17	0	0	44	1
CT	Bayrak 2008	64	4	0	32	1
MRI	Bedaux 2002	7	1	0	1	1
MRI	Bogaert 2003	12	3	3	1	1
CT	Bonmassari 2006	12	2	0	8	1
CT	Brodoefel 2008	73	5	0	22	1
CT	Budoff 2008	52	30	3	142	1
CT	Cademartiri 2007	20	1	0	51	1
CT	Carrascosa 2007	13	1	1	5	1
MRI	Cheng 2006	21	0	4	3	1
CT	Chow 2007	18	0	1	7	1
CT	Coles 2007	77	13	7	16	1
CT	Cornily 2007	9	1	0	23	1
CT	Davin 2007	42	4	12	30	1
CT	Deetjen 2007	31	3	2	26	1
CT	Dewey 2006	62	5	4	46	0
MRI	Dewey 2006	42	2	7	39	0
CT	Dewey 2009	11	1	0	17	1
CT	Ehara 2006	59	1	1	6	1
CT	Erdogan 2006	33	2	3	5	1
CT	Garcia 2006	58	58	1	70	1
CT	Gaudio 2008	16	2	1	48	1
MRI	Gerber 2005	17	1	2	6	1
CT	Ghersin 2006	29	11	6	13	1
CT	Ghostine 2006	28	2	1	35	1
CT	Gilard 2006	11	9	0	35	1
CT	Grosse 2007	29	0	1	10	1
MRI	Hackenbroch 2004	18	5	4	13	1
CT	Hacker 2007	19	1	1	9	1
CT	Halon 2007	72	10	13	16	1
CT	Hausleiter 2007	101	35	1	106	1
CT	Henneman 2006	12	1	1	6	1
CT	Henneman 2008	28	0	0	12	1
CT	Herzog 2007a	19	6	0	30	1
CT	Herzog 2007b	16	1	0	23	1
CT	Herzog 2008	18	2	0	10	1
CT	Hoffmann 2004	19	3	2	9	1
CT	Hoffmann 2005	43	2	2	28	1
MRI	Ichikawa 2007	11	8	6	33	1
MRI	Ikonen 2003	42	15	5	7	1
CT	Johnson 2007	17	2	0	16	1
CT	Kaiser 2005	97	18	16	18	1
CT	Kefer 2005	32	6	2	12	0
MRI	Kefer 2005	30	9	4	9	0
MRI	Kim 2001	56	25	4	18	1
MRI	Klein 2008	20	11	2	13	1
CT	Kolnes 2006	33	8	1	8	1
CT	Laissy 2007	11	2	2	25	1
CT	Langer 2009	25	2	1	40	0
MRI	Langer 2009	18	15	8	27	0
CT	Leber 2007	20	7	1	60	1
CT	Leschka 2005	47	0	0	20	1
CT	Leschka 2008a	69	8	2	35	1
CT	Leschka 2008b	35	5	1	33	1
CT	Maintz 2007	15	2	1	2	0
MRI	Maintz 2007	15	1	1	3	0
CT	Manghat 2007	3	0	0	12	1
CT	Marano 2008	179	17	12	119	1
CT	Martuscelli 2004	43	9	0	9	1
CT	Maruyama 2008	75	5	2	65	1
MRI	McCarthy 2007	13	6	2	8	1
CT	Meijboom 2006	18	4	0	48	1
CT	Meijboom 2007	88	4	0	12	1
CT	Meijboom 2008	244	41	2	73	1
CT	Miller 2007	139	13	24	115	1
CT	Mir-Akbari 2009	41	11	10	20	1
CT	Mollet 2004	106	3	0	18	1
CT	Mollet 2005b	31	3	0	17	1
CT	Mollet 2005a	38	1	0	12	1
CT	Moon 2005	30	2	5	21	1

	Morgan-Hughes					
CT	2005	32	1	0	24	1
CT	Nikolaou 2006	4	3	1	52	1
CT	Nikolaou 2006b	38	6	1	23	1
CT	Olivetti 2006	15	0	3	13	1
CT	Oncel 2007a	62	0	0	18	1
CT	Oncel 2007b	8	1	1	5	1
CT	Pontone 2007a	66	7	0	43	1
CT	Pontone 2007b	56	5	4	31	1
CT	Postel 2007	42	5	5	34	1
CT	Pouleur 2008	16	7	1	53	0
MRI	Pouleur 2008	17	17	0	43	0
CT	Pugliese 2006	25	1	0	9	1
CT	Pundziute 2008	53	4	1	42	1
CT	Raff 2005	38	3	2	27	1
CT	Reant 2006	12	6	1	21	1
MRI	Regenfus 2000	34	6	2	8	1
CT	Rixe 2009	40	6	0	30	1
CT	Rodevand 2006	49	37	0	15	1
CT	Romeo 2007	43	2	0	123	1
CT	Ropers 2003	35	8	6	28	1
CT	Ropers 2006	25	5	1	50	1
CT	Ropers 2007	41	11	1	47	1
MRI	Sakuma 2006	42	6	9	56	1
MRI	Sandstede 1999	10	1	1	7	1
CT	Scheffel 2006	14	0	1	15	1
CT	Scheffel 2007	13	2	0	35	1
CT	Scheffel 2008	66	4	0	50	1
CT	Schuijf 2006	29	1	2	28	1
CT	Shabestari 2007	104	10	4	20	1
CT	Stolzmann 2008	55	2	0	43	1
CT	Tsai 2007	50	5	1	22	1
CT	Turkvatan 2008	116	2	2	33	1
CT	Ulimoen 2008	32	6	4	6	1
CT	Watkins 2007	44	3	1	37	1
CT	Weustink 2007	76	3	1	20	1
MRI	Yang 2009	32	5	2	23	1

SAS Program (scheutz.sas)

```
proc import out=schuetz
   datafile='C:\chapter10\schuetz.csv'
   dbms=csv
         replace;
   getnames=yes;
run;


/* Create two separate records for the true results in each study,
   the first for the diseased group, and the second for the non-diseased group.
   The variable sens is an indicator which takes the value 1 if true=true positives and 0 otherwise,
   the variable spec is also an indicator that takes the value 1 if true =true negatives and 0 otherwise */

data schuetz;
         set  schuetz;
         testtype=0;
         if test ="CT"  then testtype=1;
         sens=1; spec=0; true=tp; n=tp+fn; output;
         sens=0; spec=1; true=tn; n=tn+fp; output;
run;


/* Ensure that both records for a study are clustered together  */
proc sort data=schuetz;
         by study_id test;
run;


/* Run the Bivariate model with no covariates
   The "cov" option requests that a covariance matrix is printed for
   all model parameter estimates. The "ecov" option requests a covariance matrix
   for all additional estimates that are computed. */

proc nlmixed data=schuetz cov ecov;

         parms msens=2 mspec=1  s2usens=0 s2uspec=0 covsesp=0 ;

         logitp=(msens+usens)*sens+(mspec+uspec)*spec;

         p = exp(logitp)/(1+exp(logitp));

         model true ~ binomial(n,p);

         random usens uspec ~ normal([0,0],[s2usens,covsesp,s2uspec]) subject=study_id out=randeffs;

run;


/* Bivariate model with test as a covariate using the indicator variable testtype.
   MRI is the reference category.
   Variances of the random effects are assumed not to vary by test type. */

proc nlmixed data=schuetz cov ecov;

         parms msens=2 mspec=1  s2usens=0 s2uspec=0 covsesp=0 se_CT=1 sp_CT=0;

         logitp=(msens+usens+se_CT*testtype)*sens+(mspec+uspec+sp_CT*testtype)*spec;

         p = exp(logitp)/(1+exp(logitp));

         model true ~ binomial(n,p);

         random usens uspec ~ normal([0,0],[s2usens,covsesp,s2uspec]) subject=study_id out=randeffs;

         /* Estimate logit(sensitivity), and logit(specificity) */
         estimate 'logitsens CT'  msens + se_CT;
         estimate 'logitspec CT'  mspec + sp_CT;

run;

/* Check assumption of normality for the random effects */
proc univariate data=randeffs plot normal;
  class effect;
  var estimate;
run;

/* Bivariate model with effect of test type on only sensitivity */

proc nlmixed data=schuetz cov ecov;
```

```
        parms msens=2 mspec=1  s2usens=0 s2uspec=0 covsesp=0 se_CT=1 ;

        logitp=(msens+usens+se_CT*testtype)*sens+(mspec+uspec)*spec;

        p = exp(logitp)/(1+exp(logitp));

        model true ~ binomial(n,p);

        random usens uspec ~ normal([0,0],[s2usens,covsesp,s2uspec]) subject=study_id out=randeffs;

run;


/* Bivariate model with effect of test type on only specificity */

proc nlmixed data=schuetz cov ecov;

        parms msens=2 mspec=1  s2usens=0 s2uspec=0 covsesp=0 sp_CT=0;

        logitp=(msens+usens)*sens+(mspec+uspec+sp_CT*testtype)*spec;

        p = exp(logitp)/(1+exp(logitp));

        model true ~ binomial(n,p);

        random usens uspec ~ normal([0,0],[s2usens,covsesp,s2uspec]) subject=study_id out=randeffs;

run;


/* DIRECT COMPARISONS */

/* Create new dataset of studies with within-study comparison of CT and MRI.
"indirect" is a binary variable in the dataset coded 1 if the study evaluated
only one test (CT or MRI) and 0 if both tests were evaluated in a study */

data schuetz_direct;
        set schuetz;
        where indirect=0;
run;


/* Fit Bivariate model without covariate */
proc nlmixed data=schuetz_direct cov ecov qpoints=10;

        parms msens=2 mspec=1  s2usens=0 s2uspec=0 covsesp=0 ;

        bounds s2usens>=0;
        bounds s2uspec>=0;

        logitp=(msens+usens)*sens+(mspec+uspec)*spec;

        p = exp(logitp)/(1+exp(logitp));

        model true ~ binomial(n,p);

        random usens uspec ~ normal([0,0],[s2usens,covsesp,s2uspec]) subject=study_id out=randeffs;

run;

/* Fit Bivariate model with fixed effects */
proc nlmixed data=schuetz_direct cov ecov qpoints=10;

        parms msens=2 mspec=1;

        logitp=(msens)*sens+(mspec)*spec;

        p = exp(logitp)/(1+exp(logitp));

        model true ~ binomial(n,p);

run;

/* Fit Bivariate model with random effect for specificity only */

proc nlmixed data=schuetz_direct cov ecov qpoints=10;

        parms msens=2 mspec=1 s2uspec=0;

        logitp=(msens)*sens+(mspec+uspec)*spec;

        p = exp(logitp)/(1+exp(logitp));

        model true ~ binomial(n,p);
        random uspec ~ normal([0],[s2uspec]) subject=study_id out=randeffs;
```

```
run;
/* Fit Bivariate model with random effect for sensitivity only */
proc nlmixed data=schuetz_direct cov ecov qpoints=10;

          parms msens=2 mspec=1 s2usens=0;

          logitp=(msens+usens)*sens+(mspec)*spec;

          p = exp(logitp)/(1+exp(logitp));

          model true ~ binomial(n,p);
          random usens ~ normal([0],[s2usens]) subject=study_id out=randeffs;

run;


/* Fit Bivariate model with covariate for test type on both sens and spec.
   Random effects only for specificity */
proc nlmixed data=schuetz_direct cov ecov qpoints=10;

          parms msens=2 mspec=1  s2uspec=0 se_CT=0 sp_CT=0;

          bounds s2uspec>=0;

          logitp=(msens+se_CT*testtype)*sens+(mspec+uspec+sp_CT*testtype)*spec;

          p = exp(logitp)/(1+exp(logitp));

          model true ~ binomial(n,p);

          random uspec ~ normal([0],[s2uspec]) subject=study_id out=randeffs;

          /* Estimate logit(sensitivity) and logit(specificity) */
          estimate 'logitsens CT'  msens + se_CT;
          estimate 'logitspec CT'  mspec + sp_CT;

run;
/* Fit Bivariate model with covariate for test type on specificity.
   Random effects only for specificity */
proc nlmixed data=schuetz_direct cov ecov qpoints=10;

          parms msens=2 mspec=1  s2uspec=0 sp_CT=0;

          bounds s2uspec>=0;

          logitp=(msens)*sens+(mspec+uspec+sp_CT*testtype)*spec;

          p = exp(logitp)/(1+exp(logitp));

          model true ~ binomial(n,p);

          random uspec ~ normal([0],[s2uspec]) subject=study_id out=randeffs;

run;
/* Fit Bivariate model with covariate for test type on sensitivity.
   Random effects only for specificity */
proc nlmixed data=schuetz_direct cov ecov qpoints=10;

          parms msens=2 mspec=1  s2uspec=0 se_CT=0;

          bounds s2uspec>=0;

          logitp=(msens+se_CT*testtype)*sens+(mspec+uspec)*spec;

          p = exp(logitp)/(1+exp(logitp));

          model true ~ binomial(n,p);

          random uspec ~ normal([0],[s2uspec]) subject=study_id out=randeffs;


run;
```

References

Arends 2008
Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MG, Stijnen T. Bivariate random effects meta-analysis of ROC curves. *Med Decis Making* 2008; 28: 621-638.

Begg 1994a
Begg CB. Publication bias. In: Cooper J HL (editors). *The Handbook of Research Synthesis*. New York: Sage Foundation, 1994.

Begg 1994b
Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics* 1994; 50: 1088-1101.

Chappell 2009
Chappell FM, Raab GM, Wardlaw JM. When are summary ROC curves appropriate for diagnostic meta-analyses? *Stat Med* 2009; 28: 2653-2668.

Chu 2006
Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *J Clin Epidemiol* 2006; 59: 1331-1332.

Deeks 2001
Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001; 323: 157-162.

Deeks 2008
Deeks JJ, Higgins JPT, Altman DG. Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins JPT, Green S (editors). *Cochrane Handbook for Systematic Reviews of Interventions.* Chichester (UK): John Wiley & Sons, 2008.

Deeks 2005
Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol* 2005; 58: 882-893.

Dickersin 1990
Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA* 1990; 263: 1385-1389.

Dukic 2003
Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. *Biometrics* 2003; 59: 936-946.

Egger 1997
Egger M, Davey SG, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997; 315: 629-634.

Hamza 2009
Hamza TH, Arends LR, van Houwelingen HC, Stijnen T. Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. *BMC Med Res Methodol* 2009; 9: 73.

Harbord 2007
Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007; 8: 239-251.

Harbord 2006
Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med* 2006; 25: 3443-3457.

Harbord 2009
Harbord RM, Whiting P. metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression. *Stata Journal* 2009; 9: 211-229.

Higgins 2003
Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003; 327: 557-560.

Ioannidis 1998
Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA* 1998; 279: 281-286.

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

Leeflang 2008
Leeflang MM, Moons KG, Reitsma JB, Zwinderman AH. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. *Clin Chem* 2008; 54: 729-737.

Littenberg 1993
Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making* 1993; 13: 313-321.

Macaskill 2004
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: 925-932.

Macaskill 2001
Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. *Stat Med* 2001; 20: 641-654.

McCullagh 1980
McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society Series B* 1980; 42: 109-142.

Moses 1993
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: 1293-1316.

Nishimura 2007
Nishimura K, Sugiyama D, Kogata Y, Tsuji G, Nakazawa T, Kawano S, Saigo K, Morinobu A, Koshiba M,

Kuntz KM, Kamae I, Kumagai S. Meta-analysis: diagnostic accuracy of anti-cyclic citrullinated peptide antibody and rheumatoid factor for rheumatoid arthritis. *Ann Intern Med* 2007; 146: 797-808.

Pepe 2003
Pepe M. *The Statistical Evaluation of Medical Tests for Misclassification and Prediction*. Oxford: Oxford University Press, 2003.

Peters 2006
Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. *JAMA* 2006; 295: 676-680.

Reitsma 2005
Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005; 58: 982-990.

Rutter 1995
Rutter CM, Gatsonis CA. Regression methods for meta-analysis of diagnostic test data. *Acad Radiol* 1995; 2 Suppl 1: S48-S56.

Rutter 2001
Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001; 20: 2865-2884.

Schuetz 2010
Schuetz GM, Zacharopoulou NM, Schlattmann P, Dewey M. Meta-analysis: noninvasive coronary angiography using computed tomography versus magnetic resonance imaging. *Ann Intern Med* 2010; 152: 167-177.

Schwarzer 2002
Schwarzer G, Antes G, Schumacher M. Inflation of type I error rate in two statistical tests for the detection of publication bias in meta-analyses with binary outcomes. *Stat Med* 2002; 21: 2465-2477.

Takwoingi 2008
Takwoingi Y, Deeks JJ. METADAS: A SAS macro for meta-analysis of diagnostic accuracy studies (available at http://srdta.cochrane.org/software-development). Cochrane Collaboration, 2008.

Tosteson 1988
Tosteson AN, Begg CB. A general regression methodology for ROC curve estimation. *Med Decis Making* 1988; 8: 204-215.

Zhou 2002
Zhou XH, Obuchowski N, McClish D. *Statistical Methods in Diagnostic Medicine*. Chichester: Wiley, 2002.

Zwinderman 2008
Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. *Stat Med* 2008; 27: 687-697.