Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy

Chapter 11 Interpreting results and drawing conclusions

Patrick Bossuyt, Clare Davenport, Jon Deeks, Chris Hyde, Mariska Leeflang, Rob Scholten.

Version 0.9 Released December 13th 2013.

Please cite this version as: Bossuyt P, Davenport C, Deeks J, Hyde C, Leeflang M, Scholten R. Chapter 11:Interpreting results and drawing conclusions. In: Deeks JJ, Bossuyt PM, Gatsonis C (editors), *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* Version 0.9. The Cochrane Collaboration, 2013. Available from: http://srdta.cochrane.org/.

Saved date and time 13/12/2013 10:32 Jon Deeks

Contents

11 Interpreting results and Drawing Conclusions

11.1 Key points

- The relative unfamiliarity of DTA methods and accuracy metrics exacerbates the challenges associated with communicating review findings to a range of audiences. Review authors should consider re-expressing results and findings in sentences and numbers which will help readers understand the key findings.

- The Summary of Findings Table (SoF) brings together the key elements of a review's findings and provides information on the quantity, quality and applicability of evidence as well as the accuracy of index test(s). The main purpose of the SoF table in a DTA review discussion is to improve ease of interpretation. SoF tables should be placed ahead of the main text of the discussion section.

- Cochrane DTA reviews use three fixed subheadings under the main text discussion section to guide the interpretation of results: 'Summary of main results' 'Strengths and weaknesses of the review', and 'Applicability of findings to review question'. The authors' conclusions section is divided into 'Implications for practice' and 'Implications for research'.

- When discussing implications for practice the intended application and role of index test(s) and the possible consequences of false positive and false negative test errors should be considered. Authors may want to refer to related effectiveness research or research associated with test reliability, cost and acceptability whilst acknowledging that this will not have been evaluated in a systematic way. After discussing the balance of benefits and harms, review authors may want to highlight specific actions that might be consistent with particular patterns of values and preferences.

- When discussing implications for research authors should place the findings of their review in the context of other research related to the clinical question and specify the nature of any further research required: further accuracy studies or other dimensions of test evaluation (for example effectiveness, cost-effectiveness).

11.2 Introduction

The purpose of Cochrane reviews is to facilitate healthcare decision-making by patients and the general public, by clinicians or other healthcare workers, administrators, and policy makers. Such people will rely on the discussion section and the authors' conclusions to make sense of the information in the review and to help them to interpret the results. Because of the importance of the discussion and conclusion sections, authors need to take great care that these sections accurately reflect the data and information contained in the review.

The meta-analysis in a systematic review of test accuracy studies may result in a summary estimate of the test's sensitivity and specificity, in a summary ROC curve and corresponding parameters, or in summary estimates of comparative accuracy. The relative unfamiliarity of DTA methods and accuracy metrics exacerbates the challenges associated with communicating review findings to a range of audiences. These challenges usually relate to the relative complexity of summary statistics,

communicating the clinical significance of unexplained heterogeneity and the applicability of review findings.

In addition, the contribution of estimation of test accuracy to evidence-based decision making needs to be made explicit. Accuracy data usually do not provide readers with clear answers about whether to buy, reimburse or order tests. Such decisions usually need more information concerning the consequences of testing (i.e. consequences for index test positive results and index test negative results), and other ways in which tests impact on patients. The discussion section of a DTA review should at least alert readers about this and indicate where the additional information might be found.

Above all, readers need to weigh the results and their implications against the quality of the body of evidence they stem from. This implies that the discussion and conclusions sections should include some summary statements about the quality of the evidence.

A 'Summary of Findings' (SoF) table, described in Section 11.8 provides key information in a quick and accessible format. Review authors must include such tables in Cochrane DTA reviews. The discussion section should provide explanatory information and complementary considerations.

The Cochrane DTA review structure has three fixed subheadings under the discussion section to guide the interpretation of the results: 'Summary of main results' 'Qualifying DTA evidence', and 'Applicability of findings to the review question'. The authors' conclusions section is divided into 'Implications for practice' and 'Implications for research'. In this chapter we provide suggestions on how to approach each of these sections.

11.3 Summary of main results

The summary of main results section should begin with a restatement of the question or questions that the review is attempting to answer. The number and essential characteristics of studies in the review should be summarized, including summary statements about the results of the quality assessment and a summary of the relevance of the findings from investigations of heterogeneity.

The review question should be followed by the Summary of Findings (SoF) table (see Section 11.8 below) which should act as a template for, and precede, the narrative discussion in DTA reviews. The main purpose of the SoF table is to improve ease of interpretation but it can also be used by review authors to ensure that general statements in the conclusions are linked to and supported by data in the results section of the review.

11.4 Summarising statistical findings

Review authors need to present the key findings of their review in the Summary of Main Results section of the discussion and the Summary of Findings Table. It is important that the findings are explained in ways that make them accessible to the different audiences who may use the review.

The complex meta-analytical methods that are used in Cochrane DTA reviews are likely to be unfamiliar to many readers, and the summary statistics and conditional probabilities used to

describe test performance (e.g. sensitivity and specificity and positive and negative predictive values) are often confused and misinterpreted. Review authors should consider re-expressing results and findings in sentences and numbers which will help readers understand the key findings whilst minimising the use of statistical terminology.

Chapter 10 illustrated the derivation of various summary statistics used to express test accuracy. In this chapter we will focus on the interpretation of these summary statistics and illustrate characteristics that determine how useful different summary statistics are when drawing conclusions from a DTA review. Authors should be discerning in the choice of metrics they report, considering the relative importance of false negative and false positive test errors and any limitations imposed on meta-analysis by the data available in primary studies.

Evaluation of test accuracy is an explicit recognition that most tests are imperfect and summary test accuracy statistics are used to communicate the size, and for some metrics, the direction (false positive or false negative) of erroneous test results. False negative and false positive test results will have different possible consequences depending on the testing context. In many situations, the impacts of false positive and false negative test results will vary in importance. Review authors should therefore be mindful of the possible consequences when interpreting results and drawing conclusions.

For example, consider the implications of tests used in cervical cancer screening programmes. Women who get false positive test results may suffer unnecessary anxiety and further, possibly invasive investigations to confirm a diagnosis. Women with false negative test results may suffer a considerable delay in diagnosis because screening intervals are typically several years. The consequences of such a delay may be a requirement for more invasive and toxic treatments and even increased mortality. By contrast, consider a test being used to diagnose high blood pressure. Individuals with false positive results will be subject to unnecessary life-long treatment and the consequences of having a label of 'hypertensive'. Those with false negative results will suffer a delay in treatment, but this is less likely to result in adverse consequences compared to a missed diagnosis of cervical cancer; there is a considerable delay between the onset of hypertension and its complications and blood pressure is measured relatively more frequently.

Test accuracy summary statistics can be broadly grouped into two types: paired and global. The use of global measures for meta-analysis has been discussed in Chapter 10. Paired summary statistics distinguish between the ability of a test in two dimensions: the ability of a test to correctly identify individuals with a condition of interest (the magnitude of false negative test errors) and the ability of a test to correctly identify individuals *without* a condition of interest (the magnitude of false positive test errors). Global summary statistics express the overall discriminatory ability of a test (the ability of a test to discriminate between those with and those without disease). Paired summary statistics are more clinically useful because they distinguish between the two dimensions of test accuracy and, as discussed above, the relative importance of the direction of test errors (false positives and false negatives) usually differs in specific testing contexts.

11.4.1 Paired summary statistics

Paired summary statistics that allow the calculation of post-test probability of disease include sensitivity and specificity, and positive and negative predictive values. These are conditional

probabilities which indicate that they are computed in a subgroup of participants that fulfil a certain criterion.

Referring to the 2x2 diagnostic table (Figure 1) it can be seen that sensitivity and the negative predictive value provide information on the magnitude of false negatives (as sensitivity and the negative predictive value increase, the proportion of false negative test errors decreases). Specificity and the positive predictive value provide information on the magnitude of false positive test errors (as specificity and positive predictive value increase, the proportion of false positive test errors decreases).

Figure 1 Diagnostic 2x2 table demonstrating the computation of sensitivity, specificity and predictive values.

	Reference standard +ve	**Reference standard -ve**	
Index test +ve	True positives (TP)	False positives (FP)	Positive Predictive Value TP/(TP+FP)
Index test -ve	False negatives (FN)	True Negatives (TN)	Negative Predictive Value TN/(FN+TN)
	Sensitivity TP/(TP+FN)	Specificity TN/(FP+TN)	

Conditional probabilities are often wrongly interpreted and misunderstood because of confusion about the subgroup to which they refer (Girotto 2001) – it is therefore essential when reporting such measures to be explicit about the subgroup to which they refer.

There is a considerable body of empirical literature demonstrating that sensitivity and specificity are not well understood (Puhan 2005; Stuerer 2002) and that probabilities conditional on index test results (predictive values) rather than actual disease status (sensitivity and specificity) may be more intuitive to decision makers (Reid 1998). Historically the use of predictive values has been discouraged because, unlike sensitivity and specificity, predictive values are mathematically dependent on the pre-test probability (prevalence) of the target disorder; (as prevalence increases, positive predictive values increase and negative predictive values decrease). This has implications for the transferability of predictive values between different health care settings. However, with increasing recognition of the variation in estimates of test accuracy caused by differences in the mix and severity of disease (spectrum of disease), even in populations of similar prevalence, authors should be mindful of transferability regardless of the type of summary statistic used.

11.4.1.1 Sensitivity and specificity

Sensitivity is calculated in relation to (conditional on) the sub-group of study participants who are reference standard positive (have the target condition) and for specificity study participants who are reference standard negative (do not have the target condition). Thus sensitivity expresses the

performance of the test in those who have the condition, and specificity in those who do not have the condition.

When sensitivity and specificity are reported as an output of meta-analysis these metrics need to be interpreted as 'average' estimates across included studies. Sensitivity and specificity vary with threshold, and computation of an average value makes sense only when the studies have used a common threshold. Thus analyses may need to be restricted to a subset of studies as explained in Chapter 10.4.1, or multiple analyses should be undertaken at different thresholds. When index tests are being compared, estimation of sROC curves may be helpful to increase statistical power; this is discussed in Sections 11.4.3 below.

11.4.1.2 Predictive values

The positive predictive value is calculated in relation to (conditional on) the sub-group of participants who test positive with the index test and the negative predictive value in relation to those who test negative with the index test. Thus the positive predictive value describes the proportion of patients with a positive result who actually have the disease and the negative predictive value describes the proportion of people with a negative test result who do not have the disease. In other words, predictive values state how good a positive test result is at ruling in disease, and a negative test result at ruling out disease.

Meta-analysis of predictive values is possible (Leeflang 2012). However, as discussed in Chapter 10, between-study variation in prevalence may complicate the investigation of heterogeneity, therefore the average predictive values calculated will relate to the use of the test at some average, but unknown, prevalence. If authors wish to use predictive values as a means of expressing test accuracy from a meta-analysis they should compute average sensitivity and specificity and then compute predictive values based on average estimates of sensitivity and specificity at a representative pre-test probability (prevalence) of the target condition.

Predictive values are most simply obtained from summary estimates of sensitivity and specificity by creating an illustrative 2x2 table and computing predictive values directly (the simple equations to do this are in Chapter 10, Section 10.2.3). This exercise can be done on paper or by using the 2x2 calculator built into the data entry tool in RevMan.

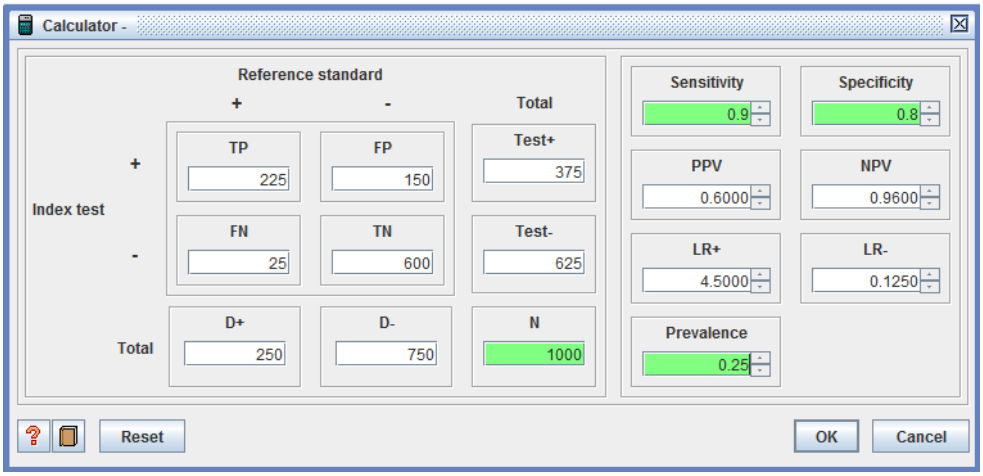
To compute predictive values, enter a fictional sample size (say 1000), the prevalence, and the estimated average sensitivity and specificity of the test – i.e. the boxes in green in Figure 2. (To access the calculator you need to be highlighting a study within the "data and analyses" section of a Cochrane Review, then use the button showing a calculator icon in the top, right-hand section of the screen).

For example, a test which has sensitivity of 0.9 and specificity of 0.8 yields the following table for a pre-test probability of disease prevalence of the target condition of 0.25 and a total sample size of 1000. This computes the positive predictive value to be 0.6 and the negative predictive value to be 0.96.

Although predictive values may be intuitive summary metrics, choosing the estimate of pre-test probability (prevalence) at which to estimate these values may not be straightforward. Estimates of a representative pre-test probability of the target disorder (prevalence) may be obtained from the

distribution of prevalence observed in the studies included in the systematic review but only if the studies are thought to be representative of the target setting. For example, the median value of prevalence might be used, although it is important to exclude case-control studies where reported prevalence is an artefact of the study design. Alternatively, authors may consider computing predictive values across a range of plausible prevalence estimates for the target setting. In some circumstances, estimates of disease prevalence may be more reliably obtained from other data sources such as disease registries. Interpretations of summary estimates of predictive values should reflect the fact that spectrum and threshold cause variation in all summary estimates of test accuracy, even when studies have a similar pre-test probability of the target disorder (prevalence).

Figure 2 Illustration of RevMan calculator conversion of sensitivity and specificity to positive and negative predictive values at a pre-test probability (prevalence) of 25%



TP: true positive; FP: false positive; FN: false negative; TN: true negative; D+ : disease positive; D- : disease negative; PPV: positive predictive value; NPV: negative predictive value; LR+ positive likelihood ratio; LR- negative likelihood ratio.

11.4.1.3 Use of normalised frequencies to present conditional probabilities

Sensitivity and specificity, and positive and negative predictive values, are typically presented as proportions or percentages. Presenting probabilities as frequencies has been shown to help readers understand their meaning (Evans 2000; Hoffrage 1998; Zhelev 2013), and this approach is encouraged both in the Summary of Main Results section of the review and in the Summary of Findings table.

A normalised frequency description expresses a proportion in terms of the number of individuals in whom an event or outcome is observed out of a group (typically 10, 100 or 1000). As with conditional probabilities, it is important to be explicit about the group to which normalised frequencies refer. For example, they may refer to all those tested, those with or without disease, or those with positive or with negative index test results.

Referring to the RevMan calculator, normalised frequency expression can be used to describe the absolute impact of a test in a population with a given prevalence (25% in Figure 2 above):

- For a test with a positive predictive value of 60%: 60 out of every 100 positive index test results will actually have disease but 40 will not (i.e. will be false positives). In a population with a pre-test probability (prevalence) of 25% (see figure 2 above) this will result in 150 false positive test results for every 1000 people tested.
- For a test with a negative predictive value of 96%: 96 out of every 100 negative index test results will not have disease but 4 will (i.e. be false negatives). In a population with a pre-test probability (prevalence) of 25% (see figure 2 above) this will result in 25 false negative test results for every 1000 people tested.

Note that if the test were applied in a setting with a different prevalence, the absolute number of false positives and false negatives would change.

There may also be advantages in using a normalised frequency representation of sensitivity and specificity. Although sensitivity and specificity do not provide information on the absolute impact of a test at a particular prevalence of disease, expressing them as normalised frequencies may help readers to interpret them. In addition, normalised frequencies explicitly illustrate that sensitivity is providing information on the false negative rate and specificity on the false positive rate. For example, referring to the RevMan calculator in Figure 2 above:

- For a test with a sensitivity of 90%: the index test will detect 90 out of every 100 with disease but 10 will be missed (i.e. will be false negatives);
- For a test with a specificity of 80%: of every 100 individuals without the disease, 20 will be wrongly diagnosed as having it (i.e. will be false positives).

Authors should remember that the absolute number of false positive and false negative test results observed in a population will depend on the prevalence of the disease being studied: as prevalence decreases the absolute number of false negatives decreases and the absolute number of false positives increases. Sensitivity and specificity are not mathematically dependent on prevalence and therefore estimates of the number of false negatives and false positives derived from these accuracy metrics will be constant across populations with different prevalence of disease.

11.4.1.4 Likelihood ratios

The use of likelihood ratios (see 10.2.3.3) to express test performance has been promoted as a metric that facilitates Bayesian probability updating (derivation of post-test probabilities) (Sackett 2000). However, evidence that likelihood ratios improve diagnostic decision making is lacking.

A positive likelihood ratio is a ratio of the proportion of index test positives in individuals with disease (sensitivity) to the proportion of index test positives in individuals without disease (1-specificity). A positive likelihood ratio therefore indicates how many more times likely positive index test results will occur in individuals with disease than in individuals without disease.

A negative likelihood ratio is a ratio of the proportion of index test negatives in individuals with disease (1-sensitivity) to the proportion of index test negatives in individuals without disease (specificity). A negative likelihood ratio therefore indicates how many times less likely negative index test results will occur in individuals with disease than in individuals without disease.

A guide for the interpretation of likelihood ratios suggests positive likelihood ratios greater than 10 as indicating a useful change (increase) in the probability of disease before and after a positive test

result and negative likelihood ratios below 0.1 have been promoted as indicating a useful decrease in the probability of disease before and after a negative test result (Jaeschke 2002). However such a universal rule has been criticised, as the usefulness of changes in pre to post test probability will be affected by the pre-test probability (prevalence) of disease. For example, for a rare (low prevalence) disease, larger positive likelihood ratios will be needed to cause a useful increase in disease probability following a positive index test result (an increase in the probability of disease that might result in a change in management). For a common (high prevalence) disease, smaller negative likelihood ratios will be needed to cause a useful decrease in the probability of disease following a negative index test result.

If review authors chose to report test accuracy using likelihood ratios, the meta-analysis macros in STATA (metandi) and SAS (metadas) automatically compute likelihood ratios with 95% confidence intervals which can be reported in the review results. If not, point estimates for likelihood ratios can be obtained from the RevMan 2x2 calculator, or by hand using the equations in Chapter 10 (10.2.3) but no confidence intervals will be available.

11.4.2 Global measures of test accuracy

Global measures of test accuracy provide information about the overall discriminatory power of a test as a single number over a range of test positivity thresholds. These characteristics have advantages for model building as part of meta-analysis and where included studies in a review evaluate tests over a range of test positivity thresholds. However, global measures of test accuracy fail to distinguish between false negative and false positive test errors

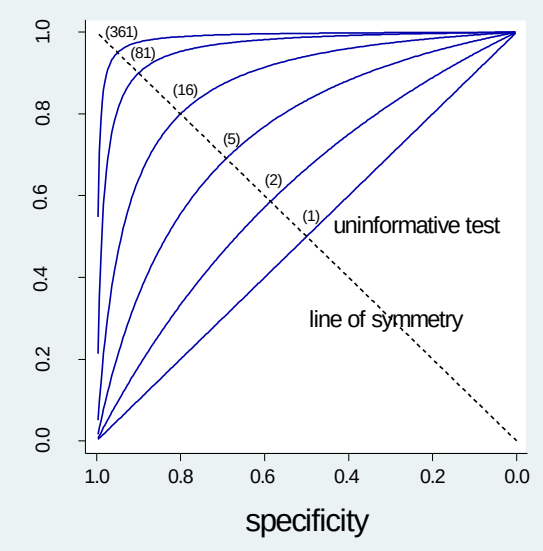
11.4.2.1 Summary Receiver Operator Characteristic Curves (sROC curves)

The summary ROC (sROC) curve is a graph showing how sensitivity and specificity values change as threshold (or some quantity related to threshold-dependent changes in test accuracy), varies across studies included in a review. Test accuracy is usefully summarised as a sROC curve when there is no common threshold or thresholds that could be used to create sub-groups of studies for separate meta analyses (see 11.4.1.1 above), or where authors wish to avoid sub-grouping studies in order to maximise statistical precision and power.

As the discriminatory power of a test increases, the sROC curve locates nearer to the top left hand corner in ROC space towards the point where sensitivity and specificity both equal 1 (100%). The sROC curve of an uninformative test would be the upward diagonal of the sROC plot.

In contrast to ROC curves plotted in individual primary studies, sROC curves do not allow identification of points on the curve that relate to a particular threshold, thus it is not possible to say what threshold a test would have to operate at to obtain a particular combination of sensitivity and specificity. However it may be helpful to identify key sensitivity/specificity pairs from the curve to illustrate performance. For example, if minimising false positives (and therefore maximising specificity) in a particular testing context is relatively more important than maximising sensitivity, the sensitivity of the test could be reported at the minimum acceptable specificity (for example a specificity of 95%). If authors choose to report sensitivity and specificity pairs from a sROC curve then the most informative and reliable estimates are likely to be points on the curve that lie within the range of the observed included study values of sensitivity and specificity rather than areas of the curve that are extrapolated from observed data.

Figure 3: Summary Receiver Operator Characteristic (sROC) curve



11.4.2.2 Diagnostic Odds Ratios (DOR) and area under the curve (AUC)

Section 10.2.6 explains how global test accuracy statistics, such as the DOR and the area under the curve (AUC) relate to sROC curves and give a single numerical value to describe test performance across all thresholds. The DOR (see also 10.2.4) is the cross product of the 2x2 diagnostic contingency table (DOR= (TP X TN) / (FP X FN)).

A diagnostic odds ratio of 1 represents an uninformative test (the upward diagonal in Fig 3 above) and as the sROC curve moves into the ideal position in the top left hand corner of the sROC plot, the DOR increases, reflecting a test with increasing discriminatory power.

When interpreting DORs, authors should note that the same DOR may be achieved by different combinations of sensitivity and specificity (as shown in Figure 4). For example a DOR of 9 could be achieved by a specificity of 90% and a sensitivity of 50% or by a sensitivity of 50% and a specificity of 90%. For this reason, and the fact that their interpretation is not intuitive (they express events in terms of odds rather than probabilities), the DOR should be considered an output statistic from hierarchical models fitted and not a suitable summary test statistic to describe test performance.

DORs are most useful in meta-analysis when making comparisons between tests or between subgroups as described below in section 11.4.3.

The AUC is the area under the ROC curve and has interpretations as “the average sensitivity across all possible specificities”, or the “probability that the test will correctly rank a randomly chosen diseased patient above a randomly chosen non-diseased patient”. An AUC of 0.5 represents an uninformative test and an AUC of 1 (where the sROC curve would be in the top left hand corner in ROC space) represents a test with 100% sensitivity and 100% specificity. Although AUC statistics are sometimes reported in primary studies, they are very rarely reported as a meta-analytical summary, and are not routinely computed by any of the meta-analytical methods reported in Chapter 10.

Figure 4: Diagnostic Odds ratios (DORs) achieved at different values of sensitivity and specificity

	Sensitivity						
Specificity	*50%*	*60%*	*70%*	*80%*	*90%*	*95%*	*99%*
50%	1	2	2	4	(9)	19	99
60%	2	2	4	6	14	29	149
70%	2	4	5	(9)	21	44	231
80%	4	6	(9)	16	36	76	396
90%	(9)	14	21	36	81	171	891
95%	19	29	44	76	171	361	1881
99%	99	149	231	396	891	1881	9801

The ringed figures indicate sensitivity-specificity combinations which have the same DOR=9.

11.4.3 Interpretation of summary statistics comparing index tests

Authors should consider two issues for reviews that compare multiple tests: the statistical measures that can be used, and the strength of evidence of the comparison. The second issue relates to whether the meta-analysis is based on within- or between- study comparisons of tests, and will be considered in section 11.6 (Qualifying the evidence). The appropriate statistical measures are not affected by this issue.

Presentation of test comparisons is facilitated by summaries of test accuracy in sROC space which allow readers to compare test performance in one figure. This may be in the form of sROC curves, (shape and relative position) or summary estimates of sensitivity and specificity. In addition, within-study (direct) test comparisons can be annotated to distinguish them from between-study (indirect) comparisons.

As for a single test, estimation and comparison of the average sensitivity and specificity of more than one index test only makes sense when each test has been evaluated at a common threshold. Although comparison of tests where studies report a mix of thresholds may most powerfully be made using the HSROC approach to maximise the number of studies included in the meta-analysis, interpretation of such comparisons is challenging and should be done with caution (see 11.4.3.2 below).

When summarising findings from a comparison of two tests, a review author should focus on describing 1) the magnitude and direction of the difference between tests and 2) the evidence that the difference is not explicable by chance.

A meta-analysis model that compares tests will produce one of two sets of output depending on whether the analysis has been undertaken using the bivariate model or the HSROC model:

11.4.3.1 Comparing tests using sensitivity and specificity (bivariate model)

For the bivariate analysis the following statistics will be reported with confidence intervals:

- Estimates of the average sensitivity and specificity for each test
- Estimates of the relative sensitivity and relative specificity expressed as odds ratios

- P-values for the difference in sensitivity and for the difference in specificity.

When the bivariate method has been used, the magnitude and direction of the difference between tests can be summarised either by reporting point estimates of the average sensitivity and specificity for the two tests, or measures of relative test sensitivity and specificity (relative measures are computed on a logit scale, and thus are technically odds ratios). It is not possible to directly translate relative measures of accuracy to the consequences of using one or other test. Therefore focusing on the size and significance (P-values) of any *difference* in estimates of average sensitivity and specificity between tests is likely to be the most accessible way of illustrating the potential impact of using different tests.

As illustrated in section 11.4.1.3 above, expression of probabilities as frequencies is also likely to be useful when discussing the consequences of any difference between tests being compared. For example, if test A has a sensitivity of 0.85 and test B a sensitivity of 0.90, test B will correctly detect 5 more patients out of every 100 with the disease than test A; while test A will result in 5 additional false negative diagnoses compared with test B. A similar approach can be used if predictive values are the summary measure being compared; at a specified prevalence, the number of false positives or false negative diagnoses generated by two tests. Note however, that *comparing* predictive values between tests is not straightforward, as predictive values are computed from the positive (or negative) test results, which will change with each test.

11.4.3.2 Comparing tests using sROC curves and diagnostic odds ratios (HSROC model)

When the HSROC model has been used, the analysis focuses on the values of the diagnostic odds ratio for the two tests and its ratio (the rDOR) and a parameter related to the proportion test positive in the study (referred to in Chapter 10 as the threshold parameter).

For the HSROC analysis the following statistics will be reported with confidence intervals:

- Estimates of the mean diagnostic odds ratio (DOR) for each test
- Estimates of the mean threshold parameter (average underlying test positivity threshold) for each test
- Estimate of the relative diagnostic odds ratio
- Estimate of the difference in the mean threshold parameter for each test
- P-values for each of the differences in DOR and threshold parameter between tests

Optionally, the model may include a term that describes the interaction between each index test and the shape of the sROC curve. This will be reported with a P-value indicating whether the SROC curves for the two tests are parallel in logit space (the same shape) or cross-over.

Comparing sROC curves of the same shape

Provided that the curves for the tests being compared have the same shape (whether symmetrical or asymmetrical), the value of the ratio of DOR will be constant all the way along the curve and therefore derivation of the rDOR at any point gives a valid comparison of tests. Interpretation of an estimated rDOR of 2.0 (1.5, 3.0) derived from sROC curves of the same shape would be that the diagnostic odds ratio for the second test is twice that of the first, and that we are 95% certain that it is between 1.5 and 3.0 times the value of the first. However, it is not possible to say in which way

any superiority in accuracy has been obtained: e.g. whether it is due to an increase in sensitivity and / or an increase in specificity. It is therefore not possible to translate differences in accuracy to the downstream consequences of adopting different tests.

As with a single test, where tests have been compared using sROC curves it may therefore be more useful to report selected sensitivity/specificity points on each of the curves to facilitate test comparisons. For example, the sensitivity of each test at the same fixed specificity could be reported. Presenting differences at several selected values might be informative. However, it is important to note that we have no information on the threshold which should be used for the tests to function at particular chosen points on the SROC curve. Particular caution should be exerted when comparing tests where the study results lie in different sections of the summary ROC space. Estimation of DORs at points to the left of the downward diagonal on the ROC plot will be achieved by a relatively higher specificity and lower sensitivity than estimation of DORs at points to the right of the downward diagonal. In addition, authors should be cautious when choosing points for comparison to distinguish between those that lie within the range of observed data from included studies and those that are extrapolated from observed data; the former are more valid estimates.

Comparing sROC curves of different shapes

If sROC curves for different tests have different shapes, the ratio of DOR will not be constant along the entire length of the curve. Comparisons of tests where the sROC curves have different shapes are therefore challenging, as the rDOR will vary along the curve, and will even switch in terms of the direction of superiority of one test over another at the point where the curves cross. Interpretation of meta-analytical models for these situations needs to be done carefully considering the observed range of the data. Again, quoting particular values from the fitted curves may assist interpretation provided that these lie within the observed range of the data.

11.4.4 Expressing uncertainty in summary statistics

It is important to express the degree of uncertainty associated with summary estimates of test accuracy whichever metrics are used. A meta-analysis will compute confidence intervals and regions for estimates of sensitivity and specificity which should be reported alongside the point estimates in text and tables as well as being presented on the summary ROC plots in the results section. Illustrations of 95% confidence regions and prediction regions can be found in 10.5.2.2 where the 95% confidence region is a measure of within-study uncertainty (the precision of the test accuracy estimate) and the prediction region is a measure of between-study variability and defines the area in ROC space where we are confident that a test performs within a stated degree of uncertainty. Cochrane reviews can depict prediction regions with coverage probabilities of 50%, 90% or 95% of where a future test accuracy study would lie. The 50% region corresponds to depicting the equivalent of an interquartile range; 95% regions often cover large areas of ROC space

Confidence intervals for likelihood ratios are generated from the SAS and Stata meta-analysis macros. Computing confidence intervals for predictive values is more complicated. The simplest approach is to use the RevMan 2x2 calculator as for deriving point estimates of the predictive values. Using likelihood ratio outputs from SAS and Stata, the RevMan calculator can convert the lower and upper confidence limits of the LR+ into lower and upper confidence limits of the PPV at a stated prevalence, and likewise lower and upper confidence limits of the LR- into lower and upper confidence limits of the NPV.

The RevMan 2x2 calculator utilises the Bayesian updating process to achieve this. That utilises the following three simple equations:

- *Equation 1* odds =probability/(1-probability)
- *Equation 2* post-test odds = pre-test odds x likelihood ratio
- *Equation 3* probability = odds/(1+odds)

Beginning with a specified pre-test probability of disease (prevalence), it is converted into a pre-test odds (equation 1), then multiplied first by the point estimate of the positive likelihood ratio, and then the upper and lower confidence limits of the positive likelihood ratio obtained from the SAS or Stata meta-analysis macros to give values for the positive predictive value and its confidence interval in terms of odds (equation 2). Odds are then converted into probabilities (equation 3). Multiplication by the negative rather than the positive likelihood ratio gives estimates of 1-NPV (the probability of having disease if you test negative).

Box 1: Interpretation of CI and P values for single estimates and comparisons of test performance: Rapid diagnostic tests for uncomplicated P.Falciparum malaria in endemic countries (Abba 2011)

Test Type	Pooled sensitivity	Pooled specificity
Test 1: HRP2 antibody based tests	94.8 (93.0, 96.1)	95.2 (93.2, 96.7)
Test 4: pLDH antibody based tests	91.5 (84.7, 95.3)	98.6 (96.9, 99.5)
Difference (test 1- test 4)	P=0.20	P<0.001

Test 1 has an estimated average sensitivity of 95%. We are 95% confident that the true value of sensitivity lies between 93% and 96%.

Test 1 has an estimated average specificity of 95%. We are 95% confident that the true value of specificity lies between 93% and 97%.

Test 4 has an estimated average sensitivity of 92%. We are 95% confident that the true value of sensitivity lies between 85% and 95%.

Test 4 has an estimated average specificity of 99%. We are 95% confident that the true value of specificity lies between 97% and 100%.

Difference in average sensitivity test 1 and test type 4: Test 1 detects on average 3 more cases out of every 100 people with disease (94.8 - 91.5 = 3.3) compared to test 4. This difference is not statistically significant (p=0.20).

Difference in average specificity test 1 and test 4: Test 1 gives on average 4 more false positive diagnoses out of every 100 people without disease (98.6 - 95.2 = 3.6) compared to test 4. This difference is statistically significant (p<0.001).

It is important to recognise that only uncertainty in the estimation of test accuracy and not uncertainty in the actual pre-test probability (prevalence) of disease in the target population is captured by these computations. This uncertainty might best be explored by computing point estimates and 95% confidence intervals for predictive values across a range of plausible prevalence estimates.

Authors should note the potential for readers to confuse the interpretation of confidence intervals (CIs) associated with ratio measures (relative risk, odds ratios) and the interpretation of CIs of point estimates such as sensitivity and specificity or positive and negative predictive values (Zhelev 2013). In particular, systematic reviews commonly include outcome measures which are ratios for which an associated CI including 1 is interpreted as there being no evidence of a difference between interventions being compared. Test accuracy reviews more commonly have point estimates of accuracy such as sensitivity, specificity and predictive values as outcome measures where a CI including 1 is interpreted as evidence that a test may have perfect accuracy. Authors should therefore consider supplementing numerical presentation of uncertainty (CIs) with verbal explanations and a normalised frequency presentation format (see box 1 below), particularly if a test accuracy review includes both estimates of performance of single tests and a comparison of test performance.

When interpreting CIs associated with comparisons, authors are reminded that CIs that do not overlap can be assumed to represent statistically significantly differences. Confidence intervals that do overlap may or may not be statistically different and P values will be required to draw conclusions about statistical significance. Interpretation of CIs is further complicated in comparisons of test accuracy because differences in the size of diseased and non diseased populations usually result in wider CIs for sensitivity than for specificity. Further, for indirect test comparisons, differences in the variance of study samples included in each test group will affect the width of CIs. Interpretation of a test comparison is illustrated below in Box 1.

11.5 Heterogeneity

Heterogeneity exists when the estimates of test accuracy vary between studies more than would be expected from within-study sampling error alone. This is common in diagnostic test accuracy reviews. When this occurs there are two aspects that are important: 1) identifying the heterogeneity, 2) describing and reporting investigations of its impact on the interpretation of results (subgroup analyses).

11.5.1 Identifying heterogeneity

The starting point for investigation of heterogeneity in DTA reviews often is through visual assessment of study results in forest plots and in ROC space. Visual inspections may be useful in giving a review author an overall impression of patterns in study results, but should be supported by further rigorous statistical analysis.

As forest plots depict estimates with associated confidence intervals it is possible to discern the presence of high levels of heterogeneity where there is little overlap in the confidence intervals from different studies. Where study results differ but confidence intervals have considerable overlap it is more likely that differences between studies are explained by by sampling variation. However,

heterogeneity evident in forest plots can sometimes be caused by variation in thresholds between studies. Sorting the paired forest plots by either sensitivity or specificity will give a graphical impression of whether this is likely, if the reverse trend in estimates in specificities is seen to that in sensitivities. However, review authors are cautioned against drawing strong conclusions, as it is not possible to reliably quantify the degree of overlap of confidence intervals that is expected by sampling variation by visual inspection.

Apparent heterogeneity may also be seen in plots of study results in ROC space. However, such plots rarely include confidence intervals, making it impossible to judge whether differences between studies are within the bounds of what is expected by chance or caused by real differences between studies. Whilst there is an option in RevMan to display confidence intervals for each study on a forest plot, practically this is only helpful when there are few studies. As diagnostic accuracy studies typically contain fewer patients with the target condition than without, estimates of sensitivities are often made with less certainty than estimates of specificity, which means that review authors should expect the play of chance to cause greater spread of study points in a ROC plot in the vertical direction than the horizontal direction. All these complications make it difficult to reliably discern whether the scatter of points in ROC space demonstrates real heterogeneity.

As described in Chapter 10, the bivariate and HSROC hierarchical models are both random effects models, and include estimates of the between study variance observed in the meta-analysis. The bivariate model produces estimates of variance in logit sensitivity and logit specificity, the HSROC model provides estimates of the variance of the log DOR and the threshold parameter. These variance parameters are synonymous to the τ^2 parameter in DerSimonian and Laird random effects meta-analyses in meta-analyses of interventions, but have no easy interpretation, as they report numbers computed on unfamiliar log odds scales. However, if these numbers are estimated to be statistically significantly greater than zero, review authors can conclude that statistically significant heterogeneity exists, even if they cannot quantify how much heterogeneity exists in a helpful way.

The magnitude of the heterogeneity is depicted by plotting the prediction region in ROC space, centred on the average operating point as described in Section 11.4.4. Whilst the confidence region depicts uncertainty in the overall average value caused by sampling variability, the prediction region depicts variation from between study heterogeneity. Where heterogeneity is high, review authors will note that the 95% prediction region is much larger than the 95% confidence region. Prediction regions also take account of correlations in variation in sensitivity and specificity, and variation in positivity threshold.

No equivalent to the I^2 statistic is currently available for DTA meta-analysis. Computing separate I^2 statistics for sensitivity and specificity fails to account for variation explained by threshold effects, and the correlation of sensitivity and specificity, and will over estimate the degree of heterogeneity observed.

11.5.2 Investigations of sources of heterogeneity

Investigations of heterogeneity aim to assess whether test accuracy varies according to the characteristics of the participants, settings, tests, reference standards and other methodological features of the study design. Heterogeneity investigations should be undertaken by meta-regression models created by adding covariates to the HSROC or bivariate models as explained in Chapter 10. These models estimate the differences in accuracy between subgroups (or the association of

accuracy with a continuous measure) and formally test the statistical significance of the differences and associations. Findings from heterogeneity analyses are often graphically presented in sROC plots displaying average sensitivity and specificity points or summary ROC curves for each subgroup.

Care should be exercised in interpreting the findings of heterogeneity investigations. There are several points that should be considered:

First, heterogeneity investigations based on small numbers of studies are unlikely to produce useful findings. The statistical power of a comparison depends on the number of studies, as well as the precision of the estimates within each study, and will be lower where the characteristic is unevenly distributed across groups. In such circumstances it is possibly that important differences may be missed. When statistically significant differences are found, caution should be observed if they are based on evidence with only one or two studies in one subgroup, as the finding may be coincidental or explained by other features.

Second, exploratory heterogeneity investigations are less trustworthy than those that were pre-specified in the protocol. Exploratory analyses are often data driven prompted by observations made in informal data analyses rather than true independent tests of research hypotheses. True pre-specification of investigations in systematic reviews is difficult as the review authors are often aware of the findings of a number of the included studies before they commence the review.

Third, obtaining a spurious significant finding increases with the number of investigations which are undertaken. This occurs for pre-specified hypotheses, but is obviously worse with exploratory analyses which are data driven. There is no formal guide to how many investigations should be undertaken (and it would be inappropriate to not investigate important pre-specified factors according to some arbitrary numerical rule) but the number of hypotheses investigated must be borne in mind when interpreting the significance of the findings. Adjustments to P-values using rules for multiple testing are not encouraged as they will be overly conservative due to the inevitable correlations between the factors investigated.

Fourth, subgroup findings which have the greatest credibility are those that have a scientific rationale. Ideally selection of characteristics for investigation should be motivated by biological, clinical and methodological hypotheses supported by evidence from other sources. Subgroup analyses based on characteristics which are implausible or irrelevant are not likely to be useful and should be avoided.

Fifth, only characteristics that can be assessed at a study level should be investigated. Relationships with patient level factors (for example, gender, age, or severity of presentation) are not suited for investigation as sources of heterogeneity, as only aggregate statistics (for example, the mean age or proportion female) can be utilised in the analysis. This may mean that important relationships are missed. For example, if there is a relationship of accuracy with age, but all the studies included in the meta-analysis have similar mean ages, no relationship can be detected. This problem is variously known as aggregation or ecological bias. It is even possible that the opposite relationship is seen between aggregate values as seen within each study (in which case it is known as the ecological fallacy). Heterogeneity investigations are therefore best suited to investigating factors which are the same for all participants within each individual study.

Sixth, it must be remembered that subgroup comparisons are observational, and suffer the same limitations as all interpretations of observational findings. This includes difficulties in concluding causal relationships, and problems with confounding between characteristics. If a feature is observed to relate to test accuracy, it may possibly be the cause of heterogeneity, or it might be caused by second feature which is correlated with test accuracy. For example, if the reference standards used for assessment have varied over time, and there have also been changes to the patient groups, both may show a relationship with test accuracy, but it will not be possible to identify which, if either, is the cause. Multivariable analysis simultaneously investigating multiple sources of heterogeneity is usually infeasible due to the restricted number of studies available

Finally, it is important to establish and report whether differences between subgroups are statistically significant. P-values are produced by meta-regression analysis as explained in Chapter 10, and should be reported in the review text and tables.

Many review authors discover that their plans for investigating heterogeneity are infeasible, either because of there being too few studies, or because studies do not report the characteristics which they planned to investigate. Cochrane DTA reviews contain a dedicated section to describe how the review differs from the protocol, where these issues can be described.

11.6 Qualifying the evidence

The quality of a body of evidence should be assessed in the light of the intended use of the tests investigated. In this section we focus on the strengths and weaknesses of the primary evidence, the strengths and weaknesses of the systematic review and the consequences these may have for the interpretation of the review's results and conclusions. If authors are aware of factors that potentially limit or bias the results of their review, these should be pointed out to readers.

Some factors may decrease and others may increase the strength of the evidence in a review. Although no subheadings are provided for these factors, authors should consider using the following third-level subheadings, when appropriate: 'Strengths and weaknesses of the included studies' and 'Strengths and weaknesses of the review process'.

Authors should be aware that in this section they are expected to discuss the strengths and weaknesses of the review with regards to estimation of accuracy, and not the strengths and weaknesses of the evidence with regards to policy making decisions which would rely on other properties of the test, including its impact on patient outcomes and cost.

11.6.1 Strengths and weaknesses of included studies

Authors should present summary statements about the characteristics, quantity, quality, consistency of findings, and applicability of the studies included in the review. It is important to highlight the strengths of the evidence as well as its potential limitations. When the review contains many large studies with very similar results, this may be mentioned as reinforcing the strength of the evidence. For comparative questions, the evidence will be stronger if all results were obtained in fully paired (within-study) or randomised comparative accuracy studies, and if the superiority of one test over another is consistent across included studies (see section 11.4.3 above).

Limitations of included studies should be summarised with reference to each of the four quality domains in QUADAS 2 (patient selection, index test(s), reference standard, and flow and timing), highlighting those items particularly relevant to the review question as reflected by the tailoring of QUADAS 2 to the review topic. Detailed discussion of the types of bias that might occur in test accuracy studies can be found in Chapter 8. Authors should be mindful of the fact that readers of DTA reviews are likely to be less familiar with the types of bias that are encountered in test accuracy research (Zhelev 2013) and descriptions of the mechanisms underlying important, potential sources of bias may facilitate understanding. RevMan produces figures and graphs that summarise quality at domain level. The use of summary scores for studies across all domains is discouraged due to the topic-specific nature of quality assessment of test accuracy studies. Making an assessment of the impact of bias on estimates of accuracy can be challenging. Assessment should include consideration of the relative importance of the four QUADAS 2 quality domains to the review topic and the proportion of studies at risk of bias. For example index tests that require subjective interpretation (such as imaging tests) will be at greater risk of review bias in the index test domain compared to more objective tests.

11.6.2 Strengths and weaknesses of the review process

Limitations of the review process include shortcomings in the search strategy, in the selection process, in data-extraction and in the analyses. Authors may not have been able to conduct their review as originally intended in the protocol. This section should point out the potential implications of these limitations for the strength of the conclusions.

11.6.2.1 Limitations of the search strategy

If any search filters have been used (despite the recommendations made elsewhere in this Handbook), then this should be addressed as a shortcoming and as a potential source of biased results. The potential for bias caused by failure to retrieve or to translate articles should be discussed bearing in mind any information that might be available on the numbers and characteristics of studies affected (for example setting, index test type, size of study).

Even with the most refined search strategy, reporting bias may occur. This could include selective reporting of results within a study (e.g. reporting an optimal threshold), selective publication of studies, or language bias. The exact mechanisms behind publication bias or reporting bias for diagnostic test accuracy studies are not yet clear, and we do not know the likely impact of these forms of bias on the results. Furthermore, we do not have good ways of testing for publication bias (see section 10.6.3). Authors may however have an idea about these mechanisms in their particular clinical area and informing the reader about potential reporting bias occurring in this way may be relevant. For example, authors with knowledge about clinical testing pathways may be in a position to comment on the potential for selective reporting of results in studies that describe included patients undergoing ‘multiple tests’ as part of their routine care but report the results of only some or one of these tests.

The successful identification and inclusion of unpublished studies can be regarded as a strength of the review process.

Lack of consensus in the selection of included studies is another potential limitation of the review process. If there has been substantial disagreement between review authors about inclusion of studies, there is a risk of including less appropriate studies or of excluding studies that are more

difficult to extract data from. Although the effects of these shortcomings may be limited, they are still potential sources of bias.

11.6.2.2 Quality assessment and data extraction

Poor reporting in primary studies may limit assessment of the methodological quality of included studies. The potential impact of 'unclear' assessments of methodological quality items will depend on the importance of that particular quality item for the judgement of risk of bias for each of the four QUADAS 2 domains, reflected by tailoring of QUADAS 2. Where reviews include direct (within-study) comparisons of test accuracy, authors should highlight the fact that methods for quality assessment of directly comparative test accuracy studies are under development (see chapter 8). Poor reporting in primary studies may also limit data extraction to inform judgements about applicability. The resource implications and potential benefits of DTA review authors contacting study authors is currently unclear. Although contacting study authors is not a requirement of the DTA review process, the successful identification of additional data can be regarded as a strong point of the review process.

11.6.2.3 Limitations in the review analyses

Although meta-analysis in general may result in more precise estimates than those of the original studies, a small number of included studies with limited numbers of patients may still jeopardize the precision and applicability of the results of the review. This especially holds when substantial heterogeneity is seen, and when sources of heterogeneity cannot be explored, let alone explained.

With respect to heterogeneity and spectrum effects, authors should make *a priori* hypotheses about possible differences in accuracy between subgroups. The Discussion section is the place to put these differences in context. Authors may be in a position to discuss how consistent results are across different clinical settings and in different groups of individuals. This allows readers to make an assessment of the transferability (or applicability) of the results and whether summary estimates of test accuracy need adjustment before application in different clinical settings. If authors find apparent differences in accuracy between subgroups, they must decide whether or not these effects are credible and relevant to readers. These differences need to be considered carefully, as chance variation may always play a role. Furthermore, authors need to keep in mind that Cochrane reviews are written for an international audience, and the discussion should not be limited to the applicability of results to any single setting.

11.6.2.4 Within and between study comparisons

Comparisons of index tests are preferably answered using within-study comparisons, where all index tests have been evaluated in the same population and verified using the same reference standard in individual primary studies (see 11.4.3 above), or in studies where participants have been randomised to alternative index tests. However direct test comparisons are relatively rare and more commonly between-study (indirect) comparisons of index tests (comparisons of separate sub-groups of single test evaluation studies) are undertaken as part of a test accuracy review. When interpreting between-study comparisons, authors should be mindful of the potential for confounding due to differences in population characteristics, reference standards and study design. For example studies evaluating test A may be characterised by participants with more severe disease compared to studies evaluating test B, resulting in spurious overestimation of the sensitivity of test A relative to test B.

Although within study comparisons reduce the potential for confounding authors should be mindful of the potential for systematic differences in populations recruited to single test evaluations and populations recruited to within study comparisons of multiple tests. For example populations recruited to studies where they will receive multiple tests may be characterised by greater diagnostic uncertainty compared to populations recruited to studies where they only receive one test. This has implications for the applicability of review findings.

When both direct and indirect comparisons are included in the same review these should be distinguished and authors should explicitly discuss any differences in results between them. In addition to bias associated with between study comparisons (Takwoingi 2013). Authors should acknowledge that estimates of test accuracy derived from between study comparisons are typically based on a greater number of studies than within study comparisons which results in between study comparisons having greater precision and more statistical power to detect any differences between tests.

11.6.2.5 Comparison with previous research

Although a heading "Previous Research" does not exist in Review Manager, authors are advised to put their research in the context of what other reviews have shown. In many cases, a Cochrane systematic review may not be the first review on the study question. It may also be possible that other systematic reviews have been published in the Cochrane Library for the same test, but for different yet related target conditions. If so, the authors should discuss any differences in quality and results between their review and the previously published reviews. Sometimes, multiple (related) reviews may stem from one generic protocol. If that is the case, authors should mention this and put their particular review in the context of the other reviews.

If a review is based on an update of a previously existing review, the authors may want to point out any essential differences in the results from those in the previous review, in particular if test features or important technical aspects of an index or reference test have changed over time, or if, for example, the use of an index test is extended to a new target population.

11.7 Applicability of findings to the review question

Test accuracy estimates generated by random effects meta-analyses are average estimates across included studies. Authors need to discuss the applicability of the results of the review (i.e. the degree to which the studies in the review correspond to the review objectives). For intervention (or treatment) reviews, this is described as 'directness' or 'indirectness': the extent to which a review is relevant for the purpose to which it is being put (Higgins and Green 2009). For test accuracy studies the applicability of findings has been described as the degree to which they are transferable to different settings (Irwig, 2002). Assessment of applicability is particularly important for DTA reviews because of the degree to which setting, patient spectrum, index test and reference standard characteristics can affect test accuracy estimates. Therefore, estimates and judgments of their applicability should be discussed with reference to this.

Two scenarios can be distinguished with different implications for assessing the applicability of review findings. A DTA review question may have broad inclusion criteria, which complicates investigation of heterogeneity but allows exploration of variation in accuracy across various settings,

different patient groups or variations in index test application. Alternatively, narrow review inclusion criteria simplifies the investigation of heterogeneity but restricts the applicability of review findings.

The QUADAS 2 tool includes an assessment of 'concern about applicability' for 3 of the 4 domains: Domain 1: patient selection; Domain 2: index test(s) and Domain 3: reference standard. Concerns are rated as 'high', 'low' or 'unclear'. These judgements of applicability should be made with reference to the stated review question (see chapter 8).

QUADAS 2 Domain 1: patient selection

The accuracy of a diagnostic test may vary depending on the clinical spectrum of participants included in the study. Ideally the spectrum of participants should be as similar as possible to the intended population as phrased in the review question with respect to demographic features, co-morbidities, severity of the target condition, presentation (for example symptomatic or asymptomatic and in what setting), and tests done before the index test(s). Authors should assess the implications of the inclusion and exclusion criteria of the studies included in the review, the prevalence of the target condition seen in the individual studies, and the clinical setting of the studies as specified by the review question. Severity of disease will have an impact on sensitivity (since estimates of sensitivity are expected to increase with increasing disease severity) and the range of differential diagnoses present in non-diseased populations will affect specificity (with an increasing number of differential diagnoses the number of false positives is likely to increase and specificity will therefore decrease). Differences in estimates of accuracy across different settings suggest non transferability of the average estimate. Conversely, if the results are consistent across clinical settings, this suggests that the findings are robust and transferable.

QUADAS 2 Domain 2: index test(s)

The index test(s) used in the studies included in the review may vary. For example, different versions of the test may have been used, the test may have been conducted on different types of specimens, operators may have had different experience and skills in using the test, and different test positivity thresholds may have been used to determine diseased and non-diseased states. Specifying a common threshold is straightforward when it has a numeric value but more difficult when it is based on a more subjective judgement, as will occur for elements of history taking and the physical examination and often in image interpretation. Where thresholds are more subjective or dependent on operator skill, interpretation of the estimate of average test accuracy presumes that the distribution of implicit thresholds being used by those interpreting test results in the studies is representative of implicit thresholds that will be used in practice. Authors need to consider whether this is likely, and whether primary study reporting will allow readers to understand any differences between test interpretation in practice and that used in included studies. Even with explicit and objective thresholds, differences between centres and laboratories in the interpretation of tests will introduce variability. Authors should also consider whether there are reproducibility and calibration issues with particular tests which could mean that numerical thresholds will not transfer from study populations to intended settings.

QUADAS2 Domain 3: reference standard

The applicability of the reference standard depends on how closely the definition of the target disorder used in the review question compares to the definition adopted in included studies. For example, heart failure is diagnosed on the basis of the fraction of blood ejected from the heart during contractions (the ejection fraction). Ejection fraction thresholds used to define heart failure can vary between 30% and 50% in test accuracy studies and therefore these thresholds may not correspond to those adopted in practice.

11.8 Summary of findings (SoF) tables

The aim of SoF tables is to communicate key information in a quick and accessible format. At the time of writing, experience of compiling SoF tables for DTA reviews is still limited, and the difficulty of summarizing their findings is compounded by the fact that test accuracy evidence is often unfamiliar to decision makers. Current guidance is to adopt a simple format and to allow author discretion, encouraging a degree of innovation which will lead to more precise guidelines in future years. Authors might also find it useful to consult the work of the GRADE working group who are currently developing a system for grading the strength of recommendations for DTA questions (Schunemann 2008) and examples of SoF tables in published reviews. SoF tables should be placed at the beginning of the main discussion section of the review as a means of facilitating interpretation of results and so that general statements in the conclusions are linked to and supported by data in the results section of the review.

In a similar manner to SoF tables in intervention reviews, it is important that the main findings of a DTA review are presented in a transparent and simple tabular format. The SoF table should provide key information on the accuracy of the index tests under consideration (and difference in accuracy where tests are being compared) and important limitations arising from the assessment of the quality and applicability of evidence.

Ideally, DTA reviews would be expected to have a single SoF table but reviews may have to include more than one, for example if the review addresses both the accuracy of an index test individually and its accuracy compared with other tests.

Review authors have to create their own table in RevMan. The table function in RevMan allows review authors to add many extra rows and columns as necessary, and merge and split cells. Authors do not have to be constrained by the limited number of rows and columns initially offered.

Authors should include a clear statement that the SoF table for the diagnostic test cannot be safely interpreted in isolation from the original data presented in the main body of the review.

11.8.1 SoF template

Requirements for the Summary of Findings table are as follows:

1) The review question and its components (population, setting, index test(s), including role and purpose, and reference standard) should be stated in full at the head of the table
2) The summary at the head of the table should also flag up any limitations noted arising from the assessment of risk of bias and applicability, or excessive heterogeneity.

3) Separate rows in the table should be used to represent each index test or variants of the index test, such as different test positivity thresholds
4) Estimates of the accuracy of each index test may be conveyed as a meta-analytic summary and by illustrating the range of accuracy estimates across included studies. The information provided for each test should, as a minimum, convey:
    a) The number of participants (median and range or equivalent) and number of studies contributing to the estimate of accuracy;
    b) Estimates of test accuracy generated by the review. This should be expressed in terms of sensitivity and specificity although other presentation formats (e.g. normalised frequencies) may also be included to improve the accessibility of evidence to users (see 11.4.1.3 above).
    c) The statistical uncertainty associated with the summary measure of test accuracy (e.g. 95% confidence intervals expressed as proportions or using normalised frequencies).
    d) Information on the prevalence of the disease, either from the studies included in the review (preferably reported as median and interquartile range) or from other external sources. The predictive value and interpretation of the test will depend on prevalence.

5) Index test comparisons - in addition to the main principles outlined above, review authors should consider stating the following in SoF tables that include index test comparisons:
    a) Number of primary studies (and patients) contributing to direct comparisons and those contributing to indirect comparisons
    b) Estimates of test accuracy for each of the tests with measures of statistical uncertainty, if possible with estimates of the absolute difference in accuracy between tests.
    c) P-values for the comparison of the accuracy of the tests enabling a reader to distinguish differences which are explicable by chance from those where there is evidence that they relate to real differences in accuracy.

There may be additional, optional aspects of the results which review authors may think sufficiently important to include, for example, variation in test accuracy estimates by cut-off, pre-test probability (prevalence) or any other covariate; or the proportion of test failures or indeterminate test results

Authors should be careful to interpret applicability only in the context of original data (i.e. those provided by included studies) and not to extrapolate beyond observed data. Authors may want to highlight aspects of their assessment of applicability that are considered of particular importance or are common in included studies.

Indicating the consistency/inconsistency of test accuracy results from one included study to the next in a summary table is particularly difficult However, given that heterogeneity is virtually always present in DTA reviews, and it is rarely possible to explain more than a small proportion of this, it is critical to the validity of the SoF table that unexplained heterogeneity is clearly acknowledged. Notes or comment sections in the table may be the best way to do this.

Figure 5: What is the diagnostic accuracy of the Platelia© Aspergillus test for invasive aspergillosis for different cut-off values? (Leeflang 2008)					
Patients/population		Immunocompromised patients, mostly haematology patients			
Prior testing		Varied, mostly physical examination and history (fever, neutropenia)			
Settings		Mostly inpatients in haematology or cancer departments			
Index test		Platelia© Aspergillus test, a sandwich ELISA for galactomannan, an *Aspergillus* antigen			
Importance		Depends on the time-gain the test may provide			
Reference standard		Gold standard would have been autopsy, but this is virtually never done. Actual reference used: clinical and microbiological criteria			
Studies		Cross-sectional studies including an equally suspected patient sample (case-control studies) were excluded. Studies had to report cut-off values that were used (n = 29). Each study can be present in more than one subgroup.			
Test / Subgroup	Summary accuracy (95% CI)	No. of participants (studies)	Prevalence Median (range)	Implications	Quality and Comments
Cut-off 0.5	Sensitivity 0.79 (0.61-0.93) Specificity 0.82 (0.71-0.92)	901 (7)	9.9% (0.8-34%)	With a prevalence of 10%, 10 out of 100 patients will develop IA. Of these, 2 will be missed by the Platelia test (21% of 10), but will be tested again. Of the 90 patients without IA, 15 will be unnecessarily referred for CT scanning.	Low numbers of diseased patients per study (1 to 20). These studies contained a representative spectrum. Uninterpretable results and withdrawals poorly reported.
Cut-off 1.0	Sensitivity 0.71 (0.61-0.81) Specificity 0.89 (0.80-0.97)	1744 (12)	12% (0.8-44%)	Of the 10 in 100 patients developing IA, 3 will be missed. Of the 90 patients without IA, 10 will be referred for CT scanning.	Low numbers of diseased patients per study (1 to 34). These studies contained a representative spectrum. Uninterpretable results and withdrawals poorly reported.
Cut-off 1.5	Sensitivity 0.62 (0.45-0.79) Specificity 0.95 (0.92-0.98)	2600 (17)	7.4% (0.8-34%)	Of the 10 patients with IA, 4 will be missed. Of the 90 patients without IA, 5 will be referred for CT scanning.	Low numbers of diseased patients per study (1 to 17), except one (98 patients). These studies contained a representative spectrum.
CAUTION: The results on this table should not be interpreted in isolation from the results of the individual included studies contributing to each summary test accuracy measure. These are reported in the main body of the text of the review					

11.9 Conclusions

At this point in the review, the results of the meta-analysis, the quality of the evidence, and its applicability to the review question have been considered. The next step is to explain to readers how these results can be used to draw conclusions. Conclusions in Cochrane DTA reviews are divided into 'Implications for practice' and 'Implications for research'.

11.9.1 Implications for practice

Implications for practice should be as practical and unambiguous as possible. They should not go beyond the evidence that was reviewed and should be justifiable by the data presented in the review. In addition, the decision to use a diagnostic test will often be based on the accuracy of other tests not included in the review, and on evidence about the effectiveness of downstream actions, such as starting or withholding therapy to patients with specific test results.

Recommendations that depend on assumptions about resources and values should be avoided. A common mistake is for authors to confuse facts and judgements. For example, if summary sensitivity is 80%, this is a fact, but to describe this subjectively as 'high sensitivity' would be a judgement.

Because Cochrane reviews have an international audience, the implications for practice should, as far as possible, assume a broad international perspective, rather than addressing specific national or local circumstances. Authors should be particularly careful to bear in mind that different people might make different decisions based on the same evidence. The primary purpose of the review should be to present information rather than to offer advice. The implications for practice should help readers understand the implications of the evidence in relationship to practical decisions.

The inability of DTA studies to inform direct conclusions about impact on patient outcomes, cost, and cost-effectiveness is clear. However, evidence about test accuracy, and particularly comparative test accuracy, may in some circumstances provide adequate information to dictate practice (Lord 2006). For example, if one test is found to have superior accuracy to another and known to have no drawbacks (for example, by being less invasive, cheaper, quicker, and easier to deliver), the review may provide adequate evidence to support its use. In contrast, estimates of accuracy for some tests may be so poor as to indicate that conclusions can be drawn that a test has no useful role in diagnosis.

In other situations the implications for practice may not be so clear. For example, if one test has higher sensitivity but lower specificity than another, which test creates the best patient outcomes will not be discernible from evidence about test accuracy alone. The consequences of false positives and false negatives need to be known, and their trade off assessed. Sometimes the difference is large enough to be clear as to which test is best, other times deducing the best test will require decision modeling. In this case, it should be made clear that whilst a Cochrane DTA review will provide important information to use in the model, it cannot answer the question by itself. Similarly, if there are other aspects of the test that differ; for example its invasiveness, cost, speed, acceptability, and ease of delivery; the relative importance of accuracy compared to these other features should be assessed. Again this will require research outside of the study, perhaps synthesized in a decision model, or even requiring a randomised controlled trial comparing testing strategies. In such circumstances, the review authors should point out the key factors involved in

making an assessment of the value of the test, and the further research which would be needed to provide an answer under future research recommendations.

We suggest review authors think through the following issues when considering implications for practice:

How is the test positioned in the clinical pathway? The position of the index test in a clinical pathway will dictate the absolute numbers of true positives, false positives, true negatives, and false negatives and the downstream consequences of these. Tests used early on in a clinical pathway where pre-test probability (prevalence) of the target disorder is at its lowest (for example screening asymptomatic individuals) are likely to result in a relatively high absolute number of false positives at any given specificity and a relatively low absolute number of false negatives at any given sensitivity.

Increasing prevalence of the target is likely to decrease the absolute number of false positives at any given specificity and increase the number of false negatives at any given sensitivity. The effect of prevalence of the target condition on the absolute number of test errors is only communicated by a limited number of test accuracy measures (the 2x2 diagnostic contingency table; predictive values at a specified prevalence; normalised frequencies).

How does the index perform in relation to its intended role (add, replace, triage)? Chapter 3 outlined the different roles that index tests might have in a clinical testing pathway: should an index *replace* another test?; should an index test be *added* to another test?; should the index test be used to *triage* individuals to receive another test? (Bossuyt 2003). Authors need to consider whether the performance characteristics of the test(s) evaluated in the DTA review are consistent with its intended role in practice. For example, a new test intended to triage individuals for a more invasive test is likely to have to demonstrate evidence of comparable sensitivity to the more invasive test to ensure that introduction of the new test will not result in an increase in the number of false negatives. However if the new test has a lower specificity this may be acceptable as the result will still be a reduction in the number of individuals having to undergo the more invasive test.

There may be issues in addition to accuracy that should be highlighted by authors when considering the consequences of introducing the test in its intended role (Ferrante di Ruffano 2012).For example for a test replacement question, consideration of whether test A should replace test B may not rely solely on evidence of improved accuracy of test A compared to test B. If test B is less costly or less invasive than test A, then evidence of comparable accuracy of the two tests may be considered sufficient.

What are the consequences for patients of true positive, false positive, false negative or true negative test results? The importance of false negatives will depend on the consequences of missing the target condition being tested for and the importance of false positives will depend on the consequences of inappropriate treatment or testing. For example, a false negative result may be more relevant when a test is used for screening asymptomatic individuals for cancer (negative test results will not receive further follow up) then when testing hospital patients for suspected infection (negative test results will continue to be observed and re-tested if symptoms do not resolve). Similarly, false positives will have less relevance when index test-positive patients are referred to undergo unnecessarily another, non-invasive test (e.g. 24-hour blood pressure monitoring after a

single raised blood pressure measurement) than when false positives undergo treatments with adverse side effects or receive a stigmatising diagnosis (e.g. a false positive diagnosis of mental illness).

What are the consequences of introducing the index test(s), for the intended use, in the intended role, for patient outcomes? The consequences of introducing the index test(s) into practice require comparison with other tests – prior tests or additional tests – that are available for the same intended use. In addition to the implications of test errors associated with the intended use and the role of the index test, review authors may want to highlight other issues that may be relevant for readers in reaching a decision about the potential benefits and harms of adopting the index test(s). Costs, organisational outcomes, test reliability, uninterpretable results, acceptability, uptake, and direct harms and benefits of tests (physical and psychological) are examples of such issues (Ferrante di Ruffano 2012). Information on these issues may come from other (not systematically searched and assessed) sources. Authors need to be cautious when providing additional information that has not been gathered in a systematic way and that may therefore be dependent on interpretation, or even be prone to bias. However, information about test features can assist decision-makers assessing the value and use of the test and should be provided where available.

Authors need to bear in mind that the consequences (dimensions outlined above) of introducing index test(s) in practice will vary greatly between countries and that the costs in their local area may not be generalisable to other localities.

11.9.2 Implications for research

Implications for research may cover two broad areas: additional research needed on aspects of tests beyond their accuracy, and further studies of test accuracy which need to be undertaken. Research funders and commissioners will read this section of the review to help inform future funding decisions.

As discussed in 11.9.1, evidence about other routes by which tests impact on patients may need to be taken into account to assess the potential of tests to improve patient outcomes in clinical practice. Review authors should specify the characteristics of the tests which need assessment and the method by which this should be done (i.e. from other systematic reviews, routine data sources, new primary studies, and qualitative research). Additionally they should indicate whether decision modelling or randomised trials comparing different diagnostic pathways should be undertaken, and comment on the key design features of the studies required.

Where the evidence on test accuracy fails to be conclusive, review authors may recommend future test accuracy studies which should be undertaken. This may involve stating that similar but better reported studies are needed; or accuracy studies need to be of better quality, done in the appropriate patient spectrum, point in the care pathway, or setting; or comparative accuracy studies are needed. Please include as much detail of the design of future studies as possible – where better comparative accuracy studies are needed it is important to describe the important comparisons which need to be made. If there are particularly issues that need to be sorted before future accuracy studies are commissioned, such as obtaining consensus on the best reference standard or the method by which a new index test is delivered, projects to achieve this should be described.

If the review revealed new questions or generated hypotheses, then this is the place to discuss them. This may be particularly important where reviews have obtained insights through investigations of between study heterogeneity, or where heterogeneity exists but cannot be explained.

References

Abba 2011
Abba K, Deeks JJ, Olliaro PL, Naing CM, Jackson SM, Takwoingi Y, Donegan S, Garner P. Rapid diagnostic tests for diagnosing uncomplicated P. falciparum malaria in endemic countries. Cochrane Database of Systematic Reviews 2011, Issue 7. Art. No.: CD008122. DOI: 10.1002/14651858.CD008122.pub2.

Bossuyt 2006
Bossuyt P, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332:1089–92

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;58:882-93.

Evans 2000
Evans J, Handley SJ, Perham N, Over DE, Thompson VA. Frequency versus probability formats in statistical word problems. Cognition 2000; 77:197-213.

Ferrante di Ruffano 2012
Ferrante di Ruffano L, Hyde CJ, McCaffrey KJ, Bossuyt PMM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. BMJ 2012; 344:e686

Girotto 2001
Girotto V, Gonzalez M. Solving probabilistic and statistical problems: a matter of information structure and question form. Cognition 2001; 78(3):247-276.

Higgins 2009
Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2 [updated September 2009]. The Cochrane Collaboration, 2009. Available from www.cochrane-handbook.org.

Hoffrage 1998
Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. Academic Medicine 1998; 73(5):538-540.

Irwig 2002
Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. BMJ 2002;324:669-71.

Jaeschke 2002
Jaeschke R, Guyatt G, Lijmer J. Diagnostic tests. In: Guyatt G, Rennie D, eds. Users' guides to the medical literature. Chicago: AMA Press, 2002:121-40.

Leeflang 2008
Leeflang MM, Debets-Ossenkopp YJ, Visser CE, Scholten RJ, Hooft L, Bijlmer HA, Reitsma JB, Bossuyt PMM, Vandenbroucke-Grauls CM. Galactomannan detection for invasive aspergillosis in immunocompromized patients. Cochrane Database of Systematic Reviews 2008, Issue 4. Art. No.: CD007394.

Leeflang 2012
Leeflang MM, Deeks JJ, Rutjes AW, Reitsma JB, Bossuyt PM. Bivariate meta-analysis of predictive values of diagnostic tests can be an alternative to bivariate meta-analysis of sensitivity and specificity. J Clin Epidemiol. 2012 Oct;65(10):1088-97.

Lord 2006
Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med. 2006;144(11):850-5.

Puhan 2005
Puhan MA, Steurer J, Bachmann LM, ter Riet G. A randomized trial of ways to describe test accuracy: the effect on physicians' post-test probability estimates. Ann Intern Med 2005; 143(3):184-189.

Reid 1998
Reid MC, Lane DA, Feinstein AR. Academic calculations versus clinical judgments: Practicing physicians' use of quantitative measures of test accuracy. The American Journal of Medicine 1998; 104(4):374-380.

Sackett 2000
Sackett DL, Straus S, Richardson WS, RosenbergW, Haynes RB. Evidence based medicine.How to practise and teach EBM. 2nd ed. Edinburgh: Churchill Livingstone, 2000:67-93.

Schünemann 2008
Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, Williams JW Jr, Kunz R, Craig J, Montori VM, Bossuyt P, Guyatt GH; GRADE Working Group. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008;336(7653):1106-10

Steurer 2002
Steurer J, Fischer E, Bachmann LM, Koller M, ter Riet G. Communicating accuracy of tests to general practitioners: a controlled study. BMJ 2002; 324:824-826.

Takwoingi 2013
Takwoingi Y, Leeflang MMG, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Annals of Internal Medicine 2013; 158: 544-554.

Zhelev 2013
Zhelev Z, Garside R, Hyde C. A qualitative study into the difficulties experienced by healthcare decision makers when reading a Cochrane diagnostic test accuracy review. Systematic Reviews 2013, 2:32.