eunethta

EUROPEAN NETWORK FOR HEALTH TECHNOLOGY ASSESSMENT

GUIDELINE

Meta-analysis of Diagnostic Test Accuracy Studies

November 2014

The primary objective of EUnetHTA JA2 WP 7 methodology guidelines is to focus on methodological challenges that are encountered by HTA assessors while performing relative effectiveness assessments of pharmaceuticals or non-pharmaceutical health technologies.

As such the guideline represents a consolidated view of non-binding recommendations of EUnetHTA network members and in no case an official opinion of the participating institutions or individuals.

Disclaimer: EUnetHTA Joint Action 2 is supported by a grant from the European Commission. The sole responsibility for the content of this document lies with the authors and neither the European Commission nor EUnetHTA are responsible for any use that may be made of the information contained therein.

This guideline on “Meta-analysis of Diagnostic Test Accuracy Studies” has been developed by HIQA – IRELAND,

with assistance from draft group members from IQWiG – GERMANY.

The guideline was also reviewed and validated by a group of dedicated reviewers from GYEMSZI – HUNGARY, HAS – FRANCE and SBU - SWEDEN.

Table of contents

Acronyms - Abbreviations

AUC	area under the curve
DOR	diagnostic odds ratio
FN	false negative
FP	false positive
FPR	false positive rate
GRADE	Grading of Recommendations Applicability, Development and Evaluation
HSROC	hierarchical summary receiver operator characteristic
IPD	individual patient data
LR-	negative likelihood ratio
LR+	positive likelihood ratio
MESH	medical subject headings
NPV	negative predictive value
PPV	positive predictive value
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
RCT	randomized controlled trial
ROC	receiver operator characteristic
Sn	sensitivity
Sp	specificity
SROC	summary receiver operator characteristic
STARD	Standards for Reporting of Diagnostic Accuracy
TN	true negative
TP	true positive
TPR	true positive rate

Summary and table with main recommendations

Introduction

Diagnostic tests are used for a variety of purposes including to: determine whether or not an individual has a particular target condition; provide information on a physiological or pathological state, congenital abnormality, or on a predisposition to a medical condition or disease; predict treatment response or reactions; define or monitor therapeutic measures. Ideally an evaluation should be undertaken to assess the clinical utility of a test. Such an assessment is generally not supported by appropriately designed studies or by long term outcome data. In the absence of clinical utility data, diagnostic tests are evaluated on the basis of test accuracy: the ability of the test to correctly determine the disease status of an individual. Test accuracy is not a measure of clinical effectiveness and improved accuracy does not necessarily result in improved patient outcomes. A number of metrics are available to describe the characteristics of a diagnostic test, such as the sensitivity, specificity, diagnostic odds ratio, predictive values, likelihood ratios, and the receiver operator characteristic (ROC) curve. Diagnostic tests may also be subject to a threshold effect, whereby the translation of a test result into a dichotomous positive/negative result is not uniform across studies.

Problem statement

Diagnostic test accuracy may be evaluated across a number of studies; to improve the precision of the estimate, it may be desirable to combine data from a number of studies in a meta-analysis. This guideline reviews available methods for the meta-analysis of diagnostic test accuracy studies that report a dichotomous outcome, and discusses types of bias that are encountered in such meta-analyses.

Methods for meta-analysis of diagnostic test accuracy studies

The hierarchical summary receiver operator characteristic (HSROC) and bivariate random-effects techniques are considered the most appropriate methods for pooling sensitivity and specificity from multiple diagnostic test accuracy studies. Both approaches take into account any correlation that may exist between sensitivity and specificity. The two methods offer equivalent results under certain conditions, such as when no covariates are included. These two methods are considered to be more statistically rigorous than the alternative Moses-Littenberg approach.

The most appropriate choice of meta-analytical approach is context specific and also depends on the observed heterogeneity across studies, and the quantity of evidence available. The type of summary data that should be reported depends on whether or not there is a threshold effect. If a threshold effect is present and if it explains most of the observed heterogeneity, then a summary ROC curve can be presented. Alternatively, a summary point of sensitivity and specificity with corresponding confidence region should be reported.

Sources of bias

Numerous sources of bias can affect the summary estimate of diagnostic test accuracy: publication bias; heterogeneity; spectrum bias; verification bias; choice of cut-off points for dichotomising a test result. The accuracy reported in studies can also be influenced by underlying disease prevalence, dependence between combined tests, and missing data.

When conducting a meta-analysis, potential sources of bias should be identified and investigated in terms of how they influence the summary estimates of diagnostic test accuracy. Studies included in a meta-analysis should be appraised in terms of study quality and whether or not they are sufficiently equivalent to justify a meta-analysis.

Recommendations	The recommendation is based on arguments presented in the following publications and / or parts of the guideline text
1. Pooling studies of diagnostic test accuracy should only be undertaken when there are sufficient studies available. When only two studies are available, it is not recommended to undertake a meta-analysis; reporting should be restricted to a narrative description of the available evidence.	Section 2.1.6
2. The quality of studies being pooled should be assessed using a recognised and validated quality assessment tool.	Section 2.6.2
3. Pooled studies should be equivalent in terms of the index test, the reference standard, the patient population and the indication.	Section 2.1
4. Where important differences are identified across studies in terms of disease spectrum, study setting, or disease prevalence, these should be accounted for by including covariates.	Section 2.4
5. Where potential study differences occur, but cannot be readily accounted for, such as verification bias, these should be clearly identified and the potential impacts determined.	Section 2.4
6. The appropriate methods of meta-analysis are the hierarchical SROC and bivariate random effects techniques, unless there is an absence of heterogeneity in either the false positive rate or the true positive rate, in which case two separate univariate meta-analyses may be more appropriate.	Section 2.1.6
7. The appropriate approach to meta-analysis is defined with respect to the quantity of data, between-study	Section 2.1.6

heterogeneity, threshold effects, and the correlation between the true positive rate and the false positive rate.	
8. The reporting of meta-analysis should include all the information that justifies the choice of analytical approach and supports the exclusion of alternative approaches.	Section 2.2

1. Introduction

1.1. Central terms and concepts

This section describes the main concepts in diagnostic testing in terms of the test itself, and the measures used to describe the accuracy of a test. Test measures may be global (overall test performance) or specific (single aspect of accuracy), and they may be conditional (dependent on prevalence) or unconditional (independent of prevalence).

1.1.1. Diagnostic test, gold- and reference standards

Diagnostic test accuracy studies estimate the ability of a diagnostic test to correctly discriminate between patients with and without a particular target condition. To evaluate the accuracy of a diagnostic test (also called the index test), it must be compared with a reference standard test or a gold standard test.[1] A gold standard, which has perfect discriminatory power between positive and negative status, rarely exists. Hence the gold standard is typically replaced by a reference standard that approximates the gold standard as closely as possible.[1] In some cases, there may not be an appropriate reference standard. When analysing test accuracy, the same reference standard should be applied to the whole study population.

Test accuracy for a single study population that have been subject to a diagnostic test for a given target condition is generally presented in a 2x2 table indicating the test result (as positive or negative) and the true status with respect of the reference status of those tested (as positive or negative) (see Figure 1).

Figure 1. The 2x2 table

		True status	
		Positive	Negative
Test result	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

1.1.2. Sensitivity and specificity

Sensitivity and specificity are the most commonly used measures of diagnostic test performance.[2]

- Sensitivity (Sn) – the percentage of people with the target condition that are identified as having the condition by the diagnostic index test

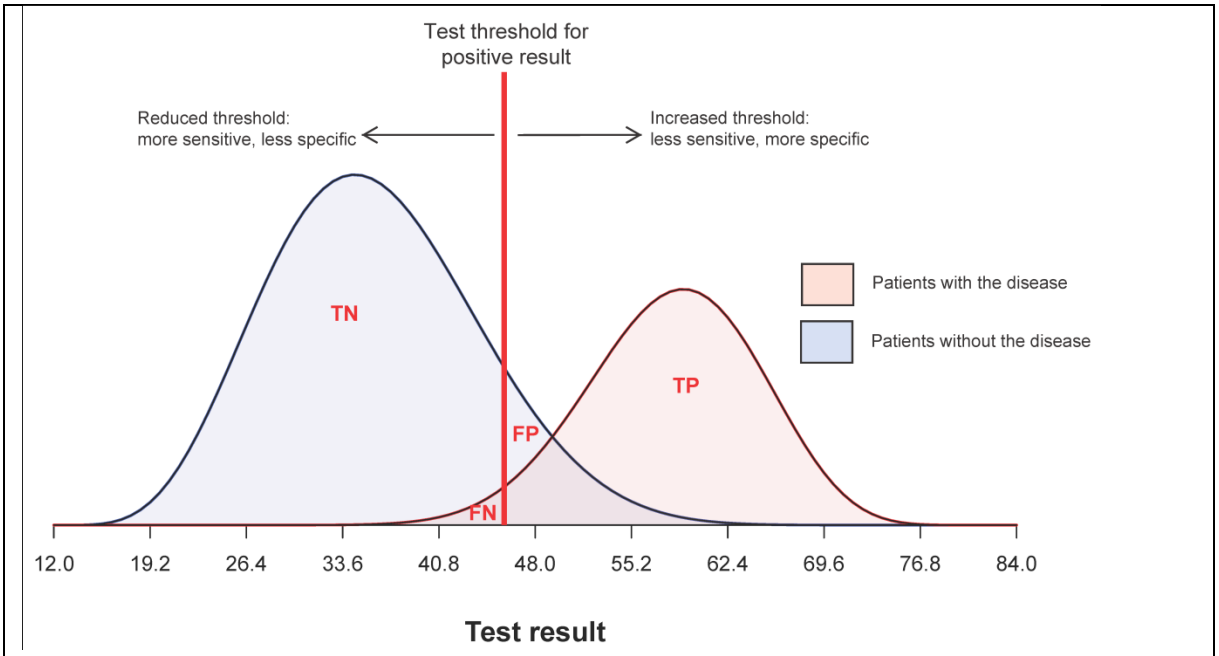
$$Sn = TP \times 100 / (TP + FN) \quad (1)$$

- Specificity (Sp) – the percentage of people that do not have the target condition that are identified as not having the condition by the diagnostic index test

$$Sp = TN \times 100 / (TN + FP) \quad (2)$$

A perfect test would have 100% sensitivity and specificity. However, in reality the two measures are almost always negatively correlated, such that increased sensitivity is associated with decreased specificity (see Figure 2). The negative correlation is often a function of the threshold beyond which a test result is considered a positive. For example, an increased threshold will result in fewer false positives (increased specificity) but more false negatives (reduced sensitivity). Different studies evaluating test accuracy may use the same test, but apply a different threshold for defining a positive test result. Decreasing the threshold decreases specificity but increases sensitivity, while increasing the threshold decreases sensitivity but increases specificity. By varying the threshold for a positive test, a correlation between sensitivity and specificity is observed which is known as a threshold effect.

Figure 2. Test threshold and impact on diagnostic accuracy



Notes: the test result values are for a hypothetical biomarker. There is an overlap between values that can be obtained for patients with and without the disease of interest. Hence perfect classification is not possible with this test.

Sensitivity and specificity are generally assumed to be independent of disease prevalence, although this is not strictly the case (see section 2.4.6). The measures have no clinical meaning and they do not apply to test results that are reported as levels rather than a dichotomous outcome. The sensitivity is also referred to as the true positive rate (TPR), while 1 minus the specificity (1-Sp) is referred to as the false positive rate (FPR).

Sensitivity and specificity are perhaps the most commonly reported measures of diagnostic test accuracy. As a concept they are relatively simple to understand. They are considered to be specific and unconditional measures of accuracy. However, sensitivity and specificity are summary test characteristics, and do not provide information about a specific patient. In other words, they provide an ‘on average’ accuracy for a given test. Sensitivity and specificity may be reported as percentages or proportions.

It is generally the case that for a test to be useful at ruling out a disease it must have high sensitivity, and for it to be useful at confirming a disease it must have high specificity.[3] The Sn-N-Out (high sensitivity, negative, rules out) and Sp-P-In (high specificity, positive, rules in) mnemonics are sometimes used to make quick diagnostic decisions, although these rules are serious simplifications and should be used with caution.[4] A high sensitivity implies very few false negatives, therefore nearly all patients labelled as negative are correctly assigned. Similarly a high specificity implies very few false positives, meaning that nearly all patients labelled positive are genuinely positive. However, a high specificity combined with a poor sensitivity may not be an informative test as many genuine positives will test negative. Ordinarily a high specificity can be used to rule in positives, but when coupled with a poor sensitivity, few genuine positives will be captured. Hence, care must be taken when interpreting sensitivity and specificity values, and both measures should be reported together when considering the accuracy of a test.

Sensitivity and specificity are sometimes presented simultaneously as Youden's Index (J_c), which is intended as a means for optimising the cut-off point (c) for a test:

$$J_c = Sn_c + Sp_c - 1 \quad (3)$$

The optimal cut-off point, c*, is the cut-off point at which J_c is maximised.[5] However, this optimisation is based on the assumption that false-positives and false-negatives are equally undesirable. In reality, the incorrect classifications of healthy and diseased persons may not be considered equally undesirable.[5] For example, for a life-threatening disease where early detection may significantly improve outcomes, there may be a preference to minimise false-negatives.

The interpretation of sensitivity and specificity can be problematic when evaluating tests that are applied repeatedly, such as for continuous monitoring of a patient's status.[6]

1.1.3. Likelihood ratios

The likelihood ratio (LR) associated with a positive test result is the probability of a positive finding in patients with the target condition divided by the probability of a positive test result in patients who do not have the target condition. Multiplying the LR by the pre-test odds of having the target condition gives the post-test odds of having the condition. The LR can be expressed for positive and negative test results:

$$\text{Likelihood ratio for positive results (LR+)} = Sn / (100 - Sp) \quad (4)$$

$$\text{Likelihood ratio for negative results (LR-)} = (100 - Sn) / Sp \quad (5)$$

As the likelihood ratios are a function of sensitivity and specificity, it is generally assumed that they do not vary with disease prevalence. Likelihood ratios can be calculated for multiple levels of test result, which can be useful in diagnostic tests for which results are presented on a continuous scale.[2] Like sensitivity and specificity, these measures are considered to be specific and unconditional measures of accuracy.

By applying the Bayes' theorem, the pre-test probability of disease (e.g., the prevalence of disease) can be converted into a post-test probability using the likelihood ratios in conjunction with the test result. As a rule of thumb, a likelihood ratio of between 0.2 and 5

gives no more than weak evidence to rule the disease in or out. A likelihood ratio of between 5 and 10, and between 0.1 and 0.2 gives moderate evidence to rule the disease in or out, respectively. A likelihood ratio of greater than 10 or less than 0.1 gives strong evidence to rule the disease in or out, respectively.[2] These ranges are only intended to provide an approximate rule of thumb and consideration must be given to the context of the results. It should also be noted that quite different combinations of sensitivity and specificity can produce the same likelihood ratio values.

1.1.4. Diagnostic odds ratio

The diagnostic odds ratio (DOR) provides a single measure of test performance that is assumed to be independent of the prevalence of the target condition.

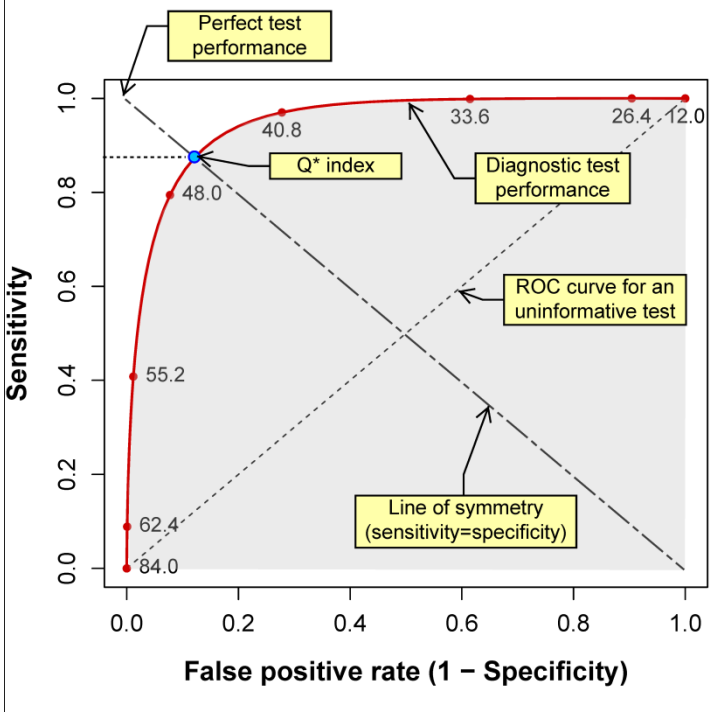
$$DOR = (TP / FN) / (FP / TN) \quad (6)$$

The diagnostic odds ratio describes the odds of a positive test results in participants with the disease compared with the odds of a positive test results in those without the disease. A single diagnostic odds ratio corresponds to a set of sensitivities and specificities depicted by a symmetrical receiver operating characteristic curve (see section 2.2).[3] The DOR is not useful in clinical practice. As it is a single measure assumed to be independent of prevalence, the DOR is referred to as a global and unconditional measure.

1.1.5. Receiver Operating Characteristic (ROC) curves

A diagnostic test may return values on a continuous scale, but this must then be converted into a dichotomous positive/negative diagnosis based on a cut-off point. The choice of cut-off point on the scale will impact on the test’s accuracy. A threshold at one extreme will result in few positives, while a threshold at the other extreme will result in many positives (see Figure 2). A ROC curve is a graphical plot used to represent the performance of a test over a range of threshold settings (see Figure 3).[3] That is, the curve shows the impact on sensitivity and specificity of varying the threshold for which a test result is labelled as a positive rather than a negative. A ROC curve plots the test sensitivity as a function of the false-positive rate (100 – Sp). As with the DOR, the ROC curve is a single measure assumed to be independent of prevalence, and hence is referred to as a global and unconditional measure.

Figure 3. Example of a Receiver Operating Characteristic (ROC) curve



Notes: the ROC curve is based on the distributions for the hypothetical test described in Figure 1. The data labels refer to cut-off point values for a positive test result. A high cut-off point results in low sensitivity and high specificity. The area in grey represents the Area Under the Curve (AUC).

The diagonal line from bottom-left to top-right in Figure 3 represents a test that is essentially uninformative, as the ability to detect genuine cases is no better than chance allocation to positive and negative. The upper left-hand corner represents a test with sensitivity and specificity of 100%, in other words a perfect dichotomous test. Clearly a desirable test is as close as possible to the upper left-hand corner and as far from the diagonal as possible. The cut-off point that yields the most upper-left point may be appropriate for clinical practice, presuming it is feasible and has been validated in (preferably multiple) independent samples. The upper left point, or maximal joint sensitivity and specificity, is also known as Q*. It is reported as the sensitivity where the line of symmetry intersects the ROC curve, The Q* point may be of little relevance if it is not possible to implement test thresholds that result in maximal joint sensitivity and specificity. Similarly, not all threshold values may be possible in practice, so the ROC curve may not be able to span the entire range of sensitivity and specificity in practice.

An area under the curve (AUC) of 1 represents a perfect test, while an AUC of 0.5 represents an uninformative test. The AUC is sometimes reported as a single summary measure of diagnostic accuracy and gives an indication of how close to perfect, or uninformative, a test is. Two tests, one with high sensitivity and the other with high specificity, may have the same AUC. The AUC does not provide any information about how the patients are misclassified (i.e., false positive or false negative) and should therefore be reported alongside another measure of test performance, such as likelihood ratios or predictive values.[2] The AUC is not useful in clinical practice as it summarises performance over a range of possible thresholds, whereas in practice a single pre-specified threshold applies. It should be noted that ROC curves of different shapes can

have the same AUC value, so an AUC value does not represent a set of unique combinations of sensitivity and specificity.[7]

1.1.6. Predictive values

The positive predictive value (PPV) is the proportion of patients with a positive test who actually have the disease, and the negative predictive value (NPV) is the proportion of patients with a negative test result who are actually free of the disease.[2]

$$PPV = \frac{Sn.P}{Sn.P+(1-Sn).(1-P)} \quad (7)$$

$$NPV = \frac{Sp.(1-P)}{(1-Sn).P+Sp.(1-P)} \quad (8)$$

Where *P* is the estimated prevalence of disease, also known as the pre-test or prior probability of disease.[8]

A patient belonging to a population with a higher prevalence of disease will have a higher PPV than a patient from a lower prevalence population. Predictive values have a strong clinical utility. However, they vary with disease prevalence and are not useful in situations where test results are reported on multiple levels rather than a dichotomous outcome. The predictive values are referred to as specific conditional measures of test accuracy.

1.1.7. Diagnostic accuracy

A single overall measure of test accuracy is also used which is expressed as the proportion of correctly classified cases:[9]

$$\text{Diagnostic accuracy} = (TP + TN) / (TP + FP + FN + TN) \quad (9)$$

A global, conditional measure of accuracy, it is not often used and is not pooled across studies.

1.2. Problem statement

Diagnostic tests are used for a variety of purposes including to: determine whether or not an individual has a particular target condition; provide information on a physiological or pathological state, congenital abnormality, or on a predisposition to a medical condition or disease; predict treatment response or reactions; define or monitor therapeutic measures. As such, the test is not a treatment, but influences a clinician when deciding on the appropriate course of action for a particular patient. Timely or correct detection of disease does not necessarily lead to timely or correct treatment of disease, hence improved diagnostic test accuracy is not synonymous with improved patient outcomes. Diagnostic tests can change patient outcomes by changing diagnostic and treatment decisions, impacting on timely treatment, modifying patient perceptions and behaviour, or putting patients at risk of direct harm.[10]

To study the association between the accuracy of a diagnostic test with regard to outcomes, a follow-up is required, but this may be at significant risk of confounding[11] unless studied in randomised controlled trials (RCTs). In the area of health technology assessment, diagnostic test accuracy is sometimes used as a surrogate for patient-relevant outcomes, although some agencies require long-term outcome data on patient outcomes.[12] In practice, diagnosis may also depend on factors other than just the results of a single diagnostic test, such as clinical history and additional testing, and hence other factors will impact on diagnosis, treatment and outcomes.[13] A linked evidence approach, whereby patient outcomes can be associated with the diagnostic test, may be a pragmatic solution although in practice there are often insufficient data to enable this approach.[14] It should also be noted that diagnostic tests may be relatively invasive (e.g., sentinel lymph node biopsy) or harmful to patients (e.g., exposure to ionising radiation), and that this information is not captured in the assessment of test accuracy.

When the sensitivity of a new diagnostic test is compared with an existing test, the detected cases may be different to those detected by the existing test. Results from treatment trials based on patients detected by the old test may not be generalisable to the cases detected by the new test. Unless clinicians can be satisfied that the new test detects the same spectrum and subtype of disease as the old test or that treatment response is similar across the spectrum of disease, it is possible that the new test will result in different outcomes.[15]

The impact of a diagnostic test can be viewed according to a number of domains (see Table 1).[16-18] The tiered model has been tailored to radiological testing, for which the resolution and sharpness of test images are relevant. For other types of tests, the resolution and sharpness may be analogous to the precision of the test.

Table 1. Tiered model of diagnostic efficacy[16-18]

Stage of efficacy	Definition
Technical capacity	Resolution, sharpness, reliability
Diagnostic accuracy	Sensitivity, specificity, predictive values, ROC curves
Diagnostic impact	Ability of a diagnostic test to affect the diagnostic workup
Therapeutic impact	Ability of a diagnostic test to affect therapeutic choices
Patient outcomes	Ability of a diagnostic test to increase the length or quality of life
Societal outcomes	Cost-effectiveness and cost-utility

The stages or tiers of efficacy answer a variety of questions about a diagnostic test, from whether or not it can work to whether or not it is worth using. This guideline is restricted to methodologies for summarising diagnostic accuracy. It must be noted that diagnostic test accuracy is not in itself a measure of clinical effectiveness, and improved accuracy does not necessarily lead to improved patient outcomes. Meta-analysis of diagnostic test

accuracy therefore estimates the pooled test accuracy and not pooled clinical effectiveness. Estimating clinical utility requires appropriate studies with longer term patient outcomes and is not considered in this guideline.

1.3. Objective(s) and scope of the guideline

This guideline presents a review of the available methods for the meta-analysis of diagnostic test accuracy studies. The aim of the guideline is to highlight the circumstances in which it is appropriate to use each of the approaches. The guideline will also elaborate on:

- thresholds for positive tests
- fixed and random effects approaches
- heterogeneity across studies
- sample sizes
- the quality and quantity of evidence required for a meta-analysis
- the case where multiple diagnostic tests may be evaluated and compared
- issues regarding study selection
- the types of bias that might arise when reviewing diagnostic test accuracy data.

The guidance is restricted to methods for pooling results from diagnostic tests that report dichotomous results (i.e., the test result is either positive or negative), as opposed to tests that report results on a continuous scale or as a number of discrete levels.

The guideline does not address issues relating to systematic reviews and meta-analysis that are not restricted or unique to diagnostic test accuracy studies. These issues include: bibliographic searching and study types. These issues are common to any meta-analysis and are comprehensively described elsewhere.[19;20] It is assumed that the meta-analysis is undertaken using comparable studies derived from a systematic review conducted according to best practice. This guideline also does not consider analysis of intermediate or longer term outcomes in relation to diagnostic test performance.

1.4. Related EUnetHTA documents

The following EUnetHTA methodological guidelines are relevant to the present guideline:

- Applicability of evidence in the context of a relative effectiveness assessment of pharmaceuticals (February 2013)
- Direct and indirect comparisons (February 2013)

It should be noted that the EUnetHTA guidelines were developed for the relative effectiveness assessment of pharmaceuticals. The extent to which the principles contained in the guidelines are also relevant to the meta-analysis of diagnostic test accuracy studies

will depend on the nature of the diagnostic test being evaluated. In general, the guideline on the applicability of evidence is most likely to be of use when pooling data from diagnostic test accuracy studies.

Also relevant are the assessment elements from the HTA Core Model Application for Diagnostic Technologies:

- Assessment element tables for HTA Core Model Application for Diagnostic Technologies (2.0), meka.thl.fi/htacore/model/AE-tables-diagnostics-2.0.pdf

2. Analysis and discussion of the methodological issue

2.1. Methods for meta-analysis of diagnostic accuracy studies

A variety of methods are available for pooling data from multiple studies of diagnostic test accuracy. The relevance of each method is influenced by the type of study data available (e.g., individual patient data, 2x2 tables, summary measures such as sensitivity and specificity). Certain data may not be available for all studies, which will also influence the approach to pooling data. The most straightforward approach is a simple pooling where the 2x2 tables from all of the studies are combined with no weighting.[7] This method assumes no correlation between sensitivity and specificity, no between-study heterogeneity, and no variability in the diagnostic threshold. As such, simple pooling can be described as a naive approach and will not be considered in these guidelines.

It is assumed that a meta-analysis is only undertaken when the available studies are considered equivalent in terms of the index test, reference standard, the patient population, and the indication. Where the studies are not equivalent it is not recommended that a meta-analysis is undertaken. A lack of study equivalence gives rise to various types of bias which are discussed in Section 2.4.

2.1.1. Separate random-effects meta-analyses of sensitivity and specificity

Sensitivity and specificity can be individually summarised across studies based on their logit transforms.[21] This approach is a random-effects method that allows for between-study heterogeneity in the two measures, but ignores the potential correlation between the two. The logit transforms are used in the analysis on the basis that an assumption of a normal distribution between studies is more reasonable on the logit scale, with an inverse logit transformation applied to the results to return them to a [0,1] interval. In addition to the point estimates of sensitivity and specificity, this approach also allows for the estimation of a ROC curve using the ratio of estimated between-study variances. This approach has been suggested for situations when there is evidence of no correlation between sensitivity and specificity across studies.[22] However, in practice, situations where it is plausible that there is no correlation are highly unlikely to arise.

2.1.2. Separate meta-analyses of positive and negative likelihood ratios

As likelihood ratios are ratios of probabilities, positive and negative likelihood ratios can be pooled separately by meta-analysis using the same mathematical methods as risk ratios.[21] Approaches can be based on either fixed-effect or random-effects models. These methods ignore the possible correlation between positive and negative likelihood ratios, and thus pooled estimates may produce values that are not possible in reality (e.g., both ratios above or below 1.0).[23] Should they be required, pooled estimates of the likelihood ratios can be computed from summary estimates of sensitivity and specificity derived using any of the other methods described in this section. Meta-analysis of predictive values is possible, although it is usually discouraged because of the influence of disease prevalence.

2.1.3. Moses-Littenberg summary receiver operating characteristic (SROC) curve

The Moses–Littenberg fixed-effects method is historically the most commonly used method for meta-analysis of diagnostic tests. A straight line is fitted to the logits of the false positive rate (FPR) and true positive rate (TPR) of each study, and its slope and intercept give the parameters of the SROC curve.[22] The SROC curve summarises pairs of

sensitivity and specificity from multiple studies. The least squares linear fit may be unweighted or weighted, although in the latter case there is uncertainty as to which weighting method to use.[21] The linear fit is then back-transformed to be plotted as the SROC curve. The Moses-Littenberg may be appropriate if all the observed heterogeneity is due to a threshold effect. That is, where all of the observed heterogeneity is due to the use of different thresholds across the included studies.

This method allows for the correlation between sensitivity and specificity, but is not statistically rigorous, as the assumptions of linear regression (constant variance, covariate measured without error) do not hold.[21] Furthermore, as it is based on an analysis of the DOR, summary measures of sensitivity and specificity are not directly available. By selecting a value for sensitivity, it is possible to compute the corresponding specificity. It is common to report the sensitivity and specificity at the Q-point(i.e., where the SROC curve intersects the diagonal that runs from the top left to bottom right of the ROC plot; sensitivity equals specificity on this diagonal).[24] However, the values at the Q-point may bear little relation to the values observed in the original studies used in the meta-analysis.

2.1.4. Hierarchical summary ROC (HSROC) model

The HSROC model for combining estimated pairs of sensitivity and specificity from multiple studies is an extension of the Moses-Littenberg fixed-effects summary ROC (SROC) model.[25] The HSROC model more appropriately incorporates both within- and between-study variability, and allows greater flexibility in the estimation of summary statistics. The HSROC model describes within-study variability using a binomial distribution for the number of positive tests in diseased and not diseased patients.

The model is specified on two levels: the within study model and the between study model. The within study model takes the following form:[26]

$$\mathrm{logit}\left(\pi_{ij}\right) = \left(\theta_i + \alpha_i X_{ij}\right)\exp\left(-\beta X_{ij}\right) \qquad (10)$$

The variable π_{ij} is the probability that a patient in study *i* with disease status *j* will return a positive test result. By defining *j*=0 for a patient without the disease and *j*=1 for a patient with the disease, it follows that for study *i*, π_{i0} is the false positive rate and π_{i1} is the true positive rate. The parameter X_{ij} is a dummy variable for the true disease status of a patient in study *i* with disease status *j*. The parameters $\theta_i$and α_i are the cut-off point and accuracy parameters, respectively, and are allowed to vary between studies. Finally, β is a scale parameter for modelling the possible asymmetry in the ROC curve.

The between-study model allows the parameters $\theta_i$and α_i to vary between studies. The following parameter definitions include a common covariate Z which affects both parameters, although they can be modelled without covariates or with multiple covariates:

$$\theta_i \sim \mathrm{N}\left(\Theta + \gamma \mathrm{Z_i}, \sigma_\theta^2\right) \qquad (11)$$

$$\alpha_i \sim \mathrm{N}\left(\Lambda + \lambda \mathrm{Z_i}, \sigma_\alpha^2\right) \qquad (12)$$

The model was originally formulated in a Bayesian framework, and hence also included specification of priors.[25] The model produces an SROC curve by allowing the cut-off point parameter to vary while holding the accuracy parameter at its mean value.

2.1.5. Bivariate random-effects meta-analysis for sensitivity and specificity

As with the HSROC method, the bivariate approach preserves the two-dimensional nature of the original data, with pairs of sensitivity and specificity jointly analysed.[24] Like the HSROC approach, this method also incorporates any correlation that might exist between these two measures using a random-effects approach. Evaluation of the bivariate model requires specification of an appropriate transformation (e.g., a generalised linear mixed model using the logit-transformation).[27] Explanatory variables can be added to the bivariate model and lead to separate effects on sensitivity and specificity, rather than a net effect on the odds ratio scale as in the SROC approach.[24]

The bivariate model is specified as follows:[26]

$$\begin{pmatrix}\mu_{Ai}\\ \mu_{Bi}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\mu_A\\ \mu_B\end{pmatrix}, \Sigma_{\mathrm{ab}}\right) \quad (13)$$

$$\Sigma_{\mathrm{AB}} = \begin{pmatrix}\sigma_{\mathrm{A}}^2 & \sigma_{\mathrm{AB}}\\ \sigma_{\mathrm{AB}} & \sigma_{\mathrm{B}}^2\end{pmatrix} \quad (14)$$

The variables μ_{Ai} and μ_{Bi} are the logit transformed sensitivity and specificity, respectively, for study *i*. Covariates affecting sensitivity and specificity can be included by replacing the means μ_A and μ_B with linear predictors in the covariates.[26] The covariates can be applied to one or both measures, and can have common or distinct effects.

2.1.6. Comparison of methods

The appropriate choice of methodology for the meta-analysis of diagnostic test accuracy studies will depend on numerous factors. The Moses-Littenburg model is considered as approximate, as the assumptions of simple linear regression are not met and because of the uncertainty around the appropriate weighting.[26] As the Moses-Littenberg model is essentially a fixed-effect model it does not provide estimates of the between study heterogeneity.[28] This method can also lead to improper SROC curves.[28]

The HSROC and bivariate methods are equivalent under certain parameterisations, such as in the absence of covariates or when the same covariates affect both sensitivity and specificity (in the bivariate model) and both the accuracy and cut-off point parameters (in the HSROC model).[26] Therefore in situations where there are no covariates, the two models will return equivalent estimates of the expected sensitivity and specificity (and also any measures derived from those two measures).

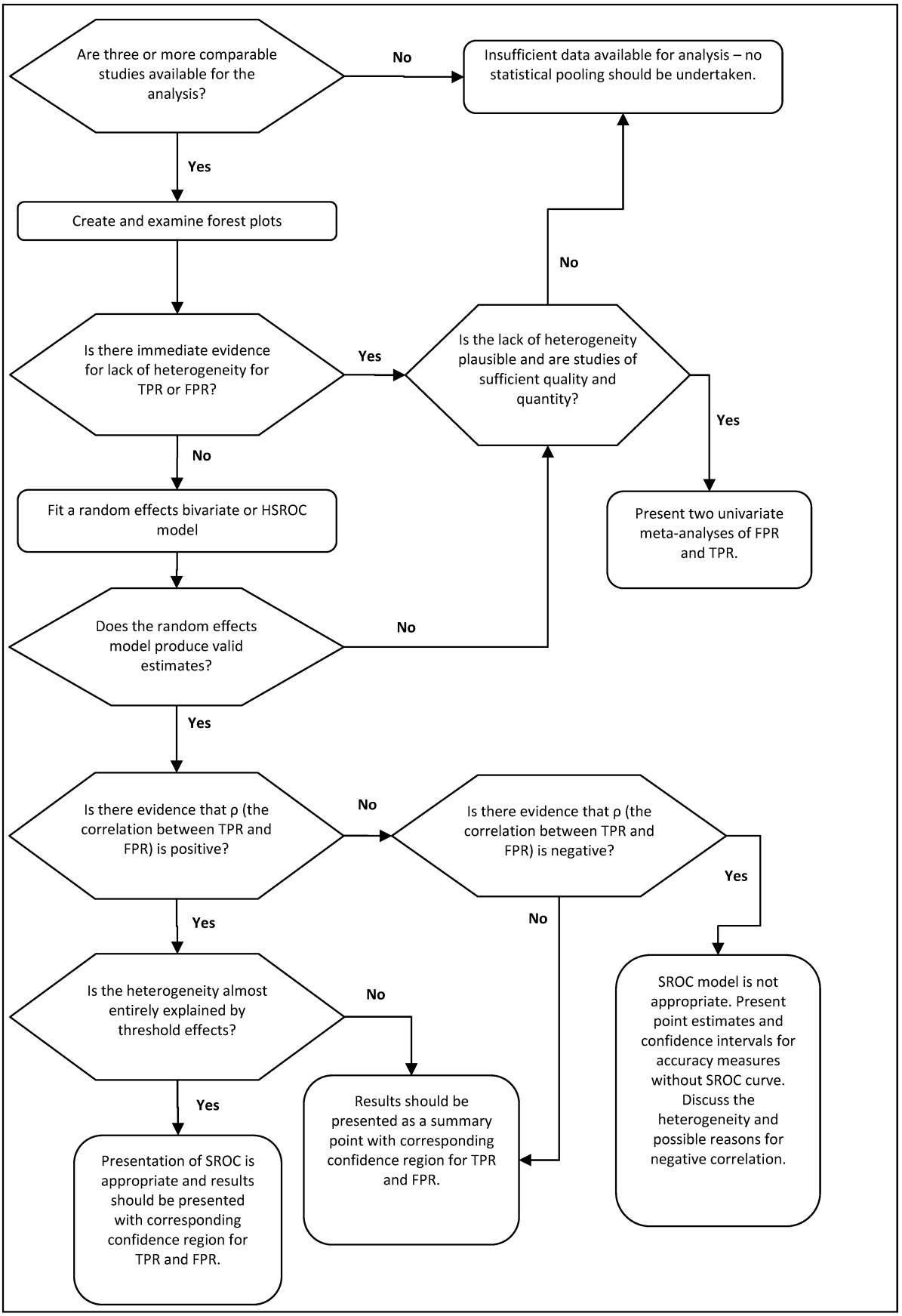
The HSROC and bivariate approaches are considered to be more statistically rigorous than the Moses-Littenberg approach,[21;28] although it has been questioned whether this necessarily translates into improved estimates of diagnostic test accuracy in all situations.[29] There is an increasing consensus that the HSROC and bivariate approaches offer the best methodologies for pooling diagnostic test accuracy studies, but there are differences between the two approaches and the nature of the underlying data may dictate which approach is more appropriate.

A first step is to separately examine the distributions of sensitivity and specificity from the included studies.[22] If either measure shows a lack of heterogeneity, then it is more appropriate to analyse the data using separate univariate meta-analyses to derive point estimates and confidence bounds for sensitivity and specificity. However, a full description of the included studies can provide contextual information that may justify a full analysis in these situations. If only one study is available, then clearly there is no basis for meta-

analysis. If only two studies are available, then there is insufficient information available to reliably estimate all of the parameters in the HSROC and bivariate models. Therefore, in the case of two studies, it is not recommended to undertake a meta-analysis and a narrative description of the studies should be presented.

The correlation between sensitivity and specificity is important and is estimated by the HSROC and bivariate methods. Ordinarily a positive correlation is expected between TPR and FPR. However, data from studies are often noisy and no correlation or a negative correlation may be estimated. The confidence bounds of the correlation estimate should be assessed. If there is a significant negative correlation, this implies that sensitivity improves with increasing specificity, which is unlikely to occur in practice due to the relationship between disease status and the test cut-off point (see Figure 2). In the event of a negative correlation, the plausibility of this finding should be discussed in relation to the nature of the test and the quantity of evidence.

Figure 3. Algorithm for the meta-analysis of diagnostic test data (adapted from Chappell et al.[22])



A key consideration is then whether or not a threshold effect is present, which is usually evidenced by a positive correlation between the false positive rate and sensitivity. When a threshold effect is present, then an SROC approach is appropriate, which can be achieved using either the HSROC or bivariate approaches. Estimates of test accuracy can be plotted in ROC space. In the absence of a threshold effect, the SROC approach is not appropriate. There will be situations were a threshold effect may or may not be plausible, depending on the nature of the test and the indication. For example, some tests explicitly depend on converting a measure on a continuous scale into a dichotomous measure of disease status (e.g., Prostate-Specific Antigen (PSA) test). These tests are likely to give rise to threshold effects. Some tests, on the other hand, may rely on a simple presence/absence measure of a biomarker which is directly interpreted as a measure of disease status (e.g., rapid strep test), or may employ an unequivocal cut-off point that is universally adopted (e.g., Ottawa Ankle rules). Due to differences in how test results are interpreted, a threshold effect may arise even when there is a universally employed cut-off point.

If a threshold effect is plausible, and heterogeneity is observed, then it must be evaluated if the heterogeneity can be attributed to a threshold effect. Determining whether observed heterogeneity is due to a threshold effect is generally based on a visual inspection of the distribution of study points in relation to the SROC curve. If study points are in close proximity to the SROC, then there will be reasonable confidence that the threshold effect is responsible for the heterogeneity. An inspection of the shape of the confidence region is also helpful, particularly to check whether the region largely encompasses and follows the shape of the SROC curve. If, on the other hand, the prediction region bears little relation to the SROC curve, or the study points are not close to the SROC curve, then it is reasonable to conclude that factors other than just a threshold effect are responsible for the observed heterogeneity.

The choice of method used must be justified by the context (e.g., the studies, inclusion of covariates, correlation between sensitivity and specificity), and the potential impact of the assumptions on the interpretation of the results must be clarified.

2.2. Presentation of results from a meta-analysis of a single diagnostic test

Reports of meta-analyses of diagnostic test accuracy must contain the requisite information for a reader to know how the analysis was undertaken, what data were used, what results were found, and whether or not the findings are reliable. To achieve this, a number of presentational features should be included, depending on the type of analysis undertaken.

2.2.1. Tables

Reports should include a table of all the included studies, and the relevant data from the 2x2 tables for each of the studies. Such tables can also include the estimated sensitivity and specificity and associated confidence bounds for the two measures for each included study.

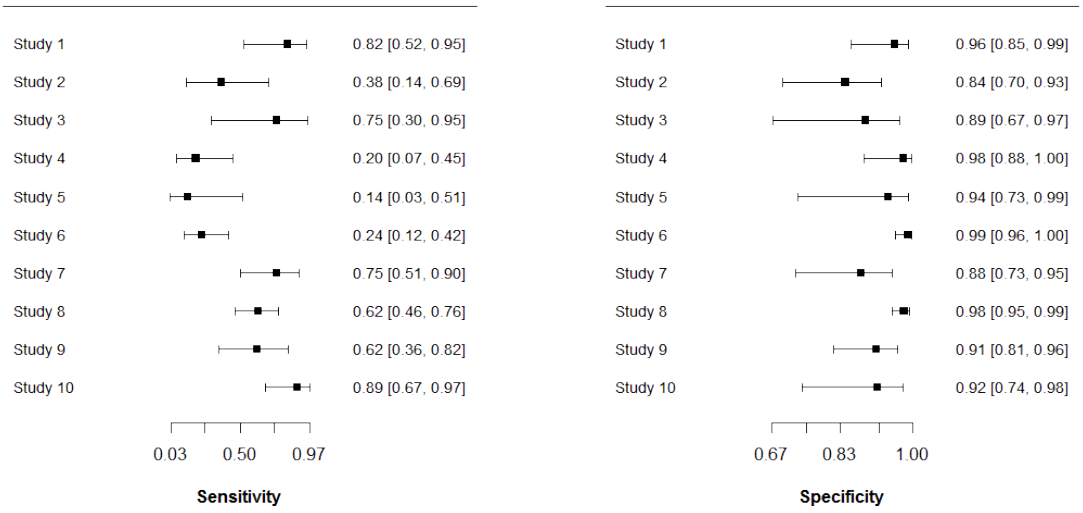
For the results of the meta-analysis, summary estimates of accuracy and their associated confidence bounds should be reported. The main result of the bivariate and the HSROC models is the pooled estimate of the summary-paired sensitivity and specificity. At a minimum, the results for sensitivity and specificity should be reported, although the DOR and likelihood ratios may also be useful. Any other useful outputs from the analysis (e.g., the estimated correlation between sensitivity and specificity) should also be reported as they may aid interpretation of the data and results.

The results of any subgroup analyses should also be tabulated. Results of sensitivity analyses can also be included in tables, as summary points and confidence bounds cannot always be easily read from graphical displays.

2.2.2. Forest plots for sensitivity and specificity

Forest plots (also called blobbograms) of sensitivity and specificity are useful for showing heterogeneity across studies. These plots give the point estimates and confidence bounds for sensitivity and specificity for the individual studies included in the analysis (see Figure 4). Studies may be ordered by sensitivity or specificity, which can aid interpretation or make it more apparent if there is a correlation between the two measures.

Figure 4. Forest plots of sensitivity and specificity for a sample meta-analysis

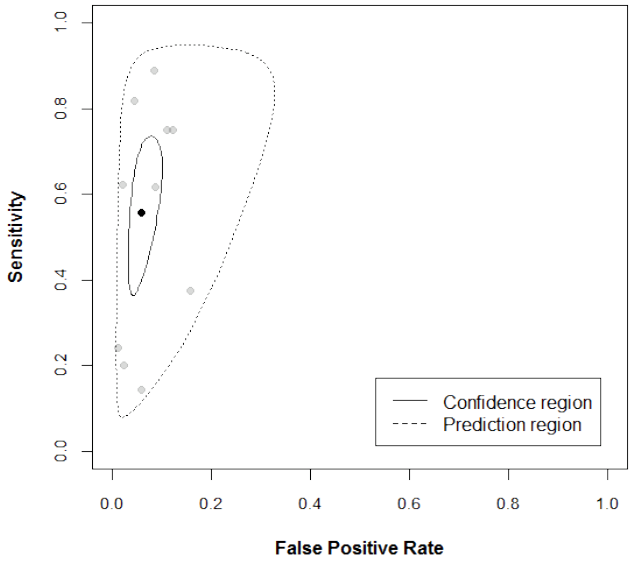


2.2.3. Confidence and prediction regions for the summary estimate of sensitivity and specificity

From the meta-analysis of diagnostic test accuracy, it is possible to generate 95% confidence and prediction regions for sensitivity and specificity. The confidence region relates to the summary point estimate based on the included studies whereas the prediction region refers to potential values of sensitivity and specificity that might be observed in a future study. If the summary values for sensitivity and specificity are to be used in a subsequent relative effectiveness assessment simulation model, the prediction region may form a more realistic basis for defining parameter uncertainty than the more narrowly defined confidence region. Furthermore, prediction regions can also be used for the purpose of identifying studies that may be statistical outliers.

Both the HSROC and bivariate models facilitate the computation of confidence and prediction regions around the summary point for sensitivity and specificity, usually in the form of a joint confidence ellipse for sensitivity and specificity.

Figure 5. An example of a summary estimate of sensitivity and false positive rate

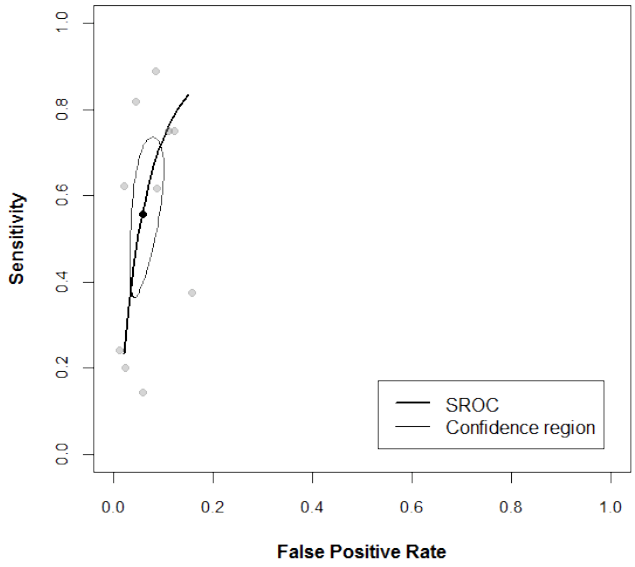


2.2.4. Summary ROC curve

The choice to display an SROC curve depends on whether or not the included studies had a common positivity threshold and the subsequent analytical approach. In instances where the threshold varies across studies, a summary estimate of sensitivity and specificity is of limited use as it represents an average across thresholds. Where the threshold varies, it is appropriate to report an SROC curve.

It must be noted that the SROC curve as specified for the HSROC model is constrained to always be positive.[25] An SROC curve can also be generated from the results of the bivariate model, although a variety of formulations are possible which can lead to quite different curves, including those with a negative slope.[30] Another drawback of the SROC curve is that uncertainty in the SROC curve is not generally calculated as a single common SROC curve is assumed. Using a Bayesian approach, it is possible to generate numerous SROC curves based on posterior densities which can be used to derive a graphical indication of uncertainty in the curve within specified quantiles.[22]

Figure 6. An example of a summary receiver operating characteristic (SROC) curve



It is suggested that the SROC curve should be restricted to the observed range of specificities in the included studies and that the analyst should not extrapolate beyond the observed data.[25] For example, if the highest upper bound for the false positive rates observed in the included studies is 0.60, then the SROC curve should not be extended beyond a FPR of 0.60 in the plot.

2.2.5. Sensitivity analysis

It is typical in relative effectiveness assessments to consider the influence of various factors on results. This is generally achieved through sensitivity analysis – an evaluation of how much the conclusions change if the included evidence is changed. In this context, sensitivity analysis refers to the quantification of uncertainty rather than an analysis of the measure of diagnostic accuracy. Sensitivity analysis may be targeted (e.g., re-analysing the data with data at risk of bias excluded) or systematic (for example, univariate sensitivity analysis with all uncertain parameters varied one at a time). The same principles apply to meta-analysis.

In any meta-analysis there is likely to be heterogeneity across the included studies. There can be many reasons for that heterogeneity, including systematic differences between the studies in terms of the patients, how the test was applied, and the choice of reference standard. Study quality can also be variable and the application of a formal risk of bias measure can be used to identify specific studies at high risk of bias. A targeted sensitivity analysis may involve excluding studies that are considered outliers in a statistical sense, or that have been evaluated as being at high risk of bias. Alternatively, the meta-analysis may be restricted to a sub-group of studies with a common characteristic, although this can also be achieved by a meta-regression approach, which is possible with both the HSROC and bivariate methods. By extension, a systematic sensitivity analysis may be based around repeating the meta-analysis with each of the studies excluded in turn. Study influence can be measured using metrics such as Cook’s distance, while statistical outliers may be identified using standardised study-level residuals.[31] In both cases, these metrics can be applied to sensitivity and specificity simultaneously.

If the results remain relatively unchanged then there can be confidence that the summary estimates are accurate. If, however, the results are sensitive to the included data, then greater attention needs to be paid to the included studies and what characteristics are impacting on differences. Evidence for a relative effectiveness assessment must be

relevant to the target population and conditions under which the diagnostic test will be used in practice.

2.3. Comparison of two diagnostic tests with respect to diagnostic accuracy (incorporate non-comparative studies in discussion of heterogeneity)

Estimating diagnostic test accuracy is often for the purpose of comparing two or more tests for the same indication. In this situation, the diagnostic accuracy of all tests has to be compared.[32] However, the evidence derived from comparative and non-comparative studies often differs. Ideally, for the purposes of comparing two diagnostic tests, robustly designed studies in which all patients receive all tests or are randomly assigned to receive one or other of the tests are preferred as evidence to guide test selection.[33] Irrespective of the comparison, the same reference standard test should be used for all patients. The use of data from non-comparative studies increases the chances of differences in the patient populations, different reference standard tests, and different interpretation of test results.

2.4. Sources of bias

Evidence has shown that diagnostic studies with methodological shortcomings may overestimate the accuracy of a diagnostic test, particularly those including non-representative patients or applying different reference standards.[34] As with the meta-analysis of interventions, the pooling of data across diagnostic test accuracy studies may be subject to numerous sources of bias, although some forms of bias are specific to diagnostic test studies.[23] In this section we outline some of the main sources of bias that can occur. In many cases, there is little that can be done to correct for bias beyond a forensic examination of the included studies, careful documentation of potential bias, and a full sensitivity analysis to examine the potential impact on results of including studies at risk of bias.

2.4.1. Data gathering and publication bias

As with the meta-analysis of any clinical intervention, meta-analysis of diagnostic test accuracy studies should be undertaken as part of a systematic review. Methods for systematic review are well described elsewhere,[19] with specific guidance available for diagnostic test accuracy studies.[35] The identification of diagnostic test accuracy studies can pose particular difficulties, due in part to the lack of consistent terminology or use of MESH terms, indeed, in some cases methodological filters can reduce the ability to find relevant studies.[36;37] Best practice is to search on the basis of the index test and target condition.[37] Difficulties can also arise where a single study publishes multiple articles using the same or overlapping cohorts; care must be taken not to include data on the same patients from several articles, which can be referred to as double data reporting bias.

Publication bias is believed to arise due to studies with poor test performance results not getting published, leading to exaggerated estimates of test performance in a systematic review.[38] As with meta-analyses of clinical interventions, asymmetry in the funnel plot (constructed using the DOR) is often taken as an indication that there may be publication bias, although there may be many other factors causing asymmetry (e.g., variations in test procedures, patients, or reference standards).[39] It is possible that publication bias may be more prevalent in studies of test accuracy than in studies of clinical effectiveness.[39] There are a number of approaches available for estimating funnel plot asymmetry, each of which may give different results in a given context. The unique features of the test accuracy study make the application of the Begg, Egger, and Macaskill tests of funnel plot asymmetry potentially misleading for typical DOR values.[40] Alternative funnel plots using

the natural logarithm DOR and functions of the effective sample size may be useful for evaluating publication bias. [40] The power of any of these statistical tests for funnel plot asymmetry decreases with increasing heterogeneity of DOR. Other factors may also be associated with sample size and hence may impact on the results of publication bias tests. Furthermore, where the number of included studies is small, the statistical methods available may be underpowered to detect asymmetry. As such, funnel plot asymmetry should be used but interpreted with caution.[38]

2.4.2. Heterogeneity in meta-analyses of sensitivity and specificity

Between-study heterogeneity refers to differences in variability in the results of studies. Clinical, methodological, and statistical heterogeneity are distinct concepts. Clinical heterogeneity refers to variability across studies in terms of participants, the intervention, and outcomes. These are legitimate differences that arise because the studies are not comparing like with like. Methodological heterogeneity is a function of variability in study design and risk of bias. Differences in methodology may include differences in the technical specifications of the test, such as the protocols for how the test is applied. This may also be referred to as technical heterogeneity. Statistical heterogeneity arises when there is greater variability in outcomes than would be expected by chance, and usually invokes a violation of underlying assumptions. For the purposes of this guideline, the outcome measure of interest is diagnostic test accuracy. Clinical and methodological heterogeneity will often, but not necessarily give rise to statistical heterogeneity.

The obvious source of statistical heterogeneity in sensitivity and specificity is due to threshold differences for test positivity.[41] If the observed between-study heterogeneity is entirely due to variation in the diagnostic threshold, estimates of summary sensitivity and specificity will underestimate diagnostic performance.[3] In these situations the appropriate meta-analytical summary is the receiver operating characteristic curve rather than a single summary point. However, it must be clear that there are no other substantial sources of heterogeneity. Where there are a variety of sources of heterogeneity, including threshold effects, the HSROC or bivariate method should be used with random effects. Presentation of an SROC may not be informative unless some attempt to measure uncertainty in the curve is included.

Another potentially important source of heterogeneity is due to observer variability. Within-study observer variability can be of the same order of magnitude as variability across studies.[42] By including studies from a wide time horizon, it is also possible that changes in how a diagnostic test is used in practice may have occurred, giving rise to heterogeneity.[42]

Although measures of heterogeneity exist for univariate meta-analyses (e.g., I^2, τ^2), there is no analogue for bivariate meta-analyses. The amount of observed heterogeneity is quantified by the random effects terms in the models, but these are not easily interpreted.[28] The distribution of study points on a plot of true versus false positive rates relative to the estimated SROC can give an indication of whether there is heterogeneity due to variation in the test threshold. The distribution of points relative to the prediction ellipse can also provide an indication of whether or not there is heterogeneity.[28]

A common approach to exploring heterogeneity is to use meta-regression whereby study-level covariates are included when estimating summary statistics. Both the HSROC and bivariate models facilitate the use of study-level covariates as either categorical (e.g., study design) or continuous (e.g., average patient characteristics).[28] In the bivariate model, covariates can be incorporated to affect summary sensitivity or summary specificity, or both measures. The HSROC model, on the other hand, allows covariates to be added to

affect the test positivity, position of the curve, and shape of the curve. A covariate may be associated with some, but not all three model parameters.[28]

2.4.3. Spectrum bias

As test performance often varies across population subgroups, diagnostic tests should be evaluated in a clinically relevant population. The performance of the test may vary depending on the mix of patients, most particularly due to differences in disease severity. Inappropriate use of patient populations can occur, introducing a form of heterogeneity referred to as spectrum bias.[43] When there is spectrum bias the diagnostic test performance varies across patient subgroups and a study of that test's performance does not adequately represent all subgroups. The impact of spectrum bias on the estimated test accuracy will depend on the difference between included patients and the actual target population.

2.4.4. Verification/work-up bias and variable gold standard

Verification bias (also called selection or workup bias), occurs when not all recipients of the index test also receive the reference or gold-standard test.[44] This will often occur where a primary study uses a two stage design, where all patients receive the index test in the first stage, but only a subsample receives the reference test in the second stage. The reference test is required to verify if the tested individuals did or did not have the target indication. When selection of subjects for the reference standard is not completely random, verification bias will occur.[44] When verification bias is present, it will often lead to an overestimate of the sensitivity of the index test.[23] To prevent misleading comparisons, estimates from a trial with a series or multi-stage design must always be described in the context of the trial design and study population.[45]

2.4.5. Bias resulting from choice of cut-off points

A data-driven approach to the selection of the optimal cut-off values can result in overly optimistic estimates of sensitivity and specificity, particularly in small studies.[46] Using simulation, it has been shown that data-driven cut-off points frequently exaggerate test performance, and that this bias probably affects many published diagnostic validity studies.[47] Bias can be reduced by optimising cut-off points using a training dataset and then applying those cut-off points to a second test set of data. However, such an approach is reliant on sufficient data availability, which is frequently problematic when considering diagnostic test accuracy studies. Pre-specified cut-off points improve the validity of diagnostic test research, and this is particularly the case for studies with small samples. Alternative methods can be used to reduce this bias, but finding robust estimates for cut-off values and accuracy requires considerable sample sizes.[46]

2.4.6. Disease prevalence

Although contrary to typical assumptions, the sensitivity and specificity of a diagnostic test can vary with disease prevalence.[48] This effect is likely to be the result of a number of mechanisms, including patient spectrum, which affect prevalence, sensitivity and specificity. Trivariate generalised linear mixed models have been applied to jointly model prevalence, sensitivity and specificity, enabling the assessment of correlations between the three parameters.[49;50]

2.4.7. Potential for dependence in combined tests

For the purpose of this guideline, it is presumed that combined tests are not repeated applications of the same test, as might happen in a screening programme, but rather the use of a variety of tests with the aim of increasing the overall diagnostic accuracy.

When investigating combined tests, or tests carried out in sequence, the correlation between test results is important. Two perfectly correlated tests will return the same results, and hence the second test does not add any information from a diagnostic point of view. This is important for the clinician: if two correlated tests are treated as independent, then the post-test probability of disease will be over-estimated by two positive tests.[51] From a meta-analytic point of view, combined tests can give rise to a number of problems, not least a multiplication of the issues for single diagnostic tests.

Where multiple tests are used for diagnosis, it is highly likely that the tests will not perform independently.[52] That is, in the case of two tests, the performance of the second test may depend on the results of the first test. When the assumption of dependence between tests is ignored, this can lead to erroneous disease probability estimates.[52]

A further issue is that patients testing positive may be removed from the tested population to receive treatment. This change to the population may affect the disease prevalence and may also introduce spectrum bias.

2.4.8. Missing data/non-evaluable results

Reports of diagnostic test accuracy studies will sometimes refer to missing data or non-evaluable results. This may be done explicitly in the text or it may be apparent from the 2x2 tables where the numbers of tests are inconsistent. A potential for bias exists if the number of patients enrolled differs from the number of patients included in the 2x2 table of results, as patients lost to follow-up differ systematically from the remaining patients.[53] Missing data can occur for a variety of legitimate reasons. For example, if a patient is to receive two different tests and is clearly positive after the first, it may be unethical to subject them to the second test if it causes a delay in treatment. Non-evaluable results can occur where the results, of what is intended to be a dichotomous measure, cannot be unequivocally classified. The exclusion of non evaluable results leads to the overestimation of diagnostic accuracy.[54] One potential solution is to adopt an intention-to-diagnose approach, which can be formulated as a 3x2 table in which non-evaluable results are included.[54] Such an analysis can significantly decrease the estimate of diagnostic performance.

2.4.9. Individual patient data analysis

Individual patient data meta-analysis enables the evaluation of diagnostic test accuracy in relation to other relevant information.[55] This approach could increase the efficiency of the diagnostic work-up by, for example, reducing the need for invasive confirmatory tests.[13;55;56] The addition of clinical information when interpreting the results of diagnostic tests can also improve accuracy;[58] if this has been applied to some, but not all patients in a study, it could be recorded as a covariate and used in an individual patient analysis. Allowing for individual patient characteristics can also allow for proper accounting of differences in the patient spectrum, and enable test results to be interpreted based on additional patient information.[26]

2.5. Meta-analysis of the prognostic utility of a diagnostic test

A prognostic factor is typically a biomarker that is used to predict future events, such as disease progression or mortality. Studies of prognostic factors aim to estimate the

relationship between the prognostic factor and an outcome. In some instances, prognosis is based on a measure such as a relative risk or hazard ratio, in which case the meta-analytic approach would be a univariate analysis. Where studies present prognostic information as a 2x2 table then methods used for diagnostic test accuracy studies may be appropriate.

As they are similar to diagnostic tests in a number of regards, the meta-analysis of prognostic factors face similar issues to those of diagnostic tests. Systematic reviews of prognostic factors are often affected by the difficulty in comprehensively identifying relevant studies. More so than diagnostic test accuracy studies, there is a relatively high risk of publication bias.[59] There is also a likelihood that many prognostic factors may be evaluated in a single study, but only those that show a high predictive value are reported. The selective reporting can mean that although the same prognostic factor may have been evaluated in numerous studies, it may be selectively reported giving a biased impression of its predictive power. Equally, although the relevant biomarker may be consistently used, the method of measurement may vary substantially.

Data extraction from identified studies can also be problematic because different methods of presentation may have been used. The prognostic measure, and often the outcome, are frequently measured on a continuous scale (e.g., tumour size) and may be recorded as longitudinal data. The manner in which these data are handled and presented can vary substantially. Results may be adjusted for relevant covariates, including other prognostic variables. Different studies will vary because of different choices of covariates (if any) and different methods of adjustment.

One solution is to use individual patient data (IPD), as this facilitates incorporation of detailed data specific to the individual patients. Grouped data can lose the associations between different prognostic measures that may be very important.

2.6. Assessing the quality of studies and meta-analysis

An important component of any systematic review or meta-analysis is a formal assessment of study quality, and the detailed reporting of methodology and findings. A number of initiatives have taken place with a view to improving the quality of published studies for both diagnostic accuracy studies and subsequent meta-analyses.

2.6.1. STARD

The Standards for Reporting of Diagnostic Accuracy (STARD) initiative was started with a view to improving the accuracy and completeness of reporting of studies of diagnostic accuracy.[60] In doing so, it was hoped that readers would be able to assess the potential for bias in a study, and to evaluate a study's generalisability. The STARD checklist was published in a number of journals and adopted by some as a requirement for submitting diagnostic test accuracy studies. However, the impact of the initiative on the quality of reporting has been questioned.[61;62] While the STARD initiative applies to the reporting of primary research, poor reporting can be indicative of poor study quality.

2.6.2. QUADAS

The Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool was originally developed in 2003 and subsequently refined and updated in 2011 as QUADAS-2.[53] The tool assesses study quality in four domains: patient selection, index test, reference standard, and flow and timing. Each domain is assessed in terms of risk of bias, and concerns regarding applicability (for the first three domains). Signalling questions are used to assist judgement regarding risk of bias. Application of the tool results in a judgement of

risk of bias for each study categorised as low, high, or unclear. These judgements can be used to exclude studies from the primary analysis or to guide sensitivity analyses. Although it is the only validated tool for assessing the quality of diagnostic test accuracy studies, it should be noted that QUADAS-2 does not include specific criteria for assessing comparative studies, although it is possible to adapt the tool for this purpose.[63]

2.6.3. PRISMA

Having identified relevant studies for a meta-analysis, assessed their risk of bias and undertaken evidence synthesis through meta-analysis, the results must then be reported. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement outlines an evidence-based minimum set of items for reporting in systematic reviews and meta-analyses.[64] The PRISMA checklist was designed for systematic reviews and meta-analyses in general, and the authors acknowledged that the checklist may need to be modified when the research question related to diagnostic or prognostic interventions. One of the key principles underpinning the PRISMA statement is that authors ensure that their methods are reported with sufficient clarity and transparency so that readers can critically judge the presented evidence and replicate the research. An analysis of the impact of PRISMA on the reporting of meta-analyses in diagnostic research has suggested that there are still issues in the quality of reporting such studies.[65] However, a modified version of PRISMA for the reporting of diagnostic test accuracy meta-analyses provides the best prospect of achieving good quality reporting.

2.6.4. GRADE

The Grading of Recommendations Applicability, Development and Evaluation (GRADE) approach provides a framework for considering the quality of evidence regarding interventions, and can be applied to diagnostic tests in terms of their impact on patient-relevant outcomes.[66] GRADE is often used in the context of developing clinical guidelines and recommendations regarding the appropriate use of a technology or health intervention.

2.7. Software

A variety of software packages have been used in the literature for carrying out the meta-analysis techniques described in these guidelines. In terms of the bivariate and HSROC approaches, implementations have been documented for the proprietary programmes SAS® and Stata®. Coded implementations in SAS® have been published in a number of studies, while the metandi module for Stata® computes results for both methods (without covariates). MLwiN is a package created by the University of Bristol for fitting multilevel models and can be applied to both techniques. The techniques can also be applied through free software packages R and WinBUGS. The latter programme is for analyses in a Bayesian framework and code for both methods has been published. Functions for the bivariate and HSROC methods in R are provided through a number of freely available packages. It should be noted that a variety of implementations are available in R with different default parameterisations, so users should pay careful attention to what methodology is coded into each function.

In all cases, it is critical that the user understands how the method has been implemented and what parameters can be set and what outputs are provided. Correct interpretation of the output is contingent on understanding how the computations have been carried out and whether the underlying assumptions are correct. Other than reporting convergence, most packages will give limited information on whether the pooled estimates and parameter values are valid. Application of the HSROC model can be associated with convergence problems when the sample sizes are small or there is too much

heterogeneity. It is important to examine the validity of the model and the software used should support such investigation.

Table 1. Software implementations of methods for the meta-analysis of diagnostic test accuracy studies

Software	Meta-analysis method		
	Moses-Littenburg	Hierarchical SROC	Bivariate random effects
RevMan*	✓	×	×
Meta-DiSc*	✓	×	×
SPSS®	✓	×	×
SAS®	✓	✓	✓
Stata®	✓	✓	✓
MLwiN[+]	✓	✓	✓
R*	✓	✓	✓
WinBUGS/OpenBUGS*	✓	✓	✓

Notes: * Free software; [+] free to UK academics.

Of the software packages listed in Table 1, some (RevMan, metaDisc, R, Stata®) contain specific commands with implemented versions of meta-analysis methods, while some (SAS®, SPSS®, Stata®, R) allow for the computation of the corresponding algorithms. The latter may allow for greater flexibility in how the algorithms are applied.

3. Conclusion and main recommendations

The meta-analysis of diagnostic test accuracy studies can be used to generate a more precise estimate by pooling data from a number of studies. Diagnostic test accuracy is not a measure of clinical effectiveness and improved accuracy does not necessarily imply improved patient outcomes. There are a variety of metrics available for describing diagnostic test accuracy, although the measures most commonly summarised in a meta-analysis are sensitivity and specificity (or the corresponding true positive rate and false positive rate). Due to the likelihood of a negative correlation between sensitivity and specificity, a meta-analysis of the two measures should take this relationship into account. While a number of methodological approaches are available for the meta-analysis of diagnostic test accuracy studies, the HSROC and bivariate methods are the most appropriate. These techniques have been implemented in a variety of software environments. There are numerous forms of bias that can affect estimates of diagnostic test accuracy in individual studies. All studies included in a meta-analysis should be carefully scrutinised to ensure they are equivalent and suitable for meta-analysis. Sensitivity analysis is a useful approach for testing the influence of studies with a high risk of bias.

Based on the preceding sections, a number of recommendations are proposed:

1. Pooling studies of diagnostic test accuracy should only be undertaken when there are sufficient studies available. When only two studies are available, it is not recommended to undertake a meta-analysis: reporting should be restricted to a narrative description of the available evidence.
2. The quality of studies being pooled should be assessed using a recognised and validated quality assessment tool.
3. Pooled studies should be equivalent in terms of the index test, the reference standard, the patient population and the indication.
4. Where important differences are identified across studies in terms of disease spectrum, study setting, and disease prevalence, these should be accounted for by including covariates.
5. Where potential study differences occur but cannot be readily accounted for, such as verification bias, these should be clearly identified and the potential impacts determined.
6. The appropriate methods of meta-analysis are the hierarchical SROC and bivariate random effects techniques, unless there is an absence of heterogeneity in either FPR or TPR, in which case two separate univariate meta-analyses may be more appropriate.
7. The appropriate approach to meta-analysis is defined with respect to the quantity of data, between-study heterogeneity, threshold effects, and the correlation between TPR and FPR.
8. The reporting of meta-analysis should include all the information that justifies the choice of analytical approach and supports the exclusion of alternative approaches.

Recommendations for those undertaking meta-analyses

For researchers undertaking a meta-analysis of diagnostic test accuracy studies, a minimum set of information must be reported, specifically:

1. A detailed description of the included studies in terms of both similarities and differences in key components (e.g., index test, reference test, population, indication, test threshold, prevalence).
2. The quality assessment of the included studies.
3. A clear description of the decision process that leads to the selection of the appropriate methodology.
4. All of the estimated parameter values along with their corresponding confidence or credibility intervals.
5. Appropriate graphical outputs including forest plots, SROC (if computed), and prediction regions.
6. The possible impact of different forms of bias on the results.

Recommendations for those reading meta-analyses

For those reading a meta-analysis of diagnostic test accuracy studies, certain key information must be included in order to appraise the findings:

1. Were the included studies comparable in terms of the key features (e.g., index test, reference test, population, indication, test threshold, prevalence)?
2. Were the included studies of acceptable quality?
3. Was the methodology used appropriate given the nature of the included evidence?
4. Were all of the estimated parameter values clearly reported and their values interpreted?
5. Were the relevant graphical outputs provided including forest plots, SROC (if computed), and prediction regions?
6. Were the possible effects of different forms of bias on the results clearly reported and supported with relevant sensitivity analyses?
7. Were the conclusions drawn consistent with the evidence analysed?

Annexe 1. Bibliography

(1) Knottnerus JA, van Weel C. General introduction: evaluation of diagnostic procedures. In: Knottnerus JA, editor. The evidence base of clinical diagnosis. London: BMJ Books; 2002. 81-94.
(2) Ebell MH. Evidence-based diagnosis. New York: Springer-Verlag; 2001.
(3) Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. BMJ 2001; 323(7305):157-162.
(4) Pewsner D, Battaglia M, Minder C, Marx A, Bucher, einer C. et al. Ruling a diagnosis in or out with "SpPIn" and "SnNOut": a note of caution. BMJ 2004; 329.
(5) Smits N. A note on Youden's J and its cost ratio. BMC Med Res Methodol 2010; 10(1):89.
(6) Chen L, Reisner AT, Chen X, Gribok A, Reifman J. Are standard diagnostic test characteristics sufficient for the assessment of continual patient monitoring? Med Decis Making 2013; 33(2):225-234.
(7) Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. In: Egger M, Davey Smith G, Altman DG, editors. Systematic Reviews in Health Care. London: BMJ Books; 2001. 248-284.
(8) Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ 1994; 309(6947):102.
(9) Eusebi P. Diagnostic accuracy measures. Cerebrovasc Dis 2013; 36:267-272.
(10) Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. BMJ 2012; 344:e686.
(11) Knottnerus JA, Dinant G-J, van Schayck OP. The diagnostic before-after study to assess clinical impact. In: Knottnerus JA, editor. The evidence base of clinical diagnosis. London: BMJ Books; 2002. 81-94.
(12) Staub LP, Dyer S, Lord SJ, Simes RJ. Linking the evidence: intermediate outcomes in medical test assessments. Int J Technol Assess Health Care 2012; 28(1):52-58.
(13) Broeze KA, Opmeer BC, van d, V, Bossuyt PM, Bhattacharya S, Mol BW. Individual patient data meta-analysis: a promising approach for evidence synthesis in reproductive medicine. Hum Reprod Update 2010; 16(6):561-567.
(14) Merlin T, Lehman S, Hiller JE, Ryan P. The "linked evidence approach" to assess medical tests: a critical analysis. Int J Technol Assess Health Care 2013; 29(3):343-350.
(15) 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):850-855.
(16) Jarvik JG. Fundamentals of Clinical Research for Radiologists: The Research Framework. Am J Roentgenol 2001; 176(4):873-878.
(17) Krupinski EA, Jiang Y. Anniversary paper: evaluation of medical imaging systems. Med Phys 2008; 35(2):645-659.
(18) Thornbury JR. Eugene W. Caldwell Lecture. Clinical efficacy of diagnostic imaging: love it or leave it. AJR Am J Roentgenol 1994; 162(1):1-8.
(19) Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 ed. The Cochrane Collaboration; 2011.
(20) Centre for Reviews and Dissemination. CRD’s Guidance for Undertaking Systematic Reviews. CRD; 2009. (https://www.york.ac.uk/inst/crd/SysRev/!SSL!/WebHelp/SysRev3.htm)

(21) Harbord RM, Whiting P, Sterne JAC, Egger M, Deeks JJ, Shang A et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. Journal of Clinical Epidemiology 2008; 61(11):1095-1103.
(22) Chappell FM, Raab GM, Wardlaw JM. When are summary ROC curves appropriate for diagnostic meta-analyses? Stat Med 2009; 28(21):2653-2668.
(23) Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008; 149(12):889-897.
(24) Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology 2005; 58(10):982-990.
(25) Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001; 20(19):2865-2884.
(26) Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JAC. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics 2007; 8(2):239-251.
(27) Menke J. Bivariate random-effects meta-analysis of sensitivity and specificity with SAS PROC GLIMMIX. Methods Inf Med 2010; 49(1):54-64.
(28) Macaskill P, Gatsonis CA, Deeks JJ, Harbord R, Takwoingi Y. Chapter 10: Analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration; 2010.
(29) Begg CB. Meta-analysis methods for diagnostic accuracy. Journal of Clinical Epidemiology 2008; 61(11):1081-1082.
(30) Dahabreh IJ, Trikalinos TA, Lau J, Schmid C. An Empirical Assessment of Bivariate Methods for Meta-Analysis of Test Accuracy. No. 12(13)-EHC136-EF. 2012. Rockville, Maryland, Agency for Healthcare Research and Quality.
(31) Harbord RM, Whiting P. metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression. In: Sterne JAC, Newton HJ, Cox NJ, editors. Meta-Analysis in Stata. Texas: Stata Press; 2009. 181-199.
(32) Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006; 332(7549):1089-1092.
(33) Takwoingi Y, Leeflang MMG, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Ann Intern Med 2013; 158(7):544-554.
(34) Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282(11):1061-1066.
(35) de Vet HCW, Eisinga A, Riphagen II, Aertgeerts B, Pewsner D. Chapter 7: Searching for Studies. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 0.4 ed. The Cochrane Collaboration; 2008.
(36) Doust JA, Pietrzak E, Sanders S, Glasziou PP. Identifying studies for systematic reviews of diagnostic tests was difficult due to the poor sensitivity and precision of methodologic filters and the lack of information in the abstract. J Clin Epidemiol 2005; 58(5):444-449.
(37) Whiting P, Westwood M, Beynon R, Burke M, Sterne JA, Glanville J. Inclusion of methodological filters in searches for diagnostic test accuracy studies misses relevant studies. J Clin Epidemiol 2011; 64(6):602-607.
(38) Tatsioni A, Zarin DA, Aronson N, Samson DJ, Flamm CR, Schmid C et al. Challenges in systematic reviews of diagnostic technologies. Ann Intern Med 2005; 142(12 Pt 2):1048-1055.

(39) Song F, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. Int J Epidemiol 2002; 31(1):88-95.
(40) 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. Journal of Clinical Epidemiology 2005; 58(9):882-893.
(41) Leeflang MMG, Deeks JJ, Rutjes AWS, Reitsma JB, Bossuyt PMM. Bivariate meta-analysis of predictive values of diagnostic tests can be an alternative to bivariate meta-analysis of sensitivity and specificity. Journal of Clinical Epidemiology 2012; 65(10):1088-1097.
(42) Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: Methodologic primer. Am J Roentgenol 2006; 187(2):271-281.
(43) Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med 2002; 137(7):598-602.
(44) de Groot JA, Dendukuri N, Janssen KJ, Reitsma JB, Brophy J, Joseph L et al. Adjusting for partial verification or workup bias in meta-analyses of diagnostic accuracy studies. Am J Epidemiol 2012; 175(8):847-853.
(45) Ringham BM, Alonzo TA, Grunwald GK, Glueck DH. Estimates of sensitivity and specificity can be biased when reporting the results of the second test in a screening trial conducted in series. BMC Med Res Methodol 2010; 10:3.
(46) Leeflang MMG, Moons KGM, Reitsma JB, Zwinderman AH. Bias in Sensitivity and Specificity Caused by Data-Driven Selection of Optimal Cutoff Values: Mechanisms, Magnitude, and Solutions. Clinical Chemistry 2008; 54(4):729-737.
(47) Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. Journal of Clinical Epidemiology 2006; 59(8):798-801.
(48) Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. CMAJ 2013.
(49) Kuss O, Hoyer A, Solms A. Meta-analysis for diagnostic accuracy studies: a new statistical model using beta-binomial distributions and bivariate copulas. Stat Med 2014; 33(1):17-30.
(50) Ma X, Nie L, Cole SR, Chu H. Statistical methods for multivariate meta-analysis of diagnostic tests: An overview and tutorial. Stat Methods Med Res 2013.
(51) van Walraven C, Austin PC, Jennings A, Forster AJ. Correlation between serial tests made disease probability estimates erroneous. Journal of Clinical Epidemiology 2009; 62(12):1301-1305.
(52) Novielli N, Cooper NJ, Sutton AJ. Evaluating the Cost-Effectiveness of Diagnostic Tests in Combination: Is It Important to Allow for Performance Dependency? Value in Health 2013; 16(4):536-541.
(53) Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011; 155(8):529-536.
(54) Schuetz GM, Schlattmann P, Dewey M. Use of 3x2 tables with an intention to diagnose approach to assess clinical performance of diagnostic tests: meta-analytical evaluation of coronary CT angiography studies. BMJ 2012; 345:e6717.
(55) Khan KS, Bachmann LM, ter Riet G. Systematic reviews with individual patient data meta-analysis to evaluate diagnostic tests. European Journal of Obstetrics & Gynecology and Reproductive Biology 2003; 108(2):121-125.
(56) Broeze KA, Opmeer BC, Coppus SFPJ, Van Geloven N, Alves MFC, Ã…nestad G et al. Chlamydia antibody testing and diagnosing tubal pathology in subfertile women: an individual patient data meta-analysis. Hum Reprod Update 2011; 17(3):301-310.
(57) Broeze KA, Opmeer BC, Coppus SF, Van Geloven N, Den Hartog JE, Land JA et al. Integration of patient characteristics and the results of Chlamydia antibody

testing and hysterosalpingography in the diagnosis of tubal pathology: an individual patient data meta-analysis. Human Reproduction 2012; 27(10):2979-2990.
(58) Loy CT, Irwig L. Accuracy of diagnostic tests read with and without clinical information: a systematic review. JAMA 2004; 292(13):1602-1609.
(59) Altman DG. Systematic reviews of evaluations of prognostic variables. In: Egger M, Davey Smith G, Altman DG, editors. Systematic Reviews in Health Care. London: BMJ Books; 2001. 228-247.
(60) Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003; 326(7379):41-44.
(61) Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB et al. The quality of diagnostic accuracy studies since the STARD statement: has it improved? Neurology 2006; 67(5):792-797.
(62) Wilczynski NL. Quality of reporting of diagnostic accuracy studies: no change since STARD statement publication--before-and-after study. Radiology 2008; 248(3):817-823.
(63) Wade R, Corbett M, Eastwood A. Quality assessment of comparative diagnostic accuracy studies: our experience using a modified version of the QUADAS-2 tool. Res Syn Meth 2013; 4(3):280-286.
(64) Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Ann Intern Med 2009; 151(4):W65-W94.
(65) Willis BH, Quigley M. The assessment of the quality of reporting of meta-analyses in diagnostic research: a systematic review. BMC Med Res Methodol 2011; 11:163.
(66) Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008; 336(7653):1106-1110.

Annexe 2. Documentation of literature search

Keywords

Five keywords were defined to enable identification of relevant literature:

- diagnostic
- test
- accuracy
- meta-analysis
- systematic

Search engines and sources of information

A variety of sources of information were identified to find published literature and information pertinent to the development of these guidelines.

Literature search

- EMBASE
- MEDLINE
- DARE
- Cochrane Database of Systematic Reviews
- CADTH/CEDAC
- EBSCOhost

Internet search

- Google and Google Scholar
- ScienceDirect
- Wiley-Interscience
- Hand searching of references cited in relevant documents
- The Cochrane Collaboration
- National Guideline Clearinghouse
- National Institute for Health and Clinical Excellence
- ISPOR
- Pharmaceutical Benefits Advisory Committee (PBAC)
- Centre for Reviews and Dissemination, University of York
- University of Bristol

Guidelines search

The websites of EUnetHTA member agencies and those of major international agencies were searched for relevant guidelines.

Other specifically identified sources of information

- Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006; 332: 1089-1092.
- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med. 2003; 138(1):W1-12.
- Chappell FM, Raab GM, Wardlaw JM. When are summary ROC curves appropriate for diagnostic meta-analyses? Statistics in Medicine. 2009; 28(21): 2653-2668.
- Harbord RM, Whiting P, Sterne JAC, Egger M, Deeks JJ, Shang A, et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. Journal of Clinical Epidemiology. 2008; 61: 1095-1103.
- Moher D, Liberati A, Tetzlaff J, Altman DG, the PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. Ann Intern Med. 2009; 151(4):264-269.
- Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary roc curve: Data-analytic approaches and some additional considerations. Statistics in Medicine. 1993; 12: 1293-1316.
- Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology. 2005; 58: 982-990.
- Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Statistics in Medicine. 2001; 20(19): 2865-2884.
- Simel DL, Bossuyt PMM. Differences between univariate and bivariate models for summarizing diagnostic accuracy may not be large. Journal of Clinical Epidemiology. 2009; 62(12): 1292-1300.
- Sutton AJ, Higgins JPT. Recent developments in meta-analysis. Statistics in Medicine. 2008; 27: 625-650.
- Tawoingi 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.
- Verde P. Meta-analysis of diagnostic test data: A bivariate Bayesian modeling approach. Statistics in Medicine. 2010; 29: 3088-3102
- Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. In: Egger M, Davey Smith G, Altman DG (eds). Systematic Reviews in Health Care. BMJ Books. London, 2001.
- Diagnostic Test Accuracy Working Group. Handbook for DTA Reviews Version 1.0.1. Cochrane Collaboration, 2009.
- Harbord RM, Whiting P. Metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression. In: Sterne JAC (ed). Meta-Analysis in Stata – An Updated Collection from the Stata Journal. Stata Press. Texas, 2009.
- Health Information and Quality Authority. Guidelines for Evaluating the Clinical Effectiveness of Health Technologies in Ireland. HIQA. Dublin, 2011.

Strategies of research

Reports, papers and other guidance documents were assessed on the basis of whether they described, applied or assessed methods of meta-analysis for diagnostic test accuracy

studies. Documents that only mentioned methods but did not describe, apply or assess them were disregarded after being checked for useful references. Documents that applied methods were used to determine the scope of application, utility and possible limitations of those methods. Finally, documents that assessed methods were used to compare methods directly and to elicit recommendations. Where relevant, the quality of studies was assessed using the STARD (Standards for Reporting for Reporting of Diagnostic Accuracy) or PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statements.

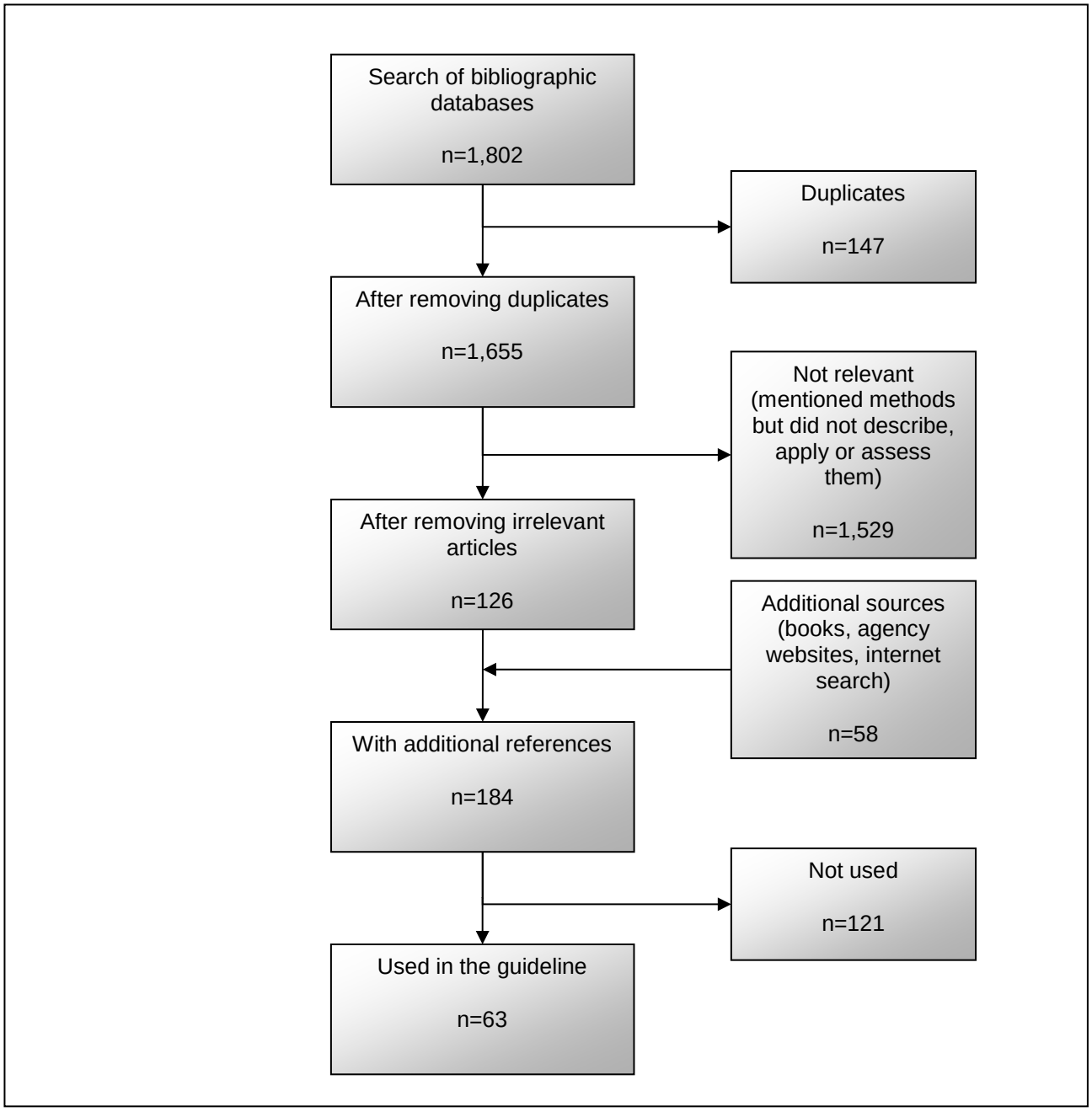
For PubMed, the search was limited to the period 1990 to date (end June 2013). In EBSCO the search was limited to 1990 to 2013 (inclusive). In both cases the search was limited to English language publications and human subjects. Database searches used the following search strategy:

> (diagnostic[Title/Abstract]) AND test[Title/Abstract]) AND accuracy[Title/Abstract]) AND (meta-analysis[Title/Abstract]) OR systematic[Title/Abstract])

Findings of literature search

The initial search returned 1,802 articles that were potentially useful. After scanning titles, abstracts and, in some cases, full text, 126 articles were retained. Of these, 63 were ultimately used and referenced in the guidelines.

Figure 7. Flowchart of literature search.



Annexe 3. Other sources of information

No other sources of information were used.