Meta-analysis of diagnostic accuracy studies in mental health

Takwoingi, Yemisi; Riley, Richard D; Deeks, Jonathan J

DOI:
10.1136/eb-2015-102228

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Takwoingi, Y, Riley, RD & Deeks, JJ 2015, 'Meta-analysis of diagnostic accuracy studies in mental health', *Evidence - Based Mental Health*, vol. 18, no. 4, pp. ebmental-2015-102228. https://doi.org/10.1136/eb-2015-102228

Evidence-Based Mental Health Online First, published on October 7, 2015 as 10.1136/eb-2015-102228

Meta-analysis of diagnostic accuracy studies in mental health

Yemisi Takwoingi,[1] Richard D Riley,[2] Jonathan J Deeks[1]

[1]Public Health, Epidemiology and Biostatistics, University of Birmingham, Birmingham, UK; [2]Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire, UK

Correspondence to Dr Yemisi Takwoingi, y.takwoingi@bham.ac.uk

ABSTRACT

Objectives To explain methods for data synthesis of evidence from diagnostic test accuracy (DTA) studies, and to illustrate different types of analyses that may be performed in a DTA systematic review.

Methods We described properties of meta-analytic methods for quantitative synthesis of evidence. We used a DTA review comparing the accuracy of three screening questionnaires for bipolar disorder to illustrate application of the methods for each type of analysis.

Results The discriminatory ability of a test is commonly expressed in terms of sensitivity (proportion of those with the condition who test positive) and specificity (proportion of those without the condition who test negative). There is a trade-off between sensitivity and specificity, as an increasing threshold for defining test positivity will decrease sensitivity and increase specificity. Methods recommended for meta-analysis of DTA studies –such as the bivariate or hierarchical summary receiver operating characteristic (HSROC) model –jointly summarise sensitivity and specificity while taking into account this threshold effect, as well as allowing for between study differences in test performance beyond what would be expected by chance. The bivariate model focuses on estimation of a summary sensitivity and specificity at a common threshold while the HSROC model focuses on the estimation of a summary curve from studies that have used different thresholds.

Conclusions Meta-analyses of diagnostic accuracy studies can provide answers to important clinical questions. We hope this article will provide clinicians with sufficient understanding of the terminology and methods to aid interpretation of systematic reviews and facilitate better patient care.

INTRODUCTION

Diagnostic accuracy is the ability of a test to correctly identify or exclude a target condition and is a fundamental part of the evaluation of medical tests. Test accuracy is estimated by comparing results of an index test (a new or existing test of interest) with a reference standard, sometimes known as a 'gold' standard, to give a 2×2 table of the number of true positives, false positives, false negatives and true negatives (table 1). The reference standard is used to verify the presence or absence of the target condition, and may be a single test or a combination of tests.[1] The ideal index test should have no misclassification errors (false negatives or false positives), but in clinical practice such a test is unlikely to exist.

Many test accuracy studies are small,[2] and even when studies are large, the number of cases may be small due to the low prevalence of the target condition. A systematic review and meta-analysis of diagnostic test accuracy (DTA) aims to identify and summarise evidence on the accuracy of tests, including an assessment of the quality and consistency of the evidence. Pooling the results of multiple studies addressing the same question using meta-analysis will provide a more precise estimate of test performance than is possible in a single study. The extent of variability in test performance between studies (heterogeneity) can be quantified in a meta-analysis and formal investigations of potential sources of heterogeneity may also be performed in order to explain why results differ between studies.

The rate of publication of systematic reviews and meta-analyses of diagnostic accuracy has risen considerably since the 1990s,[3] and they are being used to inform evidence-based clinical practice guidelines and healthcare policy.[4–7] Producing DTA reviews is more complex than reviews of interventions.[8 9] Meta-analysis is one of the challenging aspects because traditional meta-analytic methods used for intervention reviews are not appropriate for DTA reviews. It is vital that recommended methods for pooling study results are well understood to ensure appropriate application. In this paper, we summarise basic concepts in diagnostic accuracy research as a prelude to explaining the rationale for the methods recommended for DTA meta-analysis, describe the properties of the methods, and use a published example to illustrate their application in different types of analyses. We focus on commonly used methods for situations where a single 2×2 table is available, or can be derived for each study included in a meta-analysis.

BASIC CONCEPTS IN DIAGNOSTIC ACCURACY RESEARCH

It is important to be aware of both the types of data that tests produce and the statistical measures used to summarise test accuracy.

Types of data

Test results may be expressed as measurements (counts or continuous) or classifications (binary or ordered categories). Standard methods for computing test accuracy demand binary classification of the results of the index test and the reference standard (table 1). As such, for non-binary data, criteria for determining test positivity, typically referred to as thresholds, cut-offs or cut points, are needed to dichotomise the data. Thresholds may be explicit numeric values or implicit based on different criteria derived from subjective visual interpretation or measurements. For example, various numeric thresholds have been used to define a positive result for the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) for detection of dementia.[10] Higher scores are indicative of greater cognitive impairment, with a maximum average score of 5. In a Cochrane DTA review, Quinn *et al*[10] described the test accuracy of the 16-item and 26-item IQCODE for thresholds of 3.3, 3.4, 3.5 and 3.6.

Types of measures of test accuracy

The most commonly used measures are summarised in table 2. Measures are either paired or single (global) indicators of test performance. The paired measures—sensitivity and specificity, positive and negative predictive values, and positive and negative likelihood ratios—separately describe the performance of a test for ascertaining first the presence and then the absence of the target condition. Sensitivity and specificity are the most commonly reported measures in primary studies and are also used for meta-analysis.[3] A receiver operating characteristic (ROC) plot is used to show how as the test threshold decreases (for a test where the presence of disease increases the value of the test measurement such as the IQCODE example above) sensitivity increases while specificity decreases, and vice versa.[11] The position of the ROC curve depends on the discriminatory ability of the test; the more accurate the test is, the closer the curve to the upper left hand corner of the ROC plot. A test that performed no better than tossing a coin would have an ROC curve along the 45° axis.

Table 1 Cross classification of index test and reference standard results

	Reference standard positive	Reference standard negative	Total
Index test positive	a (true positives)	b (false positives)	a+b (test positives)
Index test negative	c (false negatives)	d (true negatives)	c+d (test negatives)
Total	a+c (disease positives)	b+d (disease negatives)	a+b+c+d (study total)

The most common global measures are the diagnostic OR (DOR) and the area under the curve (AUC). These measures summarise the accuracy of the test across all possible thresholds, but are not helpful in clinical practice because they do not provide information on the error rates in the diseased (false negative) and non-diseased groups (false positives). The error rates are important for judging the extent and likely impact of the downstream consequences of testing. In meta-analysis, the DOR can be a useful measure when comparing tests or subgroups, particularly if there is no preference for either superior sensitivity or specificity, and interest is in global performance.

TYPES OF ANALYSES IN A DIAGNOSTIC ACCURACY REVIEW

Three main types of analyses can be performed in a DTA review reflecting the types of questions and objectives that can be addressed.[11] These are:

1. Analysis of a single test—What is the diagnostic accuracy of the test?
2. Analysis comparing multiple tests—How does the accuracy of two or more tests differ?
3. Investigation of heterogeneity—How does test accuracy vary with clinical and methodological characteristics?

Each of these will be considered separately below, using an example. First, the broad principles of the meta-analysis approaches are discussed.

Table 2 Definition of common measures of test accuracy

Test accuracy measure	Formula*	Definition
Sensitivity	a/(a+c)	Proportion of those with the target condition correctly identified as having the condition
Specificity	d/(b+d)	Proportion of those without the target condition correctly identified as not having the condition
Positive predictive value	a/(a+b)	Proportion of those with the target condition out of the test positives
Negative predictive value	d/(c+d)	Proportion of those without the target condition out of the test negatives
Positive likelihood ratio (LR^+)	$\frac{a/(a+c)}{b/(b+d)}$	Ratio of the proportion that tests positive among those with the target condition compared to the proportion that tests positive among those without the target condition
Negative likelihood ratio (LR^-)	$\frac{c/(a+c)}{d/(b+d)}$	Ratio of the proportion that tests negative among those with the target condition compared to the proportion that tests negative among those without the target condition
Diagnostic OR	ad/bc or LR^+/LR^-	Ratio of the odds of positivity in those who have the target condition relative to the odds of positivity in those without the condition

*Expressed based on the notation used in table 1.

METHODS FOR META-ANALYSIS: BROAD PRINCIPLES

We recommend first plotting the data—the results of each study can be marked as a sensitivity and specificity point in a summary ROC (SROC) plot which will depict the location of the data (consider how close the points lie to the 100% sensitivity and 100% specificity lines); the scatter of the points; and any relationship between sensitivity and specificity obvious across multiple studies. Usually only one point is plotted per study, but see the multiple thresholds discussion later. Since the different studies included in a meta-analysis may explicitly use different thresholds, or variations in the way the test is interpreted and applied, and even the patient group may induce threshold-like differences between studies, the recommended meta-analytical methods for test accuracy all explicitly or implicitly allow for the negative correlation between sensitivity and specificity across studies that is induced by the threshold-like variation. Simple univariate meta-analytic methods pool sensitivity and specificity separately, ignoring the potential threshold effect. Such analyses can give misleading results as illustrated by Irwig *et al*.[12] An SROC curve approach was developed by Moses *et al*[13] to account for possible heterogeneity in threshold, but the approach assumes that variation is only due to the threshold effect and chance and does not allow for heterogeneity. Measures of test accuracy are not fixed properties of a test and often vary with the population, setting, characteristics and conduct of the test (including skill and experience of assessors or practitioners), and definition of the target condition. Therefore, heterogeneity is a common feature of DTA reviews. The Moses SROC approach also has other important methodological limitations[14 15] and should only be used for preliminary analyses.[11]

Special hierarchical models have been developed for DTA meta-analysis that account for the negative correlation in paired measures across studies and heterogeneity. These models are termed hierarchical models because they involve statistical distributions at two levels; within-study variability in sensitivity and specificity is taken into account at the lower level, and between-study variability at the higher level.[11] The bivariate[16 17] and the hierarchical summary receiver operating characteristic (HSROC) models[18] are the two hierarchical models used for meta-analysis when a single sensitivity and specificity pair is available for each study.[8 11]

The focus of the bivariate model is estimation of a summary point (summary sensitivity and specificity) at a common threshold. This would be useful if studies used a standard threshold and clinicians wanted to know how the test performed at that threshold (eg, the Mini-Mental State Examination at a threshold of 24/30 indicating normal). In contrast, the focus of the HSROC model is on estimating an SROC curve across different thresholds (note this is not the same curve as the Moses SROC curve). This would be useful if studies used a variety of thresholds and clinicians wanted to get an idea of how the test performed across the range of thresholds, or wanted to compare tests without restricting comparison to a single threshold. Ideally, we would like to know which threshold on the curve gives the best performance, but the position of individual thresholds cannot be identified. The bivariate and HSROC models have been shown to share statistical properties.[19] The choice of which model to use should ideally be driven by the research question (ie, focus of inference on points or curves) but is sometimes influenced by the nature of the available data (mixed thresholds) and its effect on the interpretation of summary findings, software capability and expertise of the team. Owing to their shared statistical properties, SROC curves can be computed from bivariate models and average operating points from HSROC models, so the choice of model when there is only a single test is academic. When there are comparisons between tests or subgroups, the choice of model is important.

Bivariate meta-analysis of likelihood ratios and predictive values are alternatives to bivariate meta-analysis of sensitivity and specificity but have been noted to face additional challenges.[20 21] Therefore, only

methods for summarising sensitivities and specificities are illustrated in this paper. Other measures such as likelihood ratios can be derived from functions of some of the parameters of bivariate or HSROC models shown in table 3. The example below is used to illustrate different types of analyses and the appropriate use of hierarchical models.

WORKING EXAMPLE: SCREENING TESTS FOR BIPOLAR SPECTRUM DISORDERS

A DTA systematic review compared the diagnostic accuracy of three screening questionnaires—the bipolar spectrum diagnostic scale (BSDS), the hypomania checklist (HCL-32) and the mood disorder questionnaire (MDQ)—for detection of bipolar disorder (table 4).[22] Studies used different thresholds to define test positivity for each instrument. The reference standards used were based on the Diagnostic and Statistical Manual of Mental Disorders 4th Edition (DSM-IV) criteria. The review included analyses for different settings and types of bipolar disorder. Only the analysis of bipolar disorder in general in a mental health centre setting is considered here (figure 1).

METHODS FOR META-ANALYSIS OF A SINGLE TEST

Data synthesis

Estimation of a summary sensitivity and specificity at a fixed threshold

The bivariate model jointly synthesises sensitivity and specificity to give summary estimates which are drawn as the summary point on an SROC plot. Confidence and prediction regions plotted around the summary point enable joint inferences to be made about sensitivity and specificity. These regions account for correlation between sensitivity and specificity, and are useful for illustrating uncertainty around the summary point and the extent of heterogeneity. Since summary points should only be calculated when studies share a common threshold, the available data are reduced. The choice of a common threshold is often based on the available data and may not be the threshold used in clinical practice. Furthermore, a common threshold is difficult to define for non-numeric tests.

The summary point for the MDQ can be estimated by meta-analysis of only studies that used the recommended threshold of 7. This restriction reduces the data for meta-analysis from 30 to 19 studies. The summary sensitivity and specificity were 0.65 (0.57 to 0.72) and 0.79 (0.72 to 0.84), respectively. Figure 2 shows this summary point with a 95% confidence region and 95% prediction region. The confidence region is based on the CI around the summery point and indicates that, on the basis of the available data, we would expect the 'real value' to be within that region 95% of the time. The prediction region around the summary point indicates the region where we would expect results from a new study in the future to lie, and is therefore wider than the confidence region as it goes beyond the uncertainty in the available data. Despite the use of a common threshold, considerable heterogeneity is evident from the plot.

Table 3 Basic parameters of bivariate and HSROC models

HSROC model	Bivariate model
Mean accuracy	Mean logit sensitivity
Mean threshold	Mean logit specificity
Variance of random effects for accuracy	Variance of random effects for logit sensitivity
Variance of random effects for threshold	Variance of random effects for logit specificity
Shape of SROC curve	Correlation between the logits of sensitivity and logits of specificity

Each of the models has five parameters when no covariates are included.
HSROC, hierarchical summary receiver operating characteristic; SROC, summary ROC.

Table 4 Characteristics of the index tests in the example systematic review

Characteristic	Index tests		
	Bipolar spectrum diagnostic scale	Hypomania checklist	Mood disorder questionnaire
Number of studies	8	17	30
Range of total score	0–25 points	0–32 points	0–15 points
Recommended thresholds*	13	14	7

*Thresholds recommended by the developers of each screening instrument.[32–34]

Estimation of a summary curve

The HSROC model focuses on estimation of an SROC curve. The advantage of this approach is that data from each study can be included irrespective of the threshold used, thus maximising use of the available data. Note that only one 2×2 table per study is included in a meta-analysis, and therefore a choice needs to be made for studies that report multiple thresholds.[11] Although a summary point on an SROC curve estimated using mixed thresholds is clinically uninterpretable, estimates of sensitivity and their CIs can be computed from the HSROC model at fixed values (eg, lower quartile, median and upper quartile) of specificity, or vice versa, to illustrate changes in sensitivity and specificity along the curve.

Figure 3 shows the summary curve for the MDQ. To avoid extrapolation beyond the data, the curve was drawn within the range of observed specificities (0.47 to 1.00) from the 30 included studies. For illustration of a point on the curve, the estimated sensitivity was 0.70 (0.64 to 0.77) at the median specificity of 0.77 from the included studies. Given the relationship between the bivariate and HSROC models mentioned earlier,[19] it is possible to estimate summary sensitivity and specificity by estimating an average operating point on the SROC curve. This analysis can be performed separately for studies that report data at a common threshold. For instance, the summary sensitivity and specificity from the HSROC method restricted to studies with threshold 7 are the same as those obtained from the bivariate model in the preceding section.

Measuring and explaining heterogeneity

The traditional I^2 statistic[23] is not recommended for quantifying heterogeneity in sensitivity and specificity because it is a univariate measure that does not account for potential threshold effects. To investigate whether a factor is associated with test accuracy, exploratory analyses can be performed by visual inspection of forest plots and SROC plots. Where formal investigations are possible, meta-regression can be performed by adding the factor as a covariate to a hierarchical model. If there are mixed thresholds, the analysis based on an HSROC model fitted to all relevant studies will have more power than an analysis at a fixed threshold. Owing to the complexity of hierarchical models and paucity of data, it is usual to assess the effect of covariates one at a time.

The bivariate model allows covariates to affect sensitivity, specificity or both. The HSROC model allows covariates to affect the accuracy, threshold and/or shape of the SROC curve. If SROC curves are assumed to have the same shape (ie, parallel curves), differences in test performance can be expressed as the relative diagnostic OR (rDORs) comparing the DOR of one group of studies to that of another. The fit of alternative models (effect of adding or removing covariate terms from the bivariate or HSROC model) can be assessed using statistical tests to compare models with and without covariate terms. Covariates to be

Study	TP	FP	FN	TN	Threshold	Asian	Sensitivity (95% CI)	Specificity (95% CI)
Wang 2009	5	22	1	25	2	Yes	0.83 [0.36, 1.00]	0.53 [0.38, 0.68]
Hu 2012	93	94	216	1084	3	Yes	0.30 [0.25, 0.36]	0.92 [0.90, 0.94]
Poon 2012	20	63	11	211	4	Yes	0.65 [0.45, 0.81]	0.77 [0.72, 0.82]
Gan 2012	45	16	18	43	4	Yes	0.71 [0.59, 0.82]	0.73 [0.60, 0.84]
Zaratiegui 2011	184	0	157	13	5	No	0.54 [0.49, 0.59]	1.00 [0.75, 1.00]
Shabani 2009	71	20	42	48	5	No	0.63 [0.53, 0.72]	0.71 [0.58, 0.81]
Nagata 2013	10	20	5	43	5	Yes	0.67 [0.38, 0.88]	0.68 [0.55, 0.79]
Carta 2006	18	14	6	85	6	No	0.75 [0.53, 0.90]	0.86 [0.77, 0.92]
Hardoy 2005	35	19	11	89	6	No	0.76 [0.61, 0.87]	0.82 [0.74, 0.89]
Lin 2011	81	10	14	65	6	Yes	0.85 [0.77, 0.92]	0.87 [0.77, 0.93]
van Zaane 2012	15	58	20	77	7	No	0.43 [0.26, 0.61]	0.57 [0.48, 0.66]
Miller 2004	21	12	16	24	7	No	0.57 [0.39, 0.73]	0.67 [0.49, 0.81]
Zimmerman 2009	33	64	19	364	7	No	0.63 [0.49, 0.76]	0.85 [0.81, 0.88]
Gervasoni 2009	28	17	16	85	7	No	0.64 [0.48, 0.78]	0.83 [0.75, 0.90]
Konuk 2007	23	63	13	210	7	No	0.64 [0.46, 0.79]	0.77 [0.71, 0.82]
Nallet 2013	22	27	11	92	7	No	0.67 [0.48, 0.82]	0.77 [0.69, 0.84]
Leao 2012	50	47	23	80	7	No	0.68 [0.57, 0.79]	0.63 [0.54, 0.71]
Soares 2010	57	18	24	24	7	No	0.70 [0.59, 0.80]	0.57 [0.41, 0.72]
de Sousa Gurgel 2012	36	6	15	96	7	No	0.71 [0.56, 0.83]	0.94 [0.88, 0.98]
de Dios 2008	8	13	3	63	7	No	0.73 [0.39, 0.94]	0.83 [0.73, 0.91]
Hirschfeld 2000	80	9	29	80	7	No	0.73 [0.64, 0.81]	0.90 [0.82, 0.95]
Rouget 2005	40	4	14	38	7	No	0.74 [0.60, 0.85]	0.90 [0.77, 0.97]
Meyer 2011	112	125	28	223	7	No	0.80 [0.72, 0.86]	0.64 [0.59, 0.69]
Isometsa 2003	17	9	3	8	7	No	0.85 [0.62, 0.97]	0.47 [0.23, 0.72]
Bech 2011	52	7	7	56	7	No	0.88 [0.77, 0.95]	0.89 [0.78, 0.95]
Chung 2009	0	25	6	83	7	Yes	0.00 [0.00, 0.46]	0.77 [0.68, 0.84]
Kim 2008	17	12	42	40	7	Yes	0.29 [0.18, 0.42]	0.77 [0.63, 0.87]
Chung 2008	28	2	34	38	7	Yes	0.45 [0.32, 0.58]	0.95 [0.83, 0.99]
Chou 2012	5	12	2	40	7	Yes	0.71 [0.29, 0.96]	0.77 [0.63, 0.87]
Castelo 2010b	63	13	6	32	8	No	0.91 [0.82, 0.97]	0.71 [0.56, 0.84]

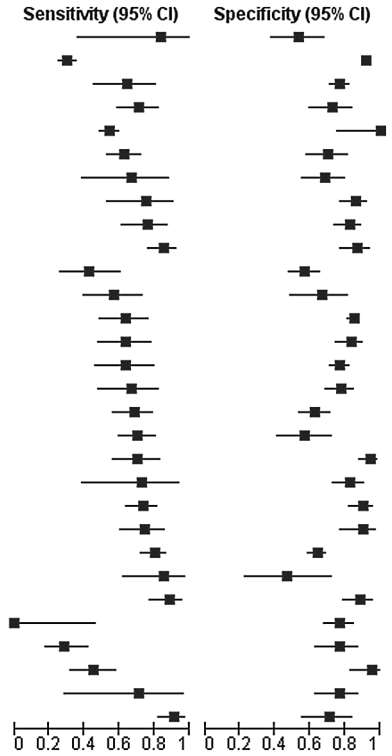

Figure 1 Forest plot of sensitivity and specificity of the MDQ for detection of any type of bipolar disorder in mental health centre settings . Two covariates (threshold and language of the instrument) are shown on the plot. The studies are ordered according to the threshold, language of the instrument (Asian, yes or no) and sensitivity (FN, false negative; FP, false positive; MDQ, mood disorder questionnaire; TN, true negative; TP, true positive).

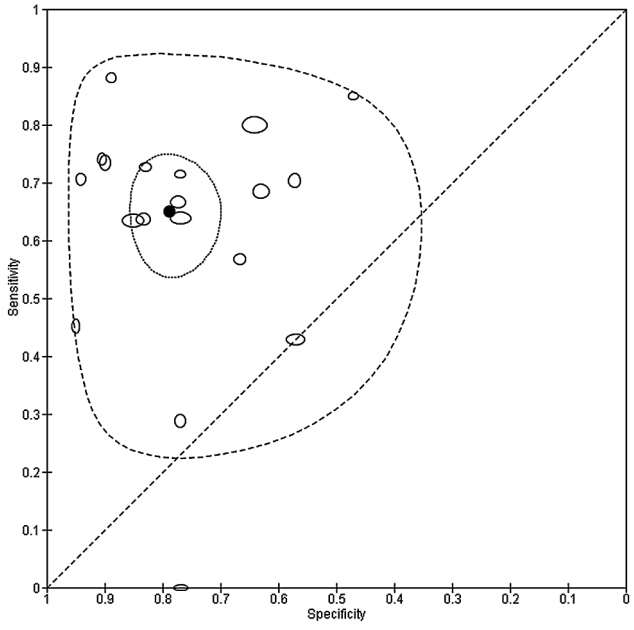

Figure 2 Summary receiver operating characteristic plot of the mood disorder questionnaire (MDQ) at a common threshold of 7 for detection of any type of bipolar disorder in mental health centre settings. The size of each point is scaled according to the precision of sensitivity and specificity for the study. The solid circle (summary point) represents the summary estimate of sensitivity and specificity for the MDQ at a threshold of 7. The summary point is surrounded by a dotted line representing the 95% confidence region and a dashed line representing the 95% prediction region (the region within which we are 95% certain that the results of a new study will lie).

investigated should be prespecified with a clear justification for their selection.

The effect of language (Asian vs non-Asian) on the accuracy of the MDQ was assessed by comparing SROC curves for the two subgroups of the covariate in one HSROC model (figure 4). Ten studies used an Asian language version while 20 studies used a non-Asian language version of the MDQ (figure 1). The model was fitted by including covariate terms for accuracy and threshold, but the curves were assumed to have the same shape. The estimated DOR (95% CI) of Asian and non-Asian language versions were 5.01 (2.64 to 9.51) and 9.14 (5.90 to 14.1), respectively. The rDOR (95% CI) of 0.55 (0.25 to 1.19) indicates that the DOR of Asian versions was 0.55 times that of the non-Asian versions, though there was no statistical evidence of a difference in accuracy ($p=0.13$).

METHODS FOR COMPARISONS OF MULTIPLE TESTS

Most DTA reviews only evaluate the accuracy of a single test, yet reviews comparing the accuracy of two or more tests are likely to be of greater relevance to clinicians and policymakers where decisions need to be made about selecting tests for use in practice. Different analytic strategies and meta-analytic models can be used for test comparisons. These are explained below.

Test comparison strategy

Ideally, test comparisons should focus on studies that have directly compared the index tests evaluated in the review.[8 24] Direct comparisons are likely to ensure an unbiased comparison, but such analyses are not always feasible due to the limited availability of comparative studies.[24] An indirect comparison uses all eligible studies that have evaluated at least one of the tests of interest, thus maximising use of the available data. However, the difference in accuracy is prone to

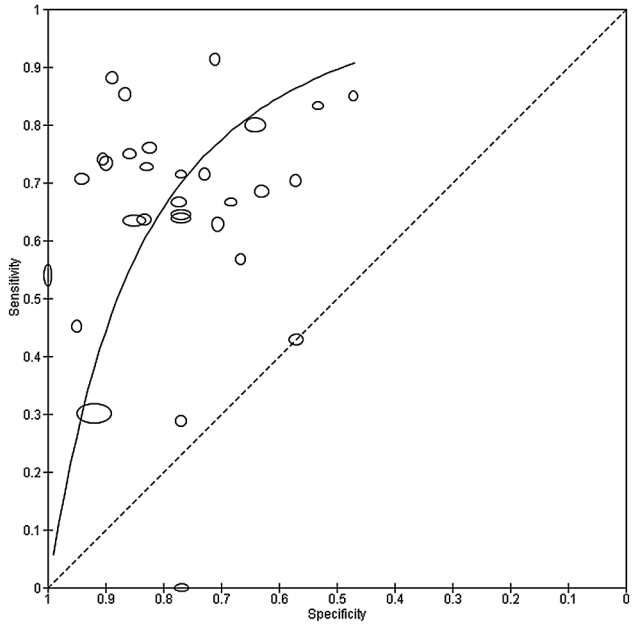

Figure 3 Summary receiver operating characteristic plot of the mood disorder questionnaire (MDQ) at different thresholds for detection of any type of bipolar disorder in mental health centre settings. Each study point was scaled according to the precision of sensitivity and specificity for the study. The summary curve was drawn restricted to the range of specificities (0.47 to 1.00) from the 30 studies included in the evaluation of the MDQ.

confounding due to differences in patient and study characteristics. A comparative review may include indirect, direct or both types of comparisons.

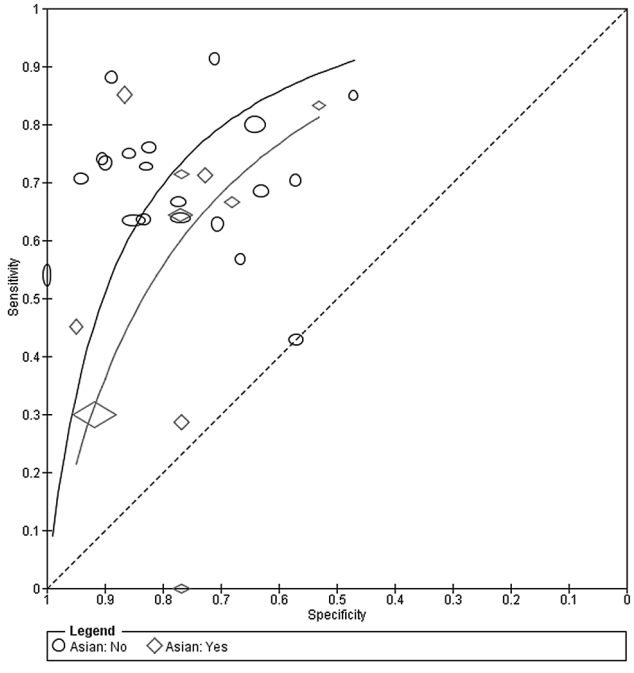

Figure 4 Summary receiver operating characteristic curves comparing Asian and non-Asian language versions of the mood disorder questionnaire. Each study point was scaled according to the precision of sensitivity and specificity for the study. The summary curves were drawn restricted to the range of specificities for each group of studies (0.47 to 1.00 for non-Asian and 0.53 to 0.95 for Asian studies).

Comparison of summary points or SROC curves

Test comparisons may be based on a comparison of summary points and/or SROC curves. Although points can be estimated at different common thresholds for each test, the ranking of tests may not be consistent at different thresholds, and therefore a comparison of curves may be more appropriate in such situations. This is evident in figure 5A which shows that the SROC curves for the three tests cross, indicating that no test is consistently more accurate than any of the others and relative accuracy depends on the choice of threshold. Current methods recommended for meta-analysis of test comparisons use a meta-regression approach, by including test type as a covariate in a hierarchical model to assess differences in test accuracy.[11] This is the same as the meta-regression approach described earlier for investigation of heterogeneity. Covariate terms can be allowed to depend on any of the model parameters shown in table 3, including the variance parameters. For an example of a comparison of summary points, see Abba *et al*.[25] A comparison of SROC curves is illustrated below.

Comparison of the MDQ, BSDS and HCL-32

The diagnostic accuracy of the three tests was compared using an HSROC meta-regression model. For a full description of the methods and interpretation of the findings, see Carvalho *et al*.[22] All 44 studies of the three tests were included in an indirect comparison (figure 5A). There was statistical evidence that the shape of the SROC curves differed, implying that the relative accuracy of the tests varies with thresholds. For the direct comparisons, eight studies compared the MDQ and HCL-32 (figure 5B panel B), and three studies compared the BSDS and MDQ. None of the 44 studies directly compared the HCL-32 and BSDS. Although the studies used different thresholds, the results from the eight studies were consistent with the HCL-32 showing higher sensitivity and lower specificity than the MDQ as shown in figure 5B. The connecting line between a pair of points identifies the results for the pair of tests from each study. A meta-analysis was performed comparing the MDQ and HCL-32. There was no evidence of a difference in accuracy.

SOFTWARE OPTIONS AND EXTENSIONS

The statistical packages available for fitting the HSROC model are currently limited to SAS, WinBUGS and R. These packages can also be used to fit the bivariate model, as can Stata and MLWin. Review Manager, the Cochrane Collaboration's review authoring tool, was used to produce the forest plot and SROC plots shown in this paper. Traditionally, the SROC plot is a plot of sensitivity against 1–specificity but because interest is often in sensitivity and specificity, Review Manager plots sensitivity against specificity on a reversed scale. Review Manager cannot be used to fit either hierarchical model but is useful for exploratory analyses, and for producing forest plots and SROC plots using parameters from hierarchical models fitted using one of the statistical packages. SAS programs for fitting hierarchical models for analysis of a single test, test comparisons and investigations of heterogeneity are available in the Cochrane Handbook for DTA reviews.[11] An SAS macro[26] that makes fitting the models more accessible and a tutorial describing how to use RevMan with Stata for different types of analyses is available at http://dta.cochrane.org/. Other user written programs are available for Stata (eg, metandi, midas) and R (eg, bamdit, DiagMeta, mada, HSROC). Extensions to the hierarchical methods discussed are emerging, in particular to deal with multiple thresholds from the same study, which require more complex modelling of correlated data per study.[27–30]

CONCLUSIONS

In this paper, we have focused on the data analysis in a DTA systematic review. For a general overview of the DTA review process, see Leeflang

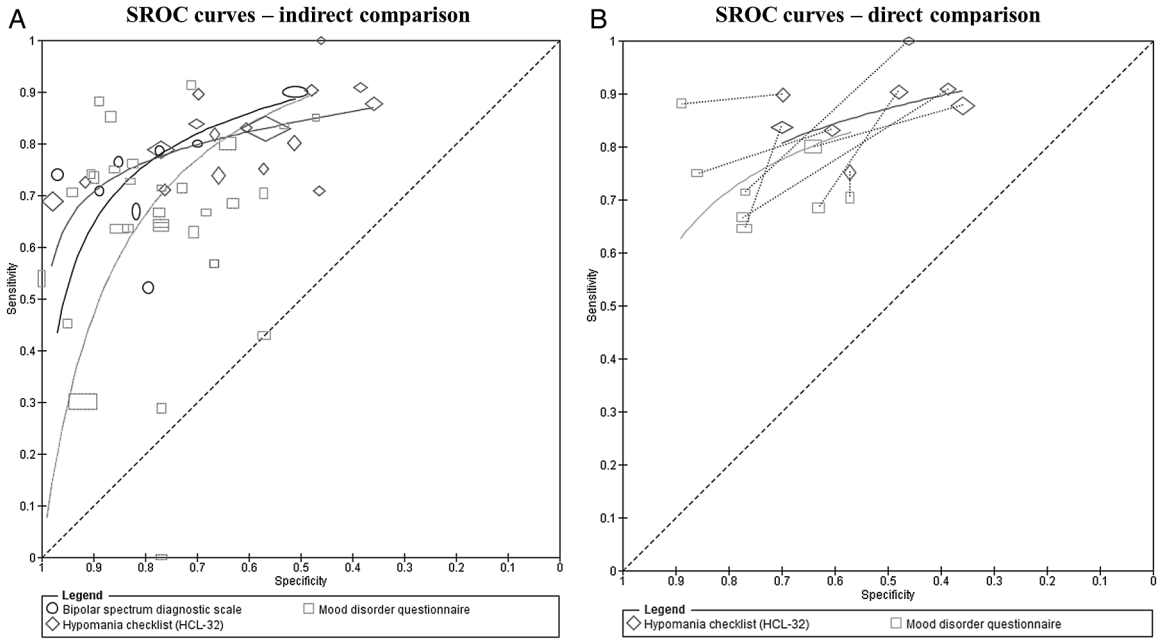

Figure 5 SROC plot comparing the accuracy of BSDS, HCL-32 and MDQ for detection of any type of bipolar disorder in mental health centre settings. For each test on an SROC plot, each symbol represents the pair of sensitivity and specificity from a study. The study points were scaled according to the precision of sensitivity and specificity in the studies. Each SROC curve was drawn restricted to the range of specificities from included studies that evaluated the test. The SROC plot in panel (A) is an indirect comparison (includes all studies that evaluated any of the tests) while panel (B) is a direct comparison where analysis were restricted to only studies that compared both tests in the same patients. A line connects the pair of points representing the two tests from each study (adapted from Carvalho *et al*[22]; BSDS, bipolar spectrum diagnostic scale; HCL-32, hypomanic checklist; MDQ, mood disorder questionnaire; SROC, Summary receiver operating characteristic). Reprinted from ref 35 with permission from Elsevier.

et al.[8 9] Meta-analyses of diagnostic accuracy studies can provide answers to important clinical questions but the methods recommended are demanding. We hope our paper will provide clinicians with sufficient understanding of the terminology and methods to aid interpretation of systematic reviews and facilitate better patient care. We advise review teams to seek the support of a statistician if statistical expertise is lacking in the team. Hierarchical models can be problematic to fit, especially when there are sparse data. The models may also need to be simplified when there are few studies and meta-analysis is judged to be appropriate.[31] The Cochrane Handbook for DTA Reviews[11] provides guidance and is available at http://dta.cochrane.org/. The website also contains links to training materials and tutorials that can aid review authors in producing methodologically rigorous reviews. For more examples of DTA reviews on topics related to mental health, we refer readers to the Cochrane Library.

Acknowledgements The authors would like to thank Dr Sam Creavin (Centre for Academic Primary Care, University of Bristol, UK); Dr André Carvalho (Translational Psychiatry Research Group, Federal University of Ceara, Brazil); Dr Terry Quinn (Cardiovascular and Medical Sciences, Glasgow Royal Infirmary, UK) and Dr Dallas Seitz (Department of Psychiatry, Queen's University Providence Care—Mental Health Services, Canada) for comments and suggestions on improving the manuscript.

Funding YT was supported by the UK National Institute for Health Research [DRF-2011-04-135]. RDR and JJD were supported by funding from a multivariate meta-analysis grant from the MRC Methodology Research Programme (grant reference number: MR/J013595/1).

Competing interests None declared.

OPEN ACCESS

doi:10.1136/eb-2015-102228

Accepted 15 September 2015

REFERENCES

1. **Naaktgeboren CA**, Bertens LC, van Smeden M, *et al*. Value of composite reference standards in diagnostic research. *BMJ* 2013;**347**:f5605.
2. **Bachmann LM**, Puhan MA, ter Riet G, *et al*. Sample sizes of studies on diagnostic accuracy: literature survey. *BMJ* 2006;**332**:1127–9.
3. **Dahabreh IJ**, Chung M, Kitsios GD, *et al*. *Comprehensive overview of methods and reporting of meta-analyses of test accuracy*. Rockville, MD: Agency for Healthcare Research and Quality, 2012.
4. **Steingart KR**, Sohn H, Schiller I, *et al*. Xpert® MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. *Cochrane Database Syst Rev* 2013;**1**:CD009593.
5. *Automated real-time nucleic acid amplification technology for rapid and simultaneous detection of tuberculosis and rifampicin resistance: Xpert MTB/RIF assay for the diagnosis of pulmonary and extrapulmonary TB in adults and children. Policy update*. Geneva: World Health Organization, 2013.
6. **Soares-Weiser K**, Takwoingi Y, Panesar SS, *et al*. The diagnosis of food allergy: a systematic review and meta-analysis. *Allergy* 2014;**69**:76–86.
7. **Muraro A**, Werfel T, Hoffmann-Sommergruber K, *et al*. EAACI food allergy and anaphylaxis guidelines: diagnosis and management of food allergy. *Allergy* 2014;**69**:1008–25.
8. **Leeflang MM**, Deeks JJ, Gatsonis C, *et al.*, Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. *Ann Intern Med* 2008;**149**:889–97.
9. **Leeflang MM**, Deeks JJ, Takwoingi Y, *et al*. Cochrane diagnostic test accuracy reviews. *Syst Rev* 2013;**2**:82.
10. **Quinn TJ**, Fearon P, Noel-Storr AH, *et al*. Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) for the diagnosis of dementia within

community dwelling populations. *Cochrane Database Syst Rev* 2014;**4**: CD010079.
11. **Macaskill P**, Gatsonis C, Deeks JJ, *et al*. Chapter 10: analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration, 2010. http://srdta.cochrane.org/, Version 1.0.
12. **Irwig L**, Macaskill P, Glasziou P, *et al*. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol* 1995;**48**:119–30; discussion 131–112.
13. **Moses LE**, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med* 1993;**12**:1293–316.
14. **Arends LR**, Hamza TH, van Houwelingen JC, *et al*. Bivariate random effects meta-analysis of ROC curves. *Med Decis Making* 2008;**28**:621–38.
15. **Ma X**, Nie L, Cole SR, *et al*. Statistical methods for multivariate meta-analysis of diagnostic tests: an overview and tutorial. *Stat Methods Med Res* Published Online First: 26 Jun 2014. doi:10.1177/0962280213492588
16. **Reitsma JB**, Glas AS, Rutjes AW, *et al*. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005;**58**:982–90.
17. **Chu H**, Guo H, Zhou Y. Bivariate random effects meta-analysis of diagnostic studies using generalized linear mixed models. *Med Decis Making* 2010;**30**:499–508.
18. **Rutter CM**, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001;**20**:2865–84.
19. **Harbord RM**, Deeks JJ, Egger M, *et al*. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007;**8**:239–51.
20. **Zwinderman AH**, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. *Stat Med* 2008;**27**:687–97.
21. **Leeflang MM**, Deeks JJ, Rutjes AW, *et al*. Bivariate meta-analysis of predictive values of diagnostic tests can be an alternative to bivariate meta-analysis of sensitivity and specificity. *J Clin Epidemiol* 2012;**65**:1088–97.
22. **Carvalho AF**, Takwoingi Y, Sales PM, *et al*. Screening for bipolar spectrum disorders: a comprehensive meta-analysis of accuracy studies. *J Affect Disord* 2014;**172C**:337–46.
23. **Higgins JP**, Thompson SG, Deeks JJ, *et al*. Measuring inconsistency in meta-analyses. *BMJ* 2003;**327**:557–60.
24. **Takwoingi Y**, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Ann Intern Med* 2013;**158**:544–54.
25. **Abba K**, Kirkham AJ, Olliaro PL, *et al*. Rapid diagnostic tests for diagnosing uncomplicated non-falciparum or Plasmodium vivax malaria in endemic countries. *Cochrane Database Syst Rev* 2014;**12**:CD011431.
26. **Takwoingi Y**, Deeks J. MetaDAS: a SAS macro for meta-analysis of diagnostic accuracy studies. User Guide Version 1.3. 2010. http://dta.cochrane.org/sites/dta.cochrane.org/files/uploads/MetaDAS Readme v1.3 May 2012.pdf (accessed 11 Aug 2015).
27. **Riley RD**, Ahmed I, Ensor J, *et al*. Meta-analysis of test accuracy studies: an exploratory method for investigating the impact of missing thresholds. *Syst Rev* 2015;**4**:12.
28. **Hamza T**, Arends L, van Houwelingen H, *et al*. Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. *BMC Med Res Methodol* 2009;**9**:73.
29. **Dukic V**, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. *Biometrics* 2003;**59**:936–46.
30. **Putter H**, Fiocco M, Stijnen T. Meta-analysis of diagnostic test accuracy studies with multiple thresholds using survival methods. *Biom J* 2010;**52**:95–110.
31. **Takwoingi Y**, Guo B, Riley RD, *et al*. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. *Stat Methods Med Res* Published Online First: 26 Jun 2015, doi:10.1177/0962280215592269
32. **Hirschfeld RM**, Williams JB, Spitzer RL, *et al*. Development and validation of a screening instrument for bipolar spectrum disorder: the Mood Disorder Questionnaire. *Am J Psychiatry* 2000;**157**:1873–5.
33. **Ghaemi SN**, Miller CJ, Berv DA, *et al*. Sensitivity and specificity of a new bipolar spectrum diagnostic scale. *J Affect Disord* 2005;**84**:273–7.
34. **Angst J**, Adolfsson R, Benazzi F, *et al*. The HCL-32: towards a self-assessment tool for hypomanic symptoms in outpatients. *J Affect Disord* 2005;**88**:217–33.
35. **Carvalho AF**, Takwoingi Y, Sales PM, *et al*. Screening for bipolar spectrum disorders: A comprehensive meta-analysis of accuracy studies. *J Affect Disord* 2014;**172C**:337–46.

Meta-analysis of diagnostic accuracy studies in mental health

Yemisi Takwoingi, Richard D Riley and Jonathan J Deeks

Evid Based Mental Health published online October 7, 2015