Number XX

An Empirical Assessment of Bivariate Methods for Meta-analysis of Test Accuracy

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

This information does not represent and should not be construed to represent an Agency for Healthcare Research and Quality or Department of Health and Human Services determination or policy.

Contract No. [redacted]

Prepared by:
[redacted]

Investigators
[redacted]

Task Order Officer
Elisabeth Kato, M.D., M.R.P.

This document is in the public domain and may be used and reprinted without permission except those copyrighted materials noted, for which further reproduction is prohibited without the specific permission of copyright holders.

Suggested Citation: [to be provided by AHRQ]. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm.

None of the investigators has any affiliations or financial involvement that conflicts with the material presented in this report. Several of the authors (<initials to be provided after peer review>) are involved in developing open source software for meta-analysis, but this software is available at no charge and the authors receive no financial benefit.

AHRQ Contact
Elisabeth Kato, M.D., M.R.P.
Task Order Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850

This report is based on research conducted by the [redacted] Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. [redacted]). The findings and conclusions in this document are those of the author(s), who are responsible for its content, and do not necessarily represent the views of AHRQ. No statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help clinicians, employers, policymakers, and others make informed decisions about the provision of health care services. This report is intended as a reference and not as a substitute for clinical judgment.

This report may be used, in whole or in part, as the basis for the development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies and strategies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base in and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by e-mail to epc@ahrq.hhs.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Stephanie Chang M.D., M.P.H.
Director, EPC Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Elisabeth Kato, M.D., M.R.P.
Task Order Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Contents

Abstract vii
Introduction 8
Methods 10
Construction of a database of meta-analysis datasets 10
Statistical analyses 10
Comparisons between alternative methods 13
Software 13
Results 13
Included studies 13
Meta-analysis of sensitivity and specificity 14
Meta-analysis in the receiver operating characteristic space 43
Discussion 49
Key findings 49
Meta-analysis of sensitivity and specificity 50
Constructing meta-analytic ROC curves 55
Limitations 55
Conclusions 56
Abbreviations 58
References 59

Tables

Table 1: Methods for meta-analysis of sensitivity and specificity used in this report 12
Table 2: Descriptive characteristics of test accuracy meta-analyses 14
Table 3: Estimated correlation by different bivariate methods 32
Table 4: Slope of ROC line in the logit-space 48
Table 5: Summary of selected previous empirical comparisons of meta-analysis methods, including simulation studies 52
Box 1: Summary of findings relevant to meta-analytic practice 57

Figures

Figure 1: Comparison of point estimates and standard errors of summary logit-transformed sensitivity and specificity (random versus fixed effects) 15
Figure 2: Comparison of point estimates and standard errors of summary sensitivity and specificity from random effects meta-analyses using a normal approximation (estimation of heterogeneity with DL versus REML) 17
Figure 3: Absolute differences of estimated sensitivity and specificity from univariate random effects meta-analyses (DerSimonian-Laird versus REML) 18
Figure 4: Comparison of point estimates and confidence interval widths of summary sensitivity and specificity from univariate random effects meta-analyses using the exact binomial likelihood versus using a normal approximation (both models fit using REML) 19

Figure 5: Absolute differences of estimated sensitivity and specificity (univariate random effects meta-analyses using the exact binomial likelihood versus a normal approximation; both models fit with REML) 20
Figure 6: Summary comparison of logit-transformed sensitivity and specificity estimates from all univariate methods considered in this report 21
Figure 7: Comparison of point estimates and confidence interval widths of summary sensitivity and specificity from bivariate random effects methods using a normal approximation (generalized DerSimonian-Laird inverse variance vs. REML) 23
Figure 8: Absolute differences of estimated sensitivity and specificity from bivariate random effects meta-analyses with generalized DerSimonian-Laird inverse variance vs. REML (both models using a normal approximation to represent within-study variability) 24
Figure 9: Comparison of point estimates and confidence interval widths of summary sensitivity and specificity from bivariate random effects meta-analyses (approximate normal versus exact binomial; both models fit with REML) 25
Figure 10: Absolute differences of estimated sensitivity and specificity from bivariate random effects meta-analyses (approximate normal versus exact binomial; both models fit with REML) 26
Figure 11: Comparison of point estimates and confidence/credibility interval widths of summary sensitivity and specificity from bivariate random effects meta-analyses (Bayesian versus maximum likelihood; both models using the exact binomial likelihood to represent within-study variability) 27
Figure 12: Absolute differences of estimated sensitivity and specificity from bivariate random effects meta-analyses fit using fully Bayesian versus REML estimation (using the exact binomial likelihood to represent within-study variability) 28
Figure 13: Summary comparison of logit-transformed sensitivity and specificity estimates from all univariate methods considered in this report 29
Figure 14: Histograms of the estimated correlation between sensitivity and specificity from the three bivariate methods compared in this report. 31
Figure 15: Matrix scatter plot of correlation estimates from the 4 bivariate methods considered in this report 32
Figure 16: Absolute differences of correlation estimates from bivariate random effects meta-analyses (generalized DerSimonian-Laird model using a normal approximation; REML using a normal approximation; REML using the exact binomial likelihood; and fully Bayesian model using the exact binomial likelihood) 33
Figure 17: Point estimates and 95% confidence intervals for the correlation of sensitivity and specificity, as estimated by the bivariate model using the exact binomial likelihood (REML estimation) 34
Figure 18: Comparison of point estimates and confidence interval widths of summary logit-transformed sensitivity and specificity (univariate random effects versus bivariate random effects inverse variance methods, both using a normal approximation for within-study variability) 36
Figure 19: Absolute differences of estimated sensitivity and specificity in univariate versus bivariate random effects inverse variance meta-analyses (both using a normal approximation for within-study variability) 37
Figure 20: Comparison of point estimates and standard errors of summary logit-transformed sensitivity and specificity from univariate versus bivariate random effects meta-analyses with REML estimation (using a normal approximation for within-study variability) 38

Figure 21: Absolute differences of estimated sensitivity and specificity from univariate and bivariate random effects meta-analyses with REML estimation (using a normal approximation to represent within-study variability for both models) 39
Figure 22: Comparison of point estimates and confidence interval widths of summary logit-transformed sensitivity and specificity from univariate versus bivariate random effects meta-analyses using the exact binomial likelihood (both models fit using REML) 40
Figure 23: Absolute differences of estimated sensitivity and specificity comparing univariate versus bivariate random effects meta-analyses fit with REML (both models using the exact binomial likelihood to represent within-study variability) 41
Figure 24: Summary comparison of logit-transformed sensitivity and specificity estimates from all methods considered in this report 42
Figure 25: Summary comparison of logit-transformed sensitivity and specificity estimates from all methods considered in this report 43
Figure 26: Scatter plot of the slopes of alternative sROC lines (logit space) 44
Figure 27: sROC curves for 24 randomly selected meta-analyses (bivariate random effects versus Moses-Littenberg methods) 45
Figure 28: HsROC curves for 24 randomly selected meta-analyses (alternative parameterizations of the HsROC curve) 47
Figure 29: Datapoints and fitted hsROC curves for an example dataset 48
Figure 30: Scatter plot of the slopes of alternative HsROC lines (logit space) 49

Appendices

Appendix A. List of included studies
Appendix B Bayesian bivariate model
Appendix C Alternative parameterizations to obtain the HsROC curve

Appendices cited in this report are available at http://ahrq.gov/XXXXX
[to be completed by AHRQ]

Abstract

Background: Multiple methods are available for meta-analysis of sensitivity and specificity pairs reported from test accuracy studies. Regression methods can be used to construct summary (meta-analytic) receiver operating characteristic (sROC) curves. Little is known about the impact of alternative methods on applied meta-analysis results.

Methods: We constructed a database of PubMed-indexed meta-analyses (1984-2003) from which 2×2 tables for each included primary study could be readily extracted. We evaluated the following methods for meta-analysis of sensitivity and specificity: fixed and random effects univariate meta-analyses using inverse variance methods; univariate random effects meta-analyses with restricted maximum likelihood (REML, both using a normal approximation and the exact binomial likelihood to describe between-study variability); bivariate random effects meta-analyses (both using a normal approximation and the exact binomial likelihood to describe between-study variability). The bivariate model using the exact binomial likelihood was also fit using a fully Bayesian approach. We constructed sROC curves using the Moses-Littenberg fixed effects method (weighted and unweighted) and the Rutter-Gatsonis hierarchical sROC (HsROC) method. We also obtained alternative HsROC curves corresponding to different underlying regression models [logit-true positive rate (TPR) over logit-false positive rate (FPR); logit-FPR over logit-TPR; difference of the logit-TPR and logit-TPR over their sum; and major axis regression of logit-TPR over logit-FPR].

Results: We re-analyzed 308 meta-analyses of test performance. Fixed effects univariate analyses produced estimates with narrower confidence intervals compared to random effects methods. Methods using the normal approximation (both univariate and bivariate, inverse variance and REML) produced estimates of summary sensitivity and specificity closer to 0.5 and smaller standard errors compared to methods using the exact binomial likelihood. Point estimates from univariate and bivariate random effects meta-analyses were similar when performing pairwise (univariate vs. bivariate) comparisons, regardless of the estimation method (inverse variance, REML with normal approximation, or REML with the exact binomial likelihood for estimation). Fitting the bivariate model using REML and fully Bayesian methods produced almost identical point estimates of summary sensitivity and specificity; however, Bayesian results indicated additional uncertainty around summary estimates. The correlation of sensitivity and specificity across studies was poorly estimated by all bivariate methods. The sROC curves produced by the Moses-Littenberg and Rutter-Gatsonis models were similar in most examples. Alternative parameterizations of the HsROC regression resulted in markedly different summary lines in a third of the meta-analyses; this depends to a large extent on the estimated covariance between sensitivity and specificity in the bivariate model.

Conclusion: Meta-analysis of sensitivity and specificity using a normal approximation produced estimates that were lower compared to meta-analysis using the exact binomial likelihood. Bivariate models are more theoretically motivated compared to univariate analyses and allow estimation of the correlation between sensitivity and specificity. Bayesian methods fully quantify uncertainty and their ability to incorporate external evidence (in the form of a prior) may be particularly useful for parameters that are poorly estimated in the bivariate model (such as the correlation between sensitivity and specificity). Alternative sROC curves provide useful global summaries of test performance; they can differ between them because each is based on different assumptions.

Introduction

Medical tests are used every day for guiding diagnosis, predicting the future course of disease, and guiding treatment selection. The effects of tests on clinical outcomes are indirect, through their influence on physicians' diagnostic thinking and treatment decisionmaking (1). Comparative studies of testing versus no testing that can answer the overarching question of test effectiveness (clinical utility) are rarely performed. Because of this, assessment of medical tests often relies only on the evaluation of test "accuracy",[a] or better test performance, typically measured by sensitivity and specificity (clinical validity of tests). Even when studies of clinical utility are available, systematic reviews of test performance are an important component of any comprehensive evidence assessment of a medical test (2,3).

In most cases tests are used to classify patients into two mutually exclusive and exhaustive groups ("test positive" and "test negative") – positive test results indicate that patients are more likely to have the condition of interest and should be targeted for additional diagnostic investigation or considered for therapeutic intervention.[b] In such cases, test accuracy can be expressed as the ability to identify as "test positives" individuals with disease (sensitivity) and as "test negatives" individuals with no disease (specificity). Additional accuracy metrics, such as the area under the receiver operating characteristic (ROC) curve, the diagnostic odds ratio (4), or the Q^* statistic (the point on the ROC curve where sensitivity equals specificity), are often reported in primary studies (5).

Individual studies of test accuracy tend to be small and are often conducted in diverse settings. Systematic reviews of medical test studies offer a natural framework for evidence synthesis. When the aim is to increase precision or quantitatively assess the impact of study-level characteristics on test sensitivity or specificity, meta-analytic methods can be used to combine the results of independent studies into summary estimates of accuracy or to identify modifiers of accuracy through meta-regression (6,7).

Meta-analysis of studies of test accuracy presents several challenges to systematic reviewers. First, meta-analysis of sensitivity and specificity requires modeling a multivariate outcome (sensitivity and specificity reported from each study) (5). Second, joint modeling of sensitivity and specificity needs to take into account the correlation of these estimates across studies induced by threshold effects (8,9). Third, studies often produce heterogeneous results, necessitating the use of random effects models when the interest is to generalize beyond the observed data (10,11). Analyses that fail to take into account threshold effects or between-study variability may produce incompatible estimates of sensitivity and specificity or spuriously precise estimates of test accuracy.

In a previous empirical investigation [citation to be provided post peer review], we found that the most common test performance metrics used in meta-analysis were sensitivity and specificity; in the majority of reviews only results from univariate

---

[a] Here we use the term "test accuracy" to denote "test performance". We do not refer to the metric "accuracy", which is the proportion of correct test classifications (true positives and true negatives) out of the total sample size in a study.

[b] Although some tests produce ordinal classifications (e.g., high-intermediate-low probability of disease) or are used as components of more complex testing algorithms, the vast majority of systematic reviews and meta-analyses focus on binary classification.

analyses were reported. Additionally, many meta-analyses used the summary receiver characteristic operating (sROC) curve method proposed by Moses and Littenberg (8,9) to assess test performance. This method is based on a regression of the difference of the logit-transformed sensitivity and specificity (i.e., the diagnostic log-odds ratio) on their sum. Although this method accounts for between-study correlation of sensitivity and specificity, it is almost always implemented in a fixed effect framework. Furthermore, it ignores the exact binomial distributions of the test results among diseased and non-diseased individuals or the uncertainty in the measurement of the independent variable (the sum of the logit-sensitivity and logit-specificity) in the regression model. Recently, several authors have advocated using bivariate models based on hierarchical regression (10,12-14). These methods can be used to estimate summary sensitivity and specificity (i.e., a summary point on the ROC plane) or to fit a line describing the bivariate distribution of sensitivity and specificity (i.e., a hierarchical sROC curve). These meta-analysis models have now been implemented in major statistical packages (15,16) and are becoming increasingly popular in meta-analytic practice (17). These implementations are based on maximization of likelihoods, but Bayesian methods have also been proposed (12,18,19). Limited empirical work suggests that these approaches yield similar conclusions in applied meta-analysis examples (16). Although theoretical arguments provide support for the use of bivariate random effects methods for the typical case of binary tests[c], the existing evidence on the practical implications of alternative methods is limited to small empirical comparisons (typically based on a few meta-analysis examples) (10,20,21). Furthermore, divergent conclusions have been drawn from considering theory and available evidence: for example, one group of methodologists suggested that "hierarchical models are necessary" (21) whereas another that "differences between univariate and bivariate models […] may not be large" (22).

This report is the second in a series of three on meta-analysis of test accuracy, conducted by the [redacted] Evidence-based Practice Center under contract with the Agency for Healthcare Research and Quality (AHRQ). For the current project we sought to perform a large-scale empirical comparison of alternative meta-analysis methods for sensitivity and specificity and for constructing sROC curves.[d] This report aims to address the following aims by using a previously established database of meta-analytic datasets:

- To compare univariate (one outcome at a time) and bivariate (joint analysis of two outcomes) methods for meta-analysis of sensitivity and specificity.
- To compare inverse variance (DerSimonian-Laird), restricted maximum likelihood (REML) and Bayesian methods for random effects meta-analysis of sensitivity and specificity.
- To compare methods using a normal approximation versus those using the exact binomial likelihood for meta-analysis of sensitivity and specificity.
- To compare alternative statistical models for constructing meta-analytic sROC curves.

[c] Similar arguments can be made in support of the use of extensions of these methods to account for multiple thresholds or multiple index tests; however, the majority of published meta-analyses are limited to the binary classification case. This report exclusively focuses on this most common case.

[d] Other reports in this series include a comprehensive survey of methods and reporting in meta-analyses of test accuracy and the development of novel methods for the analysis of diagnostic test networks.

Methods

Construction of a database of meta-analysis datasets

We used a previously described [citation to be provided post peer review] database of PubMed-indexed English-language meta-analyses of test accuracy (published between 1984 and 2003) to identify those that reported adequate information to reconstruct the 2×2 tables of included primary studies. This was done to increase the efficiency of the data extraction process by avoiding having to review all the primary papers included in each meta-analysis. The details of searches, abstract and full text screening methods, and selection criteria used to generate the original database are presented in two previous AHRQ reports produced by the [redacted] EPC. A list of included studies in provided in **Appendix A**.

From each meta-analysis we extracted the first author and year of publication, the number of included studies, and the number of index and reference standard tests reviewed. For each diagnostic outcome included in meta-analysis and each included study, we extracted the 2×2 table of true positive, false positive, true negative and false negative results, as defined by the original meta-analysis. We used these data to calculate descriptive statistics for the database, including descriptive statistics for the number of included studies in each meta-analysis, the number of patients in each component study, test sensitivity and specificity, and prevalence of the diagnostic outcomes of interest. We then repeated the meta-analyses using alternative statistical methods, as described below.

Statistical analyses

Meta-analyses of sensitivity and specificity aim to provide helpful summaries of the findings of individual studies. Sometimes, a helpful way to summarize individual studies is to provide a "summary point", a summary sensitivity and a summary specificity. For example, a summary point is helpful when the results of the studies are relatively similar, and when the studied tests do not have different explicit thresholds for positive results. Other times, it is more helpful to synthesize data using a "summary line" that describes how the average sensitivity changes with the average specificity. For example, a summary line may be a more helpful way to synthesize data when studies have different explicit thresholds and their results range widely. Of course, choosing the most helpful summary is subjective, and both summaries can be reasonably employed as they provide complementary information. Here we do not make any effort to choose the most helpful summary. We analyze all examples with all methods.

Meta-analysis of sensitivity and specificity (summary point)

Univariate meta-analysis methods

We performed univariate (one outcome at a time) meta-analyses of sensitivity and specificity using standard fixed and random effects inverse variance models to synthesize logit-transformed sensitivity and specificity values. For random effects inverse variance analyses, between-study heterogeneity was estimated using a non-iterative method (the DerSimonian-Laird method of moments (23)) and with an iterative method (REML). We

used a continuity correction of 0.5 when the observed count was zero for any of the cells of the 2×2 table, and only for the studies with a zero cell. We also performed univariate random effects analyses with the exact binomial likelihood using random effects logistic regression (a random intercept model). This model was also fit using REML.

Bivariate meta-analysis methods

We fit bivariate random effects meta-analysis models that allow for correlation of sensitivity and specificity at the between-study level. These are hierarchical models that describe the observed variability using statistical distributions of data at two levels: a within-study-level and a between-study level. We fit two variants of the bivariate model. The first used a normal approximation for the statistical distribution of 2×2 cell counts at the within-study level (for the logit transformed sensitivities and specificities) and the second used the binomial distribution (exact likelihood).

At the between-study level, both variants assume that the (true) logit-transformed sensitivities and specificities follow a bivariate normal distribution centered at their summary estimates and with a covariance matrix that represents the between-studies component of the variability of the data. For the normal approximation variant of the model, we estimated the covariance matrix (equivalently, the correlation) using the non-iterative the estimator proposed by Jackson (a multivariate generalization of the DerSimonian-Laird model) (24), as well as iteratively by REML (14). For the analysis using the exact binomial likelihood we used random effects logistic regression fit with REML (10,20).

The bivariate model with the exact binomial likelihood for within-study variability was also fit using a fully Bayesian approach, where all parameters of interest were treated as random variables. This required the specification of prior distributions for all parameters; we used non-informative priors for all analyses to encode our uncertainty about model parameters. The model structure used for our analyses is presented in **Appendix B**. For each meta-analysis we ran 10,000 iterations and then assessed convergence by inspecting trace plots and calculating the Brooks-Gelman-Rubin statistic. We considered nodes to have converged when the Brooks-Gelman-Rubin statistic was less than 1.10. Additional iterations were run until convergence was achieved (if convergence was not reached after 10,000). To sample from the posterior distribution of parameters of interest we ran the model for an additional number of iterations equal to half the number that had been required to achieve convergence. From these posterior distributions we obtained the median and 95% credibility intervals for parameters of interest. We assessed robustness to alternative prior distributions for the variance components and the between-study correlation of sensitivity and specificity (see **Appendix B**).

Table 1 summarizes the methods for meta-analysis of sensitivity and specificity assessed in this report.

Table 1: Methods for meta-analysis of sensitivity and specificity used in this report

Modeling of within-study variability	Meta-analysis method	Univariate meta-analysis (estimation of heterogeneity)	Bivariate meta-analysis (estimation of heterogeneity)
Approximate (normal for logit-transformed sensitivity, specificity)	Inverse variance	• Fixed effects • Random effects (DL)	• Random effects (multivariate DL)*
Approximate (normal for logit-transformed sensitivity, specificity)	Likelihood maximization	• Random effects (REML)	• Random effects (REML)
Exact (binomial)	Likelihood maximization	• Random effects (REML)	• Random effects (REML)
Exact (binomial)	Bayesian meta-analysis with non-informative priors	Not done	• Random effects (MCMC)

* Jackson generalization of the non-iterative DerSimonian and Laird method.
DL = DerSimonian-Laird; MCMC = Markov chain Monte Carlo; REML = restricted maximum likelihood.

Meta-analytic sROC curves (summary lines)

sROC curves depict graphically the relationship between average sensitivity and average specificity and provide a visual summary of overall test performance. The most commonly used method for generating meta-analytic sROC curves is that proposed by Moses and Littenberg, based on a regression of the difference of the logit-transformed sensitivity and specificity over their sum(8,9). We implemented both unweighted and weighted versions of this method and compared their results to the more theoretically motivated hierarchical regression methods (see below). For the weighted Moses-Littenberg sROC regression we used weights equal to the inverse of the variance of the diagnostic odds ratio of each study and calculated the sROC parameters using variance-weighted least squares regression, in a generalized least squares framework.

Hierarchical sROC (HsROC)

Rutter and Gatsonis proposed HsROC meta-analysis methods to address the limitations of the Moses-Littenberg sROC approach (12). As noted by Arends 2008 (25) several alternative parameterizations of the HsROC curve can be produced from the bivariate meta-analysis model, the Rutter-Gatsonis HsROC model being one of them (12). Different formulations of the HsROC represent alternative ways to describe the bivariate distribution of sensitivity and specificity, and can result in curves of different shape. We used the output of the bivariate meta-analysis model (fit both using maximum likelihood and Bayesian methods) to construct these curves, as described previously (25,26). Briefly, we obtained estimates of the intercept and slope of the ROC line (in the logit space) based on the following approaches: the regression of logit-sensitivity on logit-false positive rate; the regression of logit-false positive rate on logit-sensitivity; the regression of the difference of logit-sensitivity and logit-false positive rate on their sum); and a major axis regression of logit-sensitivity on logit-false positive rate (a regression obtained by minimizing the distance between data points and the fitted line). Additional details about these models are presented in **Appendix C**.

Comparisons between alternative methods

We compared summary sensitivity, specificity and the width of the corresponding confidence (of credibility) intervals[e] with scatter plots and absolute difference graphs.(28) We also graphed the ROC curves produced by the various methods and compared their results visually.

With the exception of fixed effect univariate analyses, all other models used in this report assume that the study-specific parameters (sensitivity or specificity) are random effects (i.e., they differ by study). Generalizing outside the observed studies is a key aim for most meta-analyses, and thus random effects models are appropriate. For this reason, we focus on random effects models in this report.

Software

All non-Bayesian meta-analyses were performed in Stata/IC (version 12; Stata Corp., College Station, TX). Bayesian analyses were implemented in WinBugs (29) (version 1.4.3; MRC Biostatistics Unit, Cambridge, UK), through calls from Stata or R (version 2.13.2; R Foundation for Statistical Computing, Vienna, Austria). Graphs were generated in Stata.

Results

Included studies

We included 157 systematic reviews reporting 308 meta-analyses (published between 1988 and 2003) for which complete data to reconstruct 2×2 tables of the individual studies were available. The meta-analyses contributing data to this empirical comparison represent approximately 59% of all test accuracy meta-analyses published during the study period and identified through our searches. We treated meta-analyses as independent observations even when they had common studies. The small amount of overlap does not introduce appreciable bias because overlap in studies is limited (between meta-analyses originating from the same systematic review) and uncommon (between systematic reviews).

The median number of studies in the included meta-analyses was 11 (25th-75th percentile: 8-18). Included primary studies were generally small (median of the median number of affected individuals across meta-analyses = 30; median of the median number of affected individuals across meta-analyses = 62). Additional characteristics of the included meta-analyses and their component studies are presented in **Table 2**.

[e] For all bivariate analyses we obtained non-simultaneous confidence intervals for estimates of sensitivity and specificity. One could opt to obtain simultaneous confidence intervals (27) instead; however this is not common practice in meta-analyses of test accuracy and has not been implemented in the existing software routines in common use. Generally, simultaneous confidence intervals tend to be wider than non-simultaneous confidence intervals.

Table 2: Descriptive characteristics of test accuracy meta-analyses

Characteristics	Median	25th percentile	75th percentile	Minimum	Maximum
Number of included primary studies	11	8	18	6	61
Median number of affected individuals (cases)*	30	17	46	2	468
Median number of unaffected individuals (controls)*	62	29	108	8	9979
Median ratio of affected (cases) to unaffected (control) individuals	0.46	0.21	1	0.002	15.14
Median number of "positive" test results	32	19	60	3	549
Median number of "negative" test results	59	30	99	7	9415
Median ratio of "positive" to "negative" test results	0.52	0.30	1	0.01	7.17
Median TP count	21	12	34	2	464
Median FP count	7	3	19	0	491
Median FN count	6	2	11	0	157
Median TN count	48	24	87	5	9411
"Crude"† sensitivity	0.78	0.58	0.88	0.15	0.99
"Crude"† specificity	0.86	0.77	0.93	0.18	1

* Based on the reference standard test used in each study.
† Calculated by summing the numerators and denominators of included studies (this is equivalent to a fixed effects meta-analysis using the exact binomial likelihood).

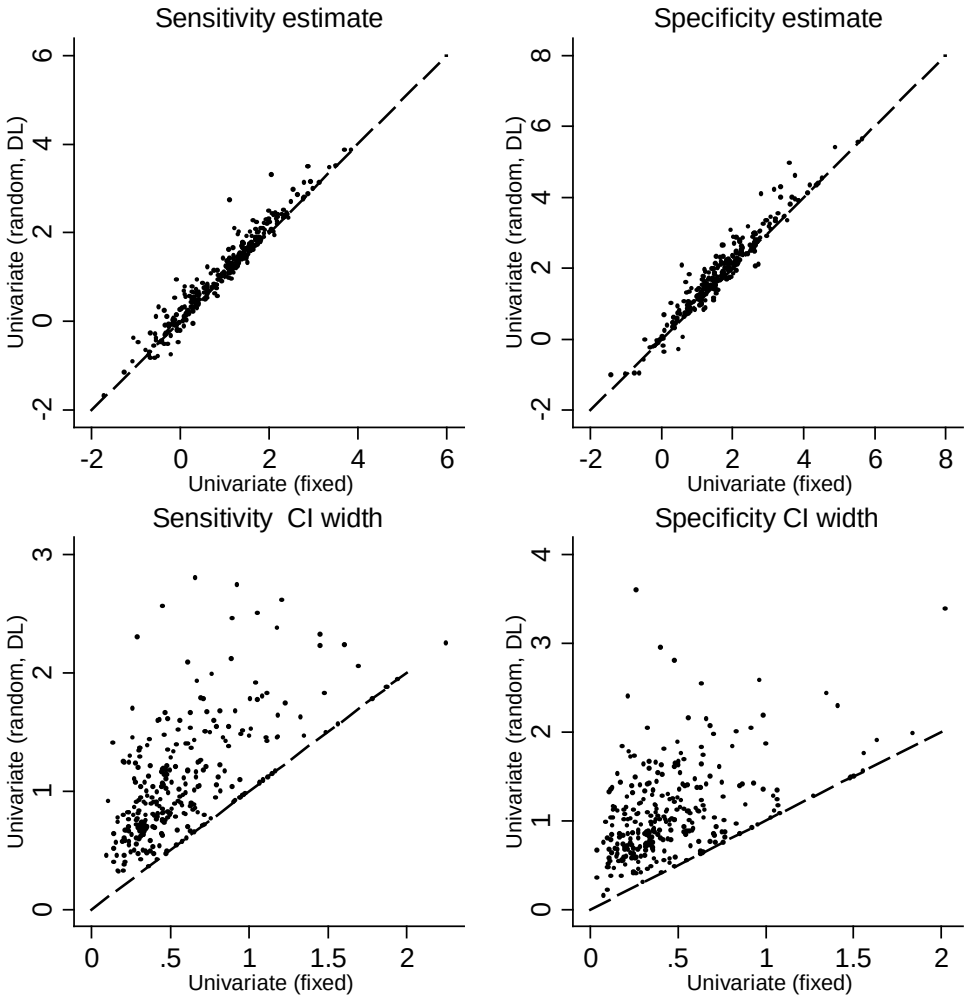
Some investigators have suggested that problems with convergence are a major concern when fitting the bivariate models for meta-analysis of test accuracy (22). In total, for 10 of 308 meta-analyses (3%) we could not obtain estimates from all methods listed in **Table 1** and they have been excluded from the comparisons presented in the following sections.

Meta-analysis of sensitivity and specificity

Fixed versus random effects univariate inverse variance meta-analyses

Figure 1 compares the point estimates (logit-transformed) and confidence interval widths for univariate meta-analyses of sensitivity and specificity using fixed versus random effects inverse variance models. Overall, point estimates from the two methods are similar; however, the estimated uncertainty around each estimate is very different between methods. This is because the random effects model incorporates between-study variability. We argue that meta-analysts are (almost) always interested in generalizable (to unobserved studies) summary estimates; in such cases random effects models are more appropriate, particularly in the presence of between-study heterogeneity (which is common in diagnostic test reviews). All subsequent comparisons in this report are limited to random effects models.

Figure 1: Comparison of point estimates and standard errors of summary logit-transformed sensitivity and specificity (univariate DL random effects versus fixed effect inverse variance)



Scatter plot of estimated logit-transformed sensitivity, specificity and their corresponding confidence interval widths from univariate random and fixed effects meta-analyses of sensitivity and specificity. CI = confidence interval; DL = DerSimonian-Laird.

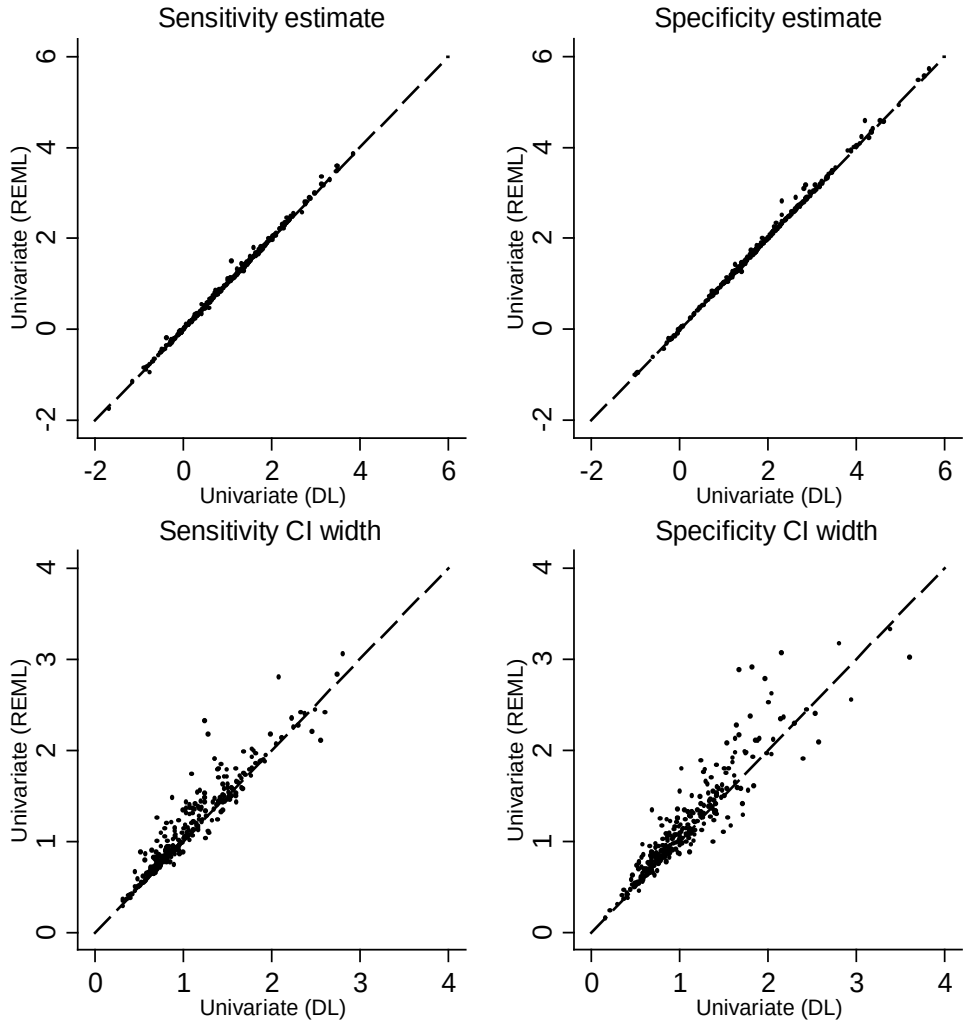
UNIVARIATE RANDOM EFFECTS META-ANALYSIS METHODS

This section compares the results of alternative univariate meta-analysis methods for sensitivity and specificity.

Inverse variance versus REML univariate random effects meta-analyses (normal approximation)

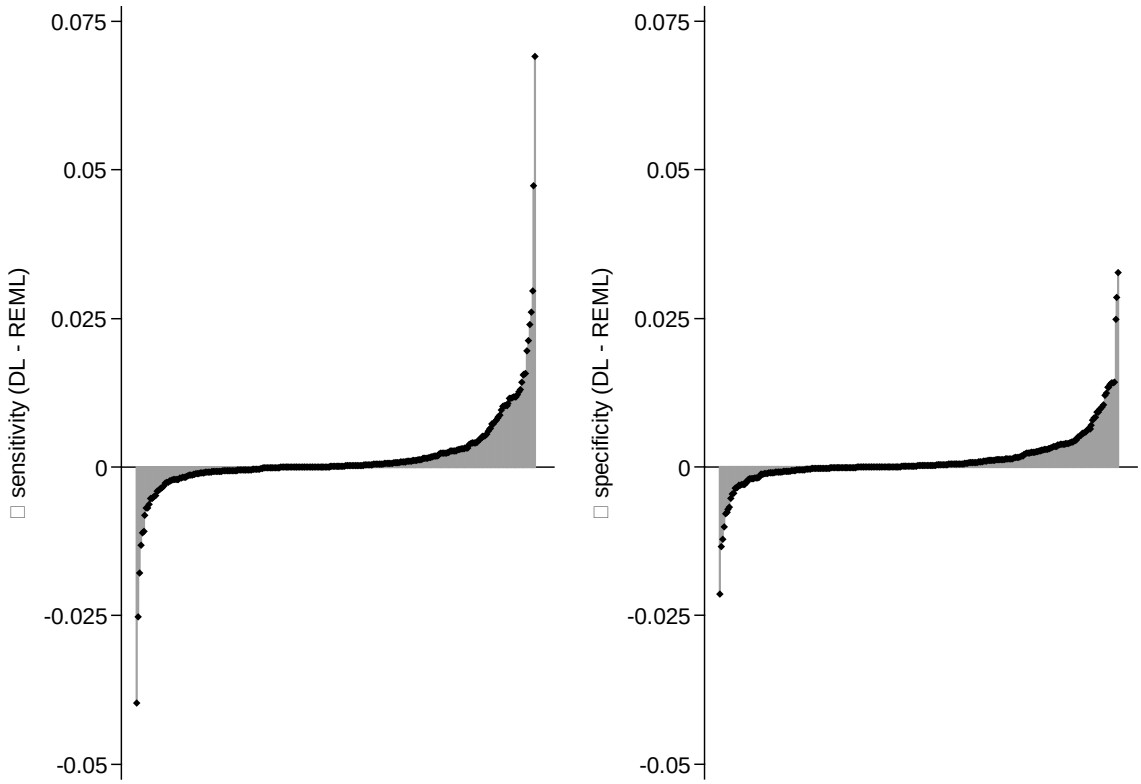
Figure 2 compares the point estimates (logit-transformed) and confidence interval widths for univariate meta-analyses of sensitivity and specificity using inverse variance versus REML random effects models based on the assumption of normal distributions for the logit-transformed probabilities. The point estimates from the two methods are very similar. However, the methods often produce different standard errors, resulting in different confidence interval widths. **Figure 3** presents the absolute difference plots (untransformed scale) between estimated sensitivities and specificities. The absolute differences between the two methods are rarely greater than 2.5%.

Figure 2: Comparison of point estimates and standard errors of summary sensitivity and specificity from random effects meta-analyses using a normal approximation (estimation of heterogeneity with DL versus REML)



Scatter plot of estimated logit-transformed sensitivity, specificity and their corresponding confidence interval widths from univariate random effects meta-analyses using the DerSimonian-Laird inverse variance method versus REML. CI = confidence interval; DL = DerSimonian-Laird; REML = restricted maximum likelihood.

Figure 3: Absolute differences of estimated sensitivity and specificity from univariate random effects meta-analyses (DerSimonian-Laird versus REML)



Difference plot of sensitivity (left panel) and specificity (right panel) estimates comparing univariate random effects meta-analyses using the DerSimonian-Laird inverse variance method versus REML (both using a normal approximation for within-study variability). DL = DerSimonian-Laird; REML = restricted maximum likelihood.

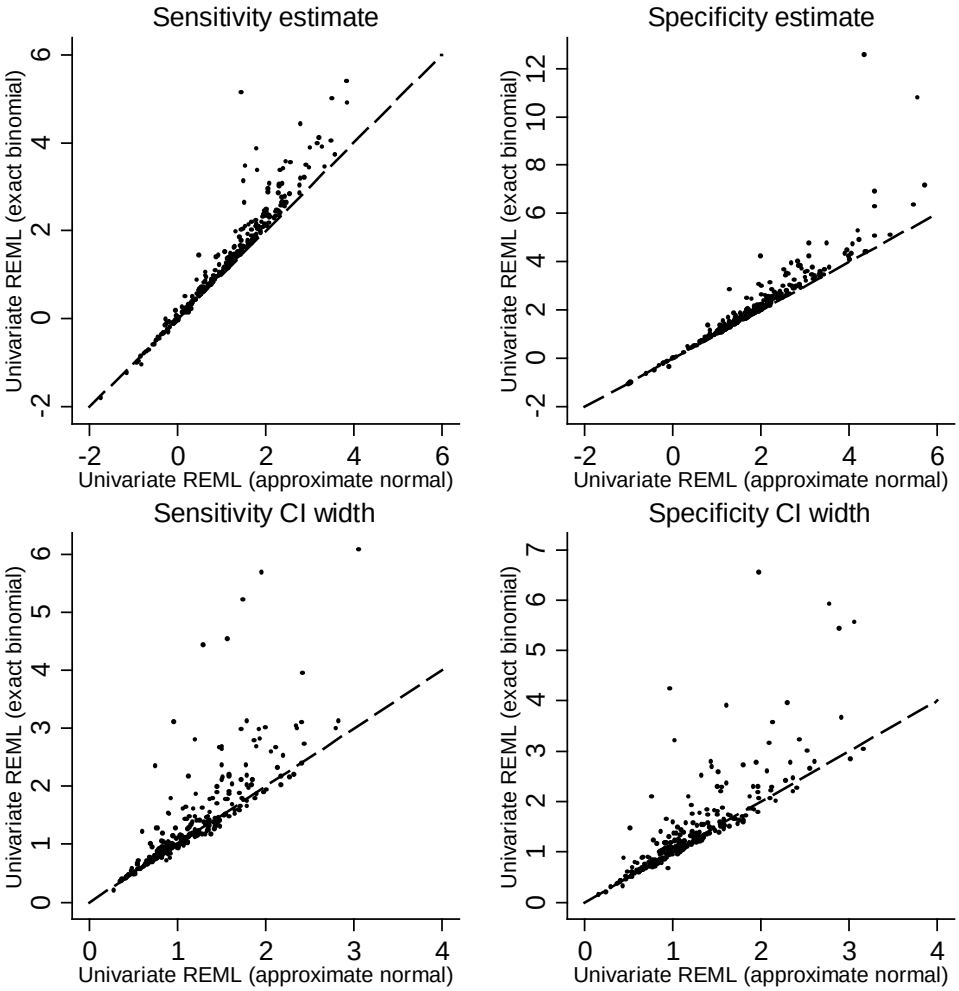
Approximate normal versus exact binomial univariate random-effects meta-analyses (REML)

Figure 4 compares the point estimates (logit-transformed) and confidence interval widths from univariate random effect meta-analyses of sensitivity and specificity using a normal approximation versus the exact binomial likelihood (both models fit with REML). The point estimates from the two models are often dissimilar, and differences are greater for sensitivity and specificity values that are closer to one. This may be explained in part by the need for continuity corrections when the normal approximation is used to model within-study variability, which biases the estimated proportions towards lower values (towards 0.5). An additional reason may be that the estimates and variances are correlated (30): the variance is a function of the estimate and the sample size. This correlation is positive for proportions larger than 0.5, and estimates near one have larger variance (and receive less weight in the meta-analysis) compared to estimates near 0.5. The net effect is

that summary sensitivity or specificity are biased towards 0.5. The aforementioned biases are not a problem for meta-analysis methods using the exact likelihood.

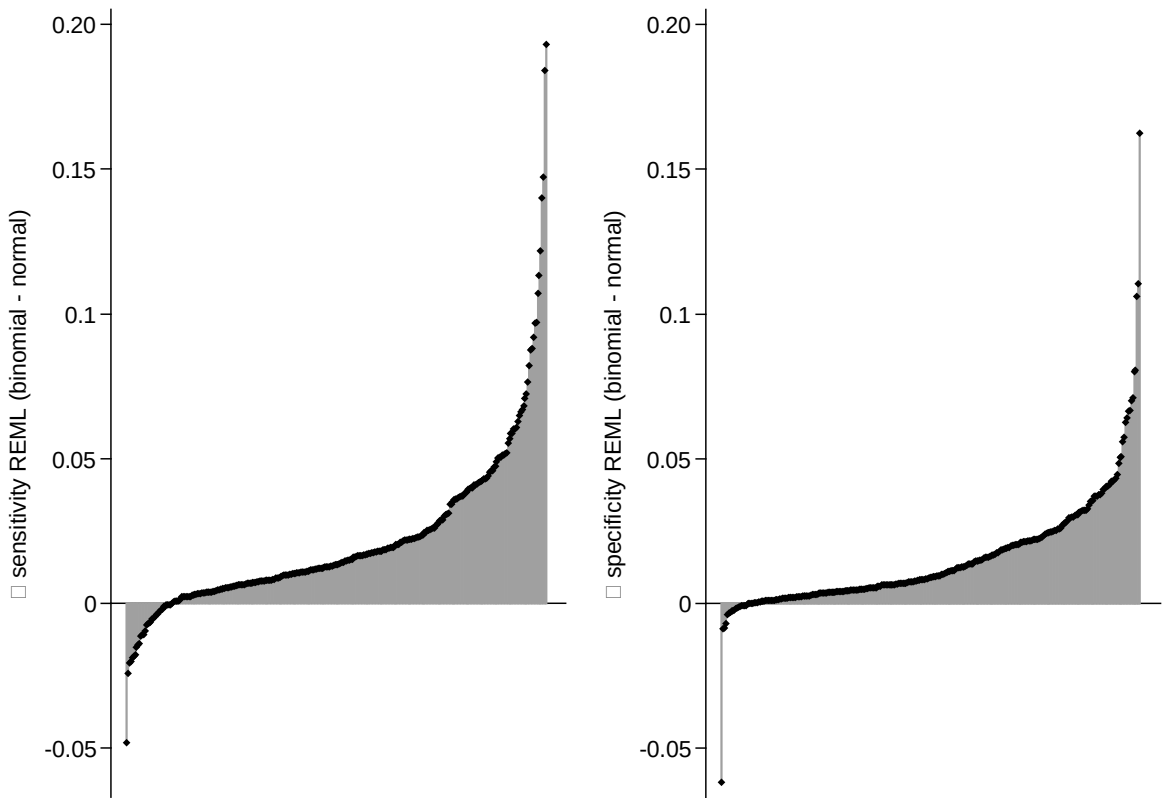
Further, in many cases the exact likelihood results in wider confidence intervals, indicating that the normal approximation leads to an underestimation of the uncertainty surrounding the summary estimates. **Figure 5** presents the absolute difference (untransformed scale) between estimated sensitivities and specificities. It is not uncommon to have differences in summary sensitivity and specificity of 5% or higher. As mentioned, summary sensitivity and specificity obtained with the normal approximation tend to smaller than those obtained from exact methods.

Figure 4: Comparison of point estimates and confidence interval widths of summary sensitivity and specificity from univariate random effects meta-analyses using the exact binomial likelihood versus using a normal approximation (both models fit using REML)



Scatter plot of estimated logit-transformed sensitivity, specificity and their corresponding confidence interval widths from univariate random effects meta-analyses using the exact binomial likelihood versus using an approximate normal likelihood to describe within-study variability. CI = confidence interval width; REML = restricted maximum likelihood.

Figure 5: Absolute differences of estimated sensitivity and specificity (univariate random effects meta-analyses using the exact binomial likelihood versus a normal approximation; both models fit with REML)

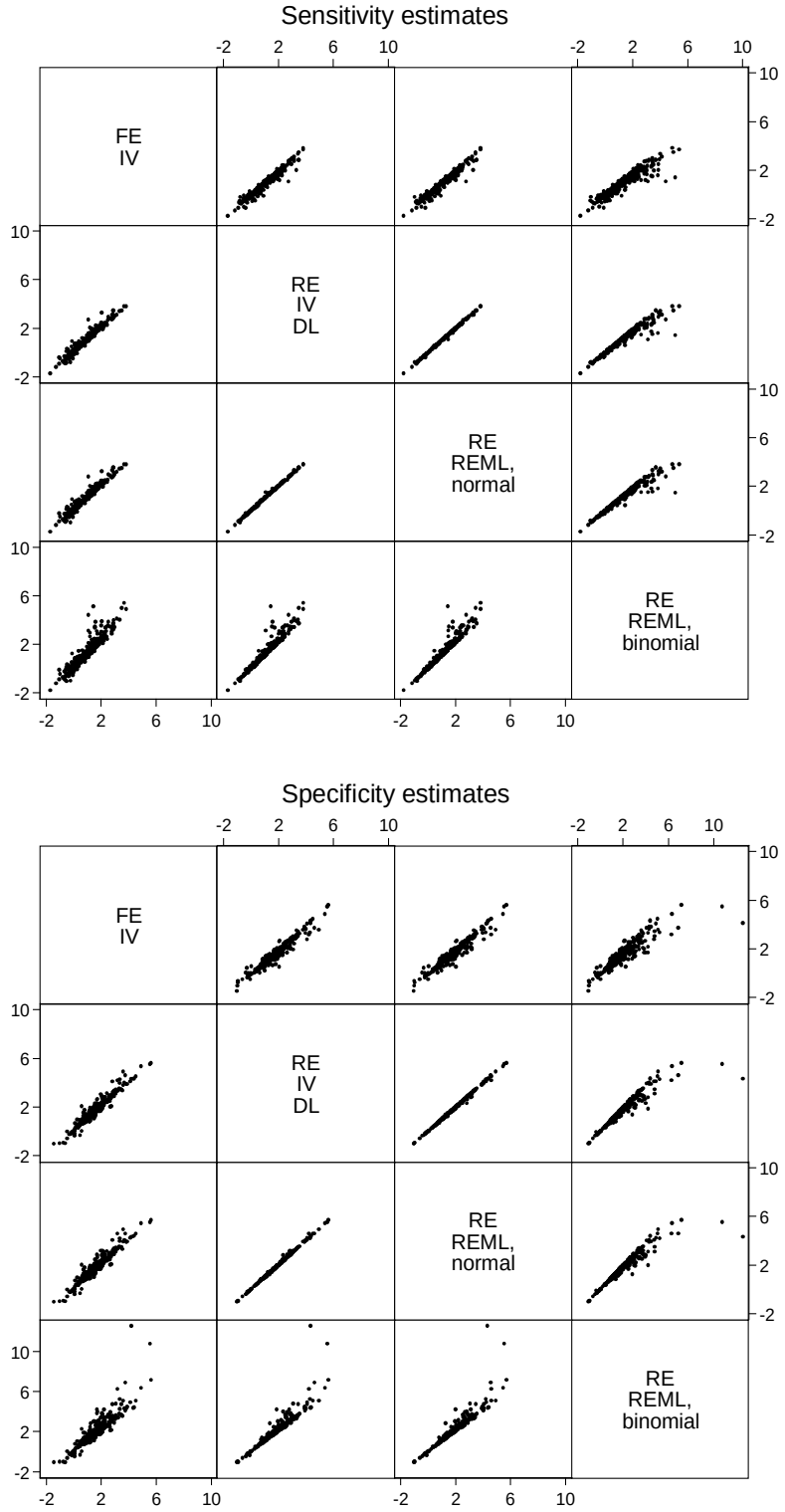


Difference plot of sensitivity (left panel) and specificity (right panel) estimates comparing univariate random effects meta-analyses using the exact binomial likelihood versus using a normal approximation for within-study variability (both models fit with REML). REML = restricted maximum likelihood.

Summary for univariate meta-analysis methods

Figure 6 summarizes the comparisons of univariate analyses of sensitivity and specificity discussed in this report [fixed effect inverse variance (normal approximation); random effects using the DerSimonian-Laird method (normal approximation); random effects using REML (normal approximation); random effects using REML (exact binomial likelihood)]. Fixed and random effects methods often produced different point estimates, reflecting the different weights assigned to each study by these methods. Among random effects methods, the greatest discrepancies were observed between methods using the exact binomial likelihood versus those relying on the normal approximation. This could be explained by the need for continuity corrections, or the fact that the logit-transformed proportions and their variance are correlated, both of which would result in a downward bias for the summary sensitivity and specificity.

Figure 6: Summary comparison of logit-transformed sensitivity and specificity estimates from all univariate methods considered in this report



Estimates are logit-transformed. Binomial = model using the exact binomial likelihood for within-study variability; DL -= DerSimonian-Laird; FE = fixed effect; IV = inverse variance; normal = model using a normal approximation for within-study variability; RE = random effects; REML = restricted maximum likelihood.

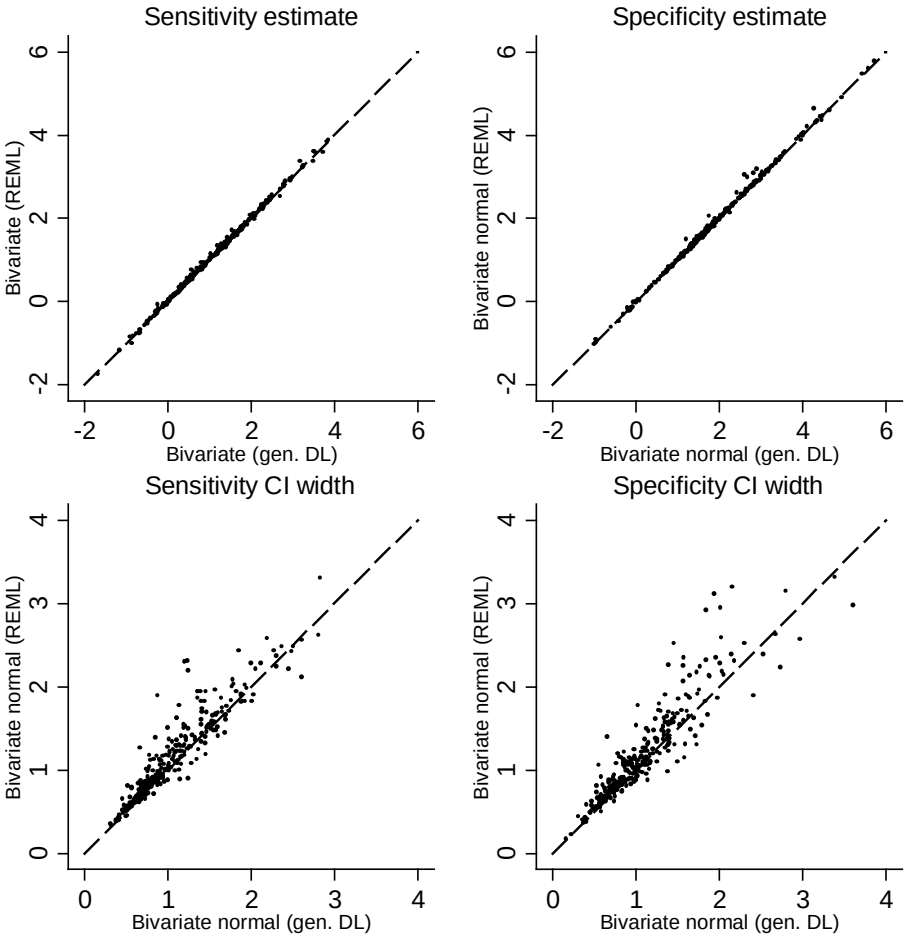
BIVARIATE META-ANALYSIS METHODS

This section compares the results of analyses based on alternative bivariate models for the joint meta-analysis of sensitivity and specificity.

Non-iterative (inverse variance) versus iterative (REML) estimation of between-study variability in bivariate random effects meta-analyses (normal approximation)

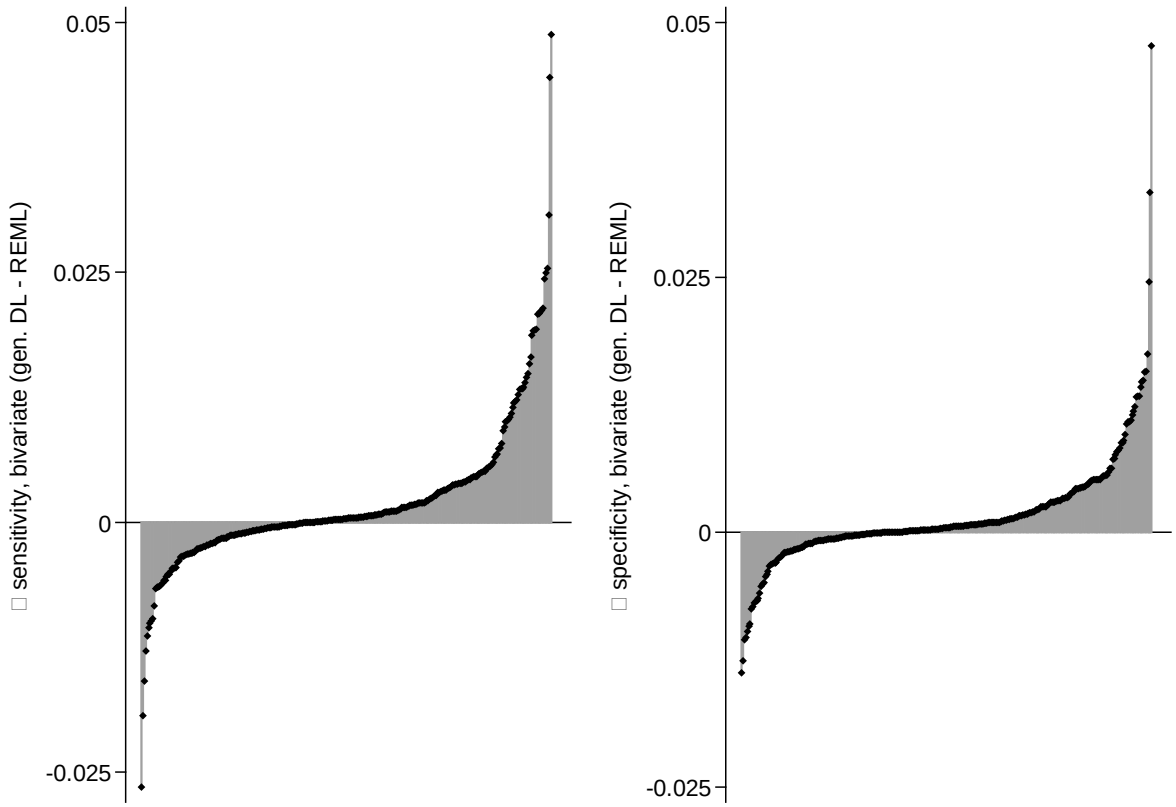
Figure 7 compares the point estimates (logit-transformed) and confidence interval widths from bivariate random effects meta-analyses of sensitivity and specificity using a normal approximation to model within-study variability. We compare models that use a non-iterative estimator of between-study variance (a generalization of the DerSimonian-Laird method) versus an iterative estimator obtained with REML (24). Point estimates from the two methods were almost identical. However, REML estimation often (but not always) resulted in greater wider confidence intervals compared to the non-iterative method. This possibly reflects the inability of the non-iterative method to incorporate the uncertainty in the estimation of between-study heterogeneity. **Figure 8** presents the absolute difference (untransformed scale) between estimated sensitivities and specificities from the two methods, indicating that differences in summary estimates higher than 2.5% are uncommon.

Figure 7: Comparison of point estimates and confidence interval widths of summary sensitivity and specificity from bivariate random effects methods using a normal approximation (generalized DerSimonian-Laird inverse variance vs. REML)



Scatter plot of estimated logit-transformed sensitivity, specificity and their corresponding confidence interval widths from bivariate random effects meta-analyses using τηηηε generalized DerSimonian-Laird and REML methods (both with a normal approximation to represent within-study variability). CI = confidence interval width; gen. DL = generalized DerSimonian-Laird; REML = restricted maximum likelihood.

Figure 8: Absolute differences of estimated sensitivity and specificity from bivariate random effects meta-analyses with generalized DerSimonian-Laird inverse variance vs. REML (both models using a normal approximation to represent within-study variability)



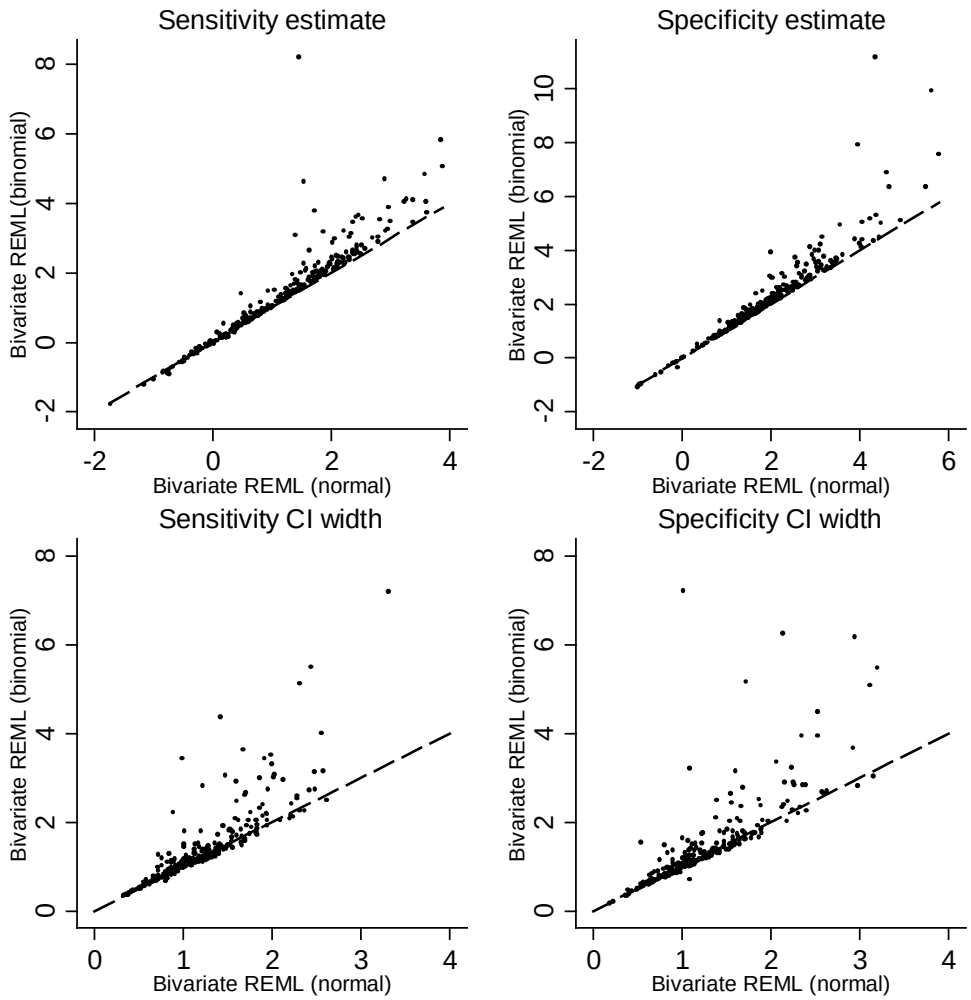
Difference plot of sensitivity (left panel) and specificity (right panel) estimates comparing bivariate random effects meta-analysis using a normal approximation for 2 estimation methods: generalized DerSimonian-Laird and REML. Gen. DL = generalized DerSimonian-Laird; REML = restricted maximum likelihood.

Approximate normal versus exact binomial bivariate random-effects meta-analyses (REML)

Figure 9 compares the point estimates (logit-transformed) and confidence interval widths for bivariate random effects meta-analyses of sensitivity and specificity using the exact binomial likelihood versus using a normal approximation for within-study variability. It is evident that point estimates from the two methods are often dissimilar, and that differences are greater toward high sensitivity and specificity values (when they approach 1); this most likely follows from the need for continuity corrections when the normal approximation is used, which will tend to bias the estimated proportion toward 0.5. Further, in many cases the exact likelihood results in wider confidence intervals,. **Figure 10** presents the absolute difference (untransformed scale) between estimated sensitivities and specificities, indicating that differences higher than 5% are not common

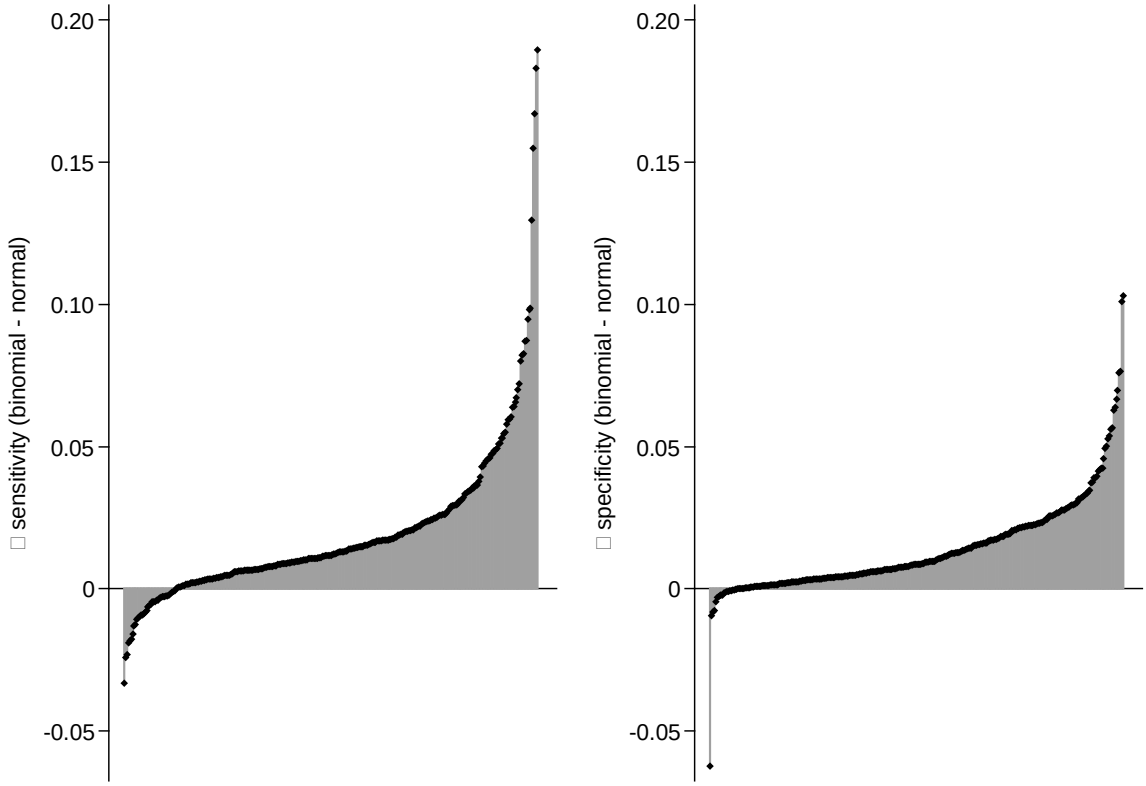
and typically suggest underestimation of sensitivity and specificity in meta-analyses using a normal approximation (compared to those using the exact binomial likelihood).

Figure 9: Comparison of point estimates and confidence interval widths of summary sensitivity and specificity from bivariate random effects meta-analyses (approximate normal versus exact binomial; both models fit with REML)



Scatter plot of estimated logit-transformed sensitivity, specificity and their corresponding confidence interval widths from bivariate random effects meta-analyses using the exact binomial likelihood versus using an approximate normal likelihood to describe within-study variability. CI = confidence interval; REML = restricted maximum likelihood.

Figure 10: Absolute differences of estimated sensitivity and specificity from bivariate random effects meta-analyses (approximate normal versus exact binomial; both models fit with REML)

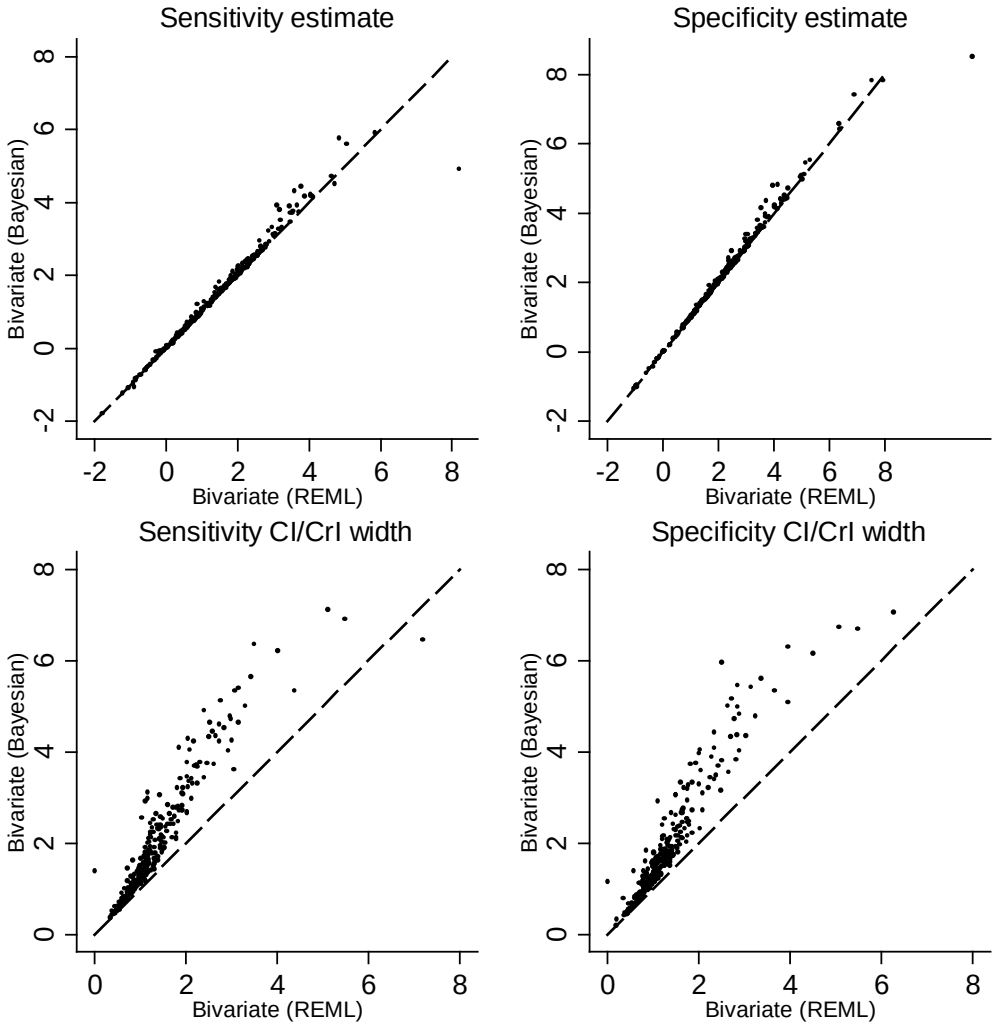


Difference plot of sensitivity (left panel) and specificity (right panel) estimates comparing bivariate random effects meta-analysis using the exact binomial likelihood versus using a normal approximation for within-study variability. REML = restricted maximum likelihood.

Bayesian versus REML fitting for the binomial bivariate random-effects model (exact likelihood)

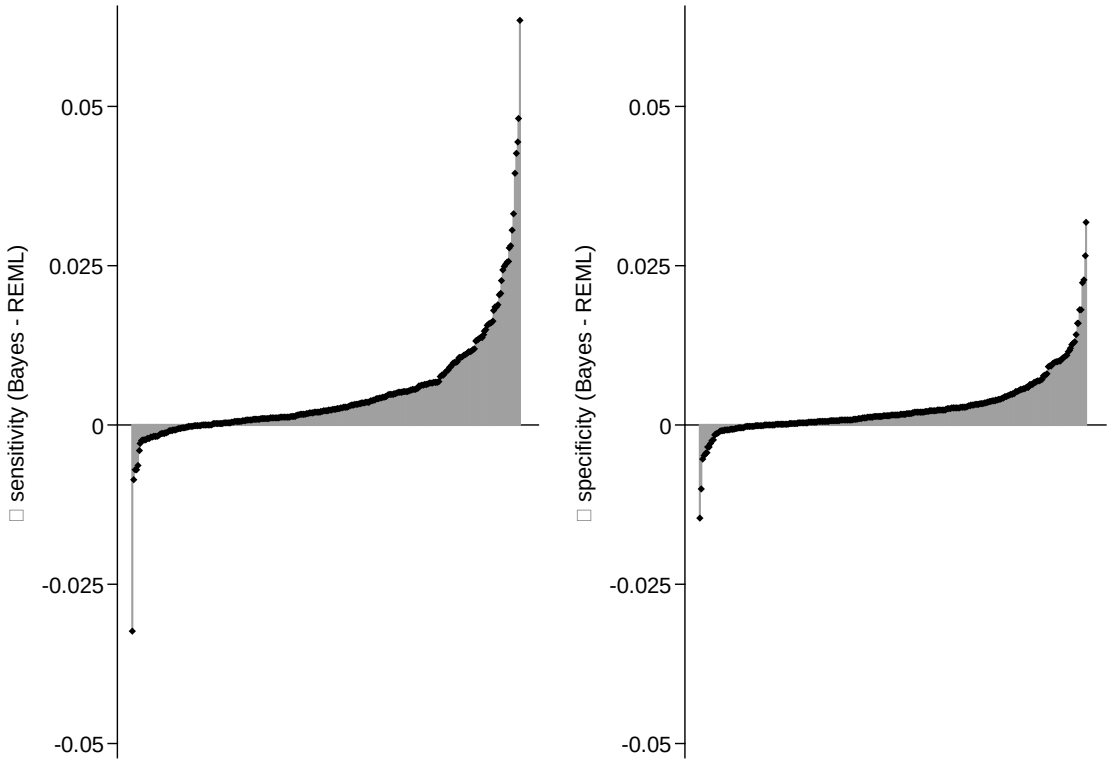
Figure 11 compares the point estimates (logit-transformed) and confidence or credibility interval widths for bivariate random effects meta-analyses of sensitivity and specificity fit using fully Bayesian and REML estimation (both based on the exact binomial likelihood for within-study variability). Point estimates from the two methods were almost identical. However, the Bayesian model typically resulted in substantially larger credibility interval widths, indicating greater uncertainty around the sensitivity and specificity estimates compared to REML. **Figure 12** presents the absolute difference (untransformed scale) between estimated sensitivities and specificities, confirming that point estimate differences higher than 5% were extremely uncommon.

Figure 11: Comparison of point estimates and confidence/credibility interval widths of summary sensitivity and specificity from bivariate random effects meta-analyses (Bayesian versus REML; both models using the exact binomial likelihood to represent within-study variability)



Scatter plot of estimated logit-transformed sensitivity, specificity and their corresponding confidence interval widths from bivariate random effects models using fully Bayesian versus REML estimation (both using the exact binomial likelihood to represent within-study variability). CI = confidence interval; CrI = credibility interval; REML = restricted maximum likelihood.

Figure 12: Absolute differences of estimated sensitivity and specificity from bivariate random effects meta-analyses fit using fully Bayesian versus REML estimation (using the exact binomial likelihood to represent within-study variability)

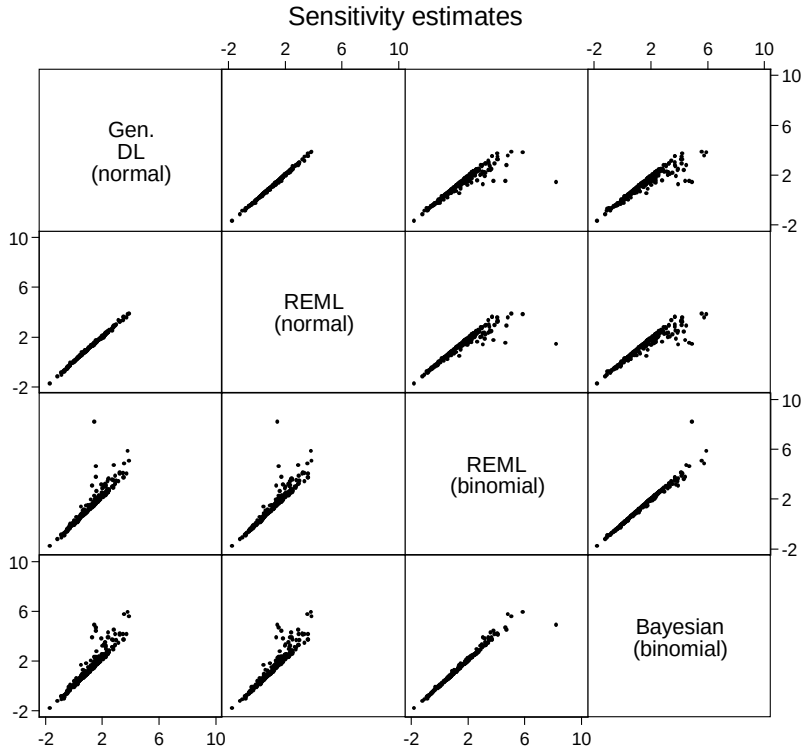


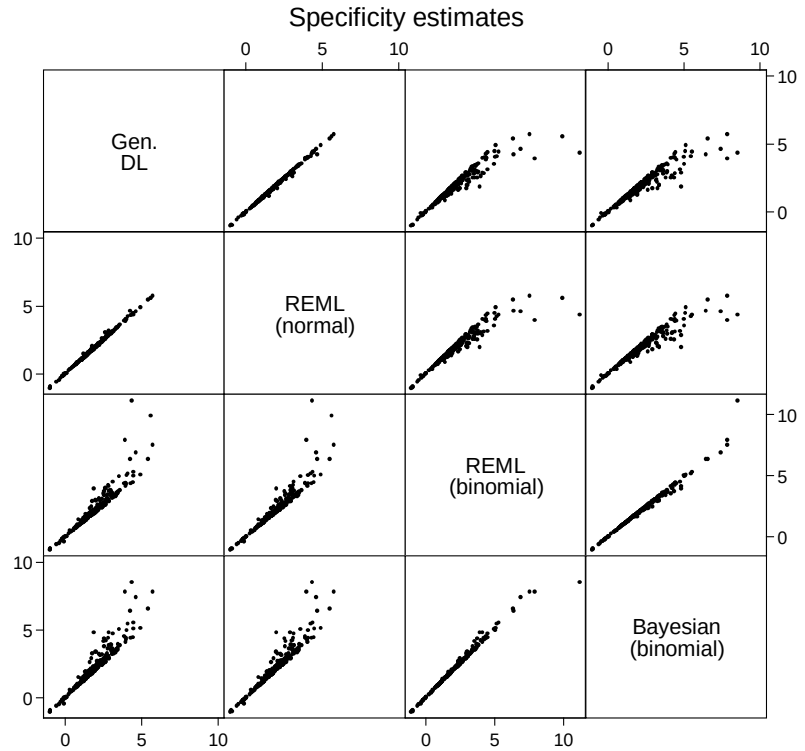
Difference plot of sensitivity (left panel) and specificity (right panel) estimates from bivariate random effects meta-analysis models fit using fully Bayesian versus REML estimation (both models used the exact binomial likelihood to represent within-study variability). REML = restricted maximum likelihood.

Summary for bivariate meta-analysis methods

Figure 13 summarizes the comparisons of bivariate analyses of sensitivity and specificity discussed in this report [random effects using a generalization of the DerSimonian-Laird method (normal approximation); random effects using REML (normal approximation); random effects using REML (exact binomial likelihood); and Bayesian random effects using the exact binomial likelihood]. The greatest discrepancies in meta-analytic point estimates were observed between methods using the exact binomial likelihood versus those relying on the normal approximation. This could be explained by the need for continuity corrections, or the fact that the logit-transformed proportions and their variance are correlated, both of which would result in a downward bias for the summary sensitivity and specificity.

Figure 13: Summary comparison of logit-transformed sensitivity and specificity estimates from all univariate methods considered in this report





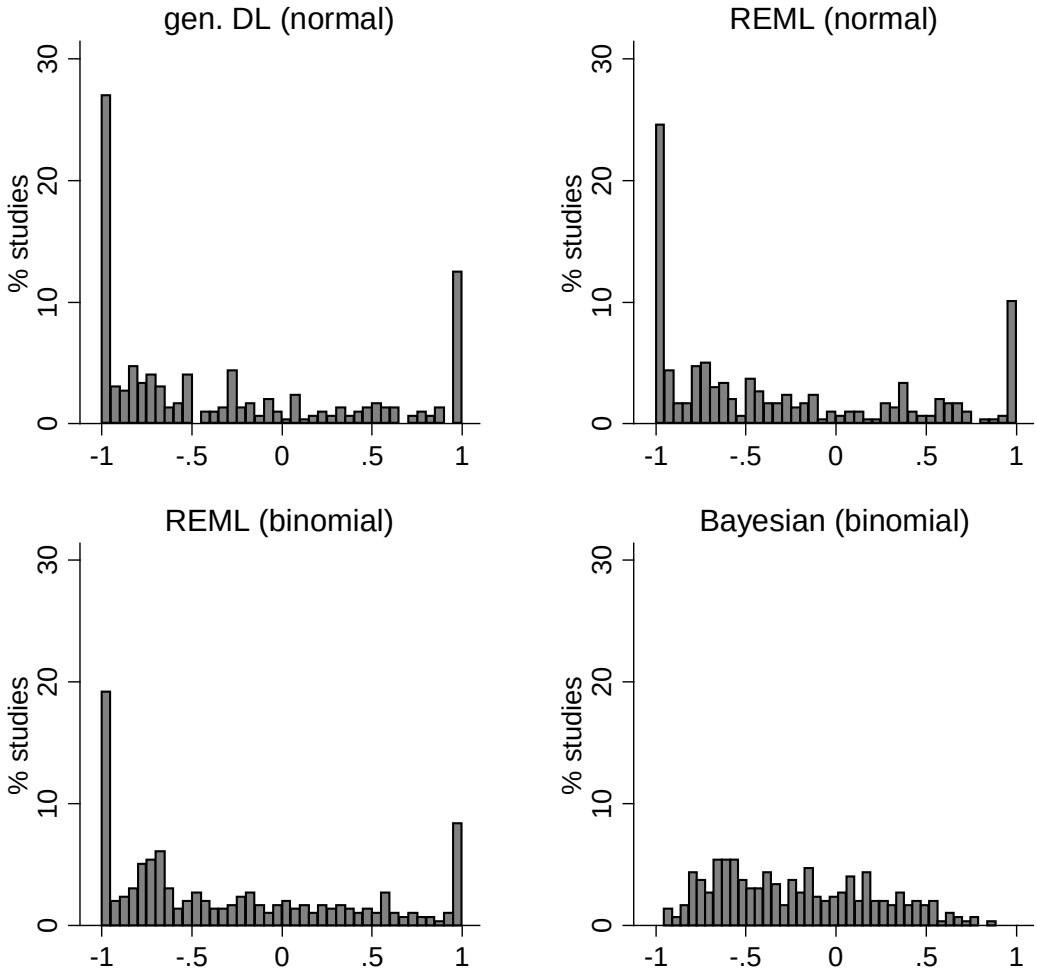
Binomial = model using the exact binomial likelihood for within-study variability; gen DL= generalized DerSimonian-Laird; normal = model using a normal approximation for within-study variability; REML = restricted maximum likelihood.

Estimation of the between-study correlation (inverse variance, REML, and Bayesian bivariate models)

The bivariate models, both using a normal approximation and the exact binomial likelihood, allow estimation of the correlation of sensitivity and specificity at the between-study level. **Figure 14** presents scatter plots of the estimated correlation from four bivariate models: the two models with a normal approximation (fit using either the generalized DerSimonian-Laird method or REML), and two models using the exact binomial likelihood to describe within-study variability (fit using REML or a fully Bayesian approach). It is obvious that these methods produce different distributions of correlations, with the maximum likelihood approaches often returning estimated correlations equal to -1 (and less frequently, 1). In contrast the Bayesian model rarely produces such extreme correlation values. Negative values are more common with all approaches, but the estimated correlations from all methods are sometimes positive (**Table 3**). Negative correlation values are consistent with the existence of threshold effects (i.e., the trade-off between sensitivity and specificity) in meta-analyses of test accuracy. The differences in the correlation estimates are highlighted by **Figure 15**, which is a matrix scatter plot of the results from the same 4 methods. **Figure 16** presents the absolute differences in the correlation estimates between alternative methods.

Generally, as is evident from the relatively wide confidence intervals, the correlation parameter is poorly estimated. As an example, **Figure 17** shows the correlation point estimates and corresponding 95% confidence intervals from 227 studies for which they could be calculated (for studies where the correlation was estimated to be -1 or +1, confidence intervals could not be calculated). The confidence interval excluded negative correlation values in 5 of the 74 cases where the point estimate of the correlation was positive; in contrast, the confidence interval excluded positive correlation values in 43 of the 153 cases where the point estimate of the correlation was negative (Fisher exact $p<0.001$ for the difference in statistical significance).

Figure 14: Histograms of the estimated correlation between sensitivity and specificity from the three bivariate methods compared in this report.



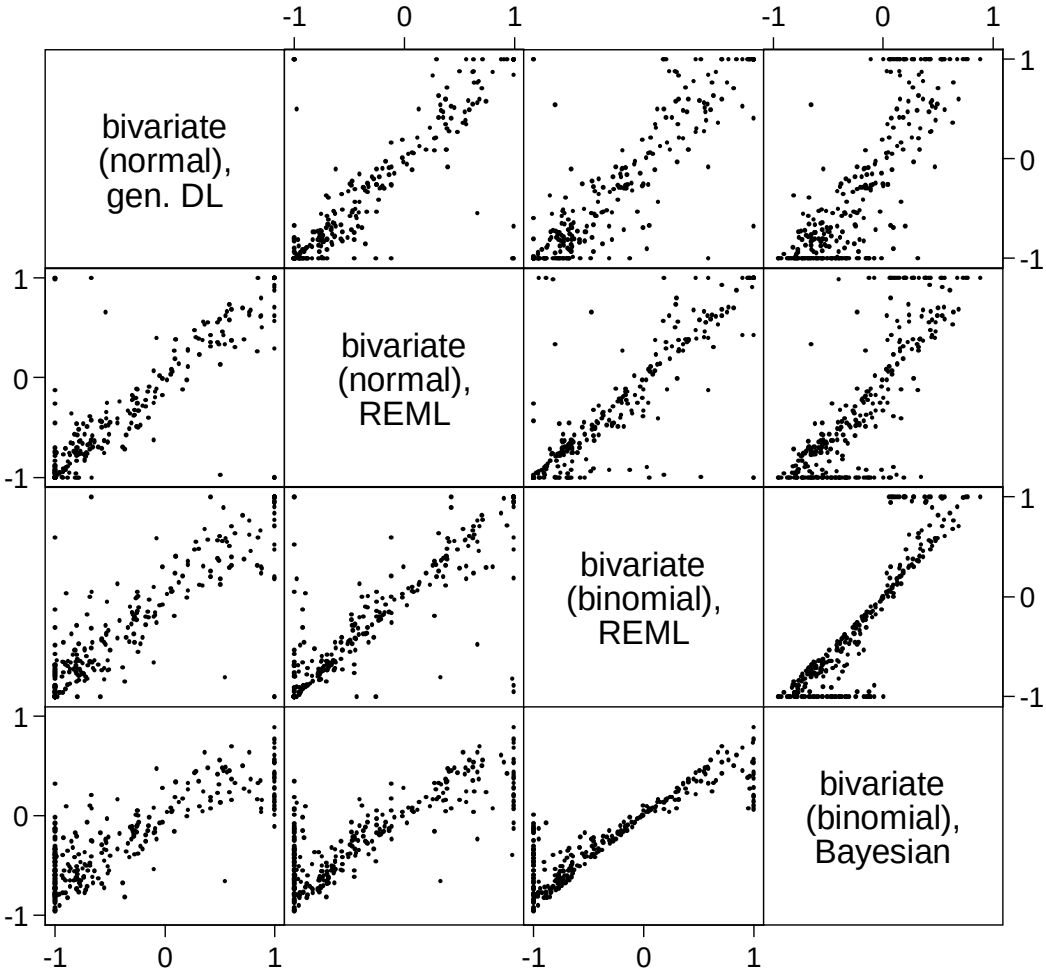
Histograms of the estimated correlation between sensitivity and specificity from bivariate random effects meta-analysis models using: (top left) normal approximation with generalized DerSimonian-Laird (gen.DL); (top right) normal approximation wth restricted maximum likelihood (REML); (bottom left) exact binomial likelihood with REML; (bottom right) exact binomial likelihood with Bayesian model..

Table 3: Estimated correlation by different bivariate methods

Method (model /approach for within-study variability/ model fitting method)	Estimated correlation <0 N (%)
BREMA, normal, gen. DL*	207 (70)
BREMA, normal, REML	209 (70)
BREMA, binomial, REML	202 (68)
BREMA, binomial, Bayesian	199 (66)

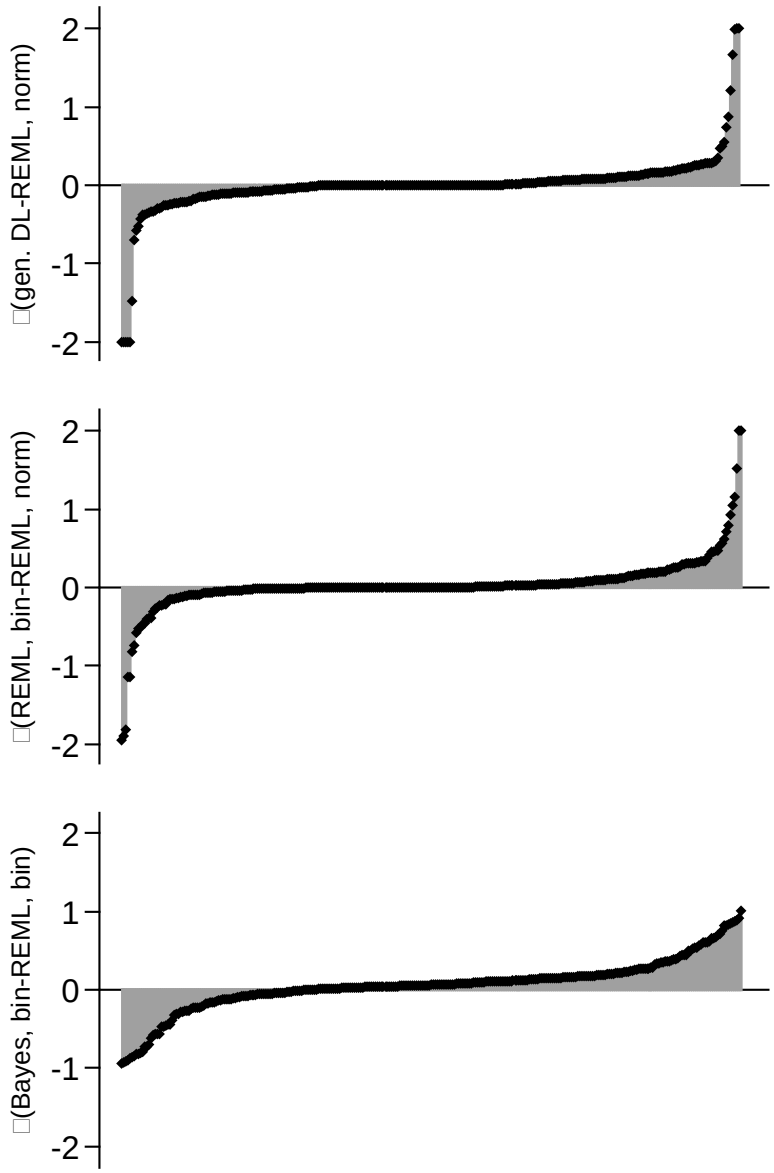
* In one meta-analysis the correlation could not be estimated.
BREMA = bivariate random effects meta-analysis; gen. DL = generalized DerSimonian-Laird; REML = restricted maximum likelihood.

Figure 15: Matrix scatter plot of correlation estimates from the 4 bivariate methods considered in this report



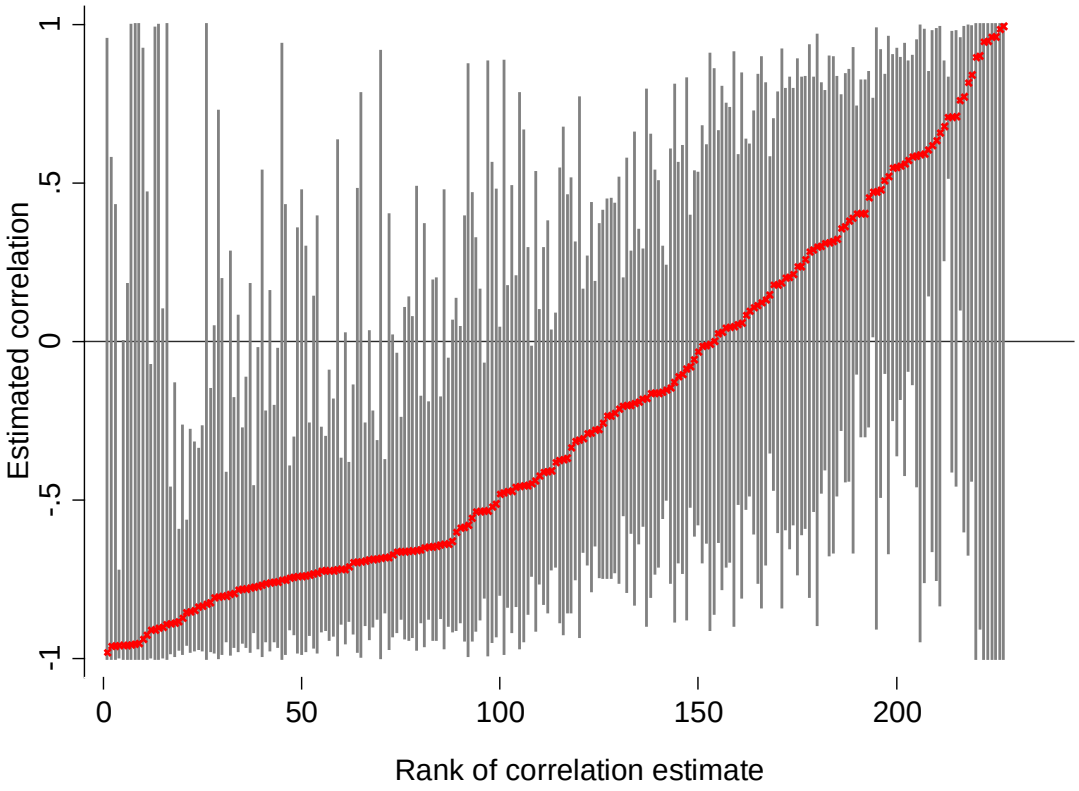
Binomial = model using the exact binomial likelihood to describe within-study variability; Gen. DL = generalized DerSimonian-Laird; normal = model using an approximate normal likelihood to describe within-study variability; REML = restricted maximum likelihood.

Figure 16: Absolute differences of correlation estimates from bivariate random effects meta-analyses (generalized DerSimonian-Laird model using a normal approximation; REML using a normal approximation; REML using the exact binomial likelihood; and fully Bayesian model using the exact binomial likelihood)



Difference plot comparing correlation estimates between models: top panel, generalized DerSimonian-Laird method (with a normal approximation) versus REML (with a normal approximation); middle panel, REML (with the exact binomial likelihood) versus REML (with a normal approximation); bottom panel, Bayesian model (exact binomial likelihood) versus REML (exact binomial likelihood). ρ = correlation; bin. = model using the exact binomial likelihood; norm. = model using a normal approximation; REML = restricted maximum likelihood.

Figure 17: Point estimates and 95% confidence intervals for the correlation of sensitivity and specificity, as estimated by the bivariate model using the exact binomial likelihood (REML estimation)



Estimates from the bivariate model using the exact binomial likelihood for within-study variability. Point estimates are shown as red "x" symbols, extending lines represent 95% confidence intervals. REML = restricted maximum likelihood.

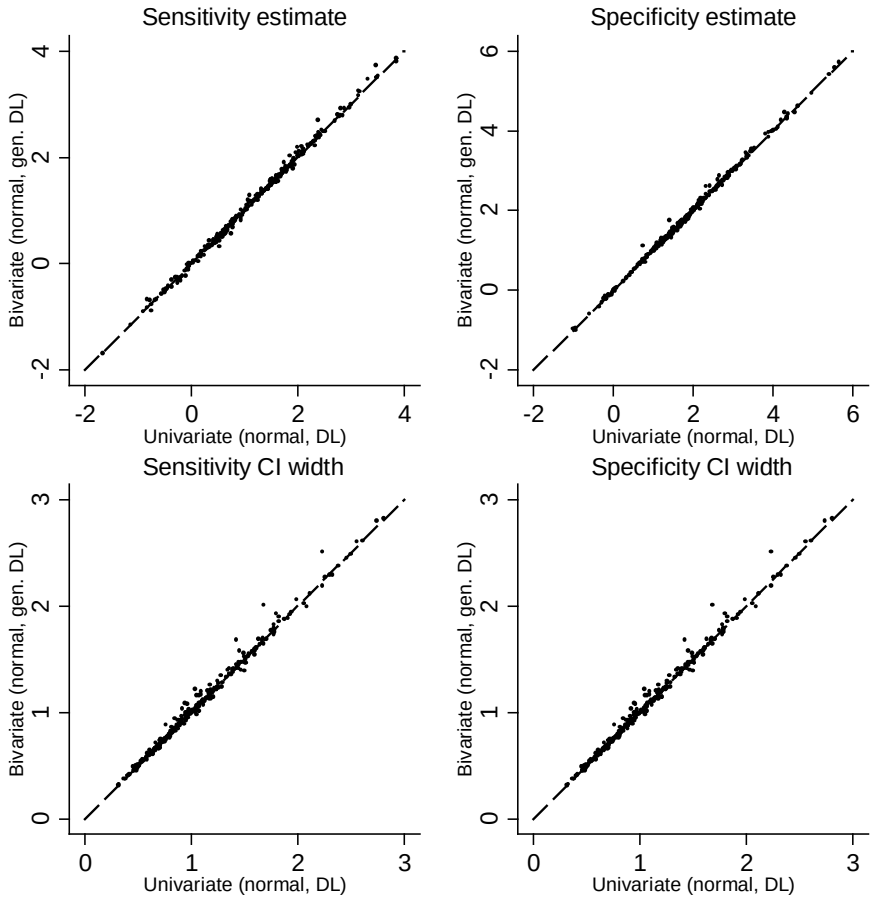
COMPARISON OF UNIVARIATE AND BIVARIATE METHODS

This section compares the results of univariate and bivariate methods for meta-analysis of sensitivity and specificity. We present comparisons that – to the extent possible – "isolate" the effect of bivariate analysis from the effect estimation method (inverse variance versus REML) or the likelihood used to represent within-study variability (exact binomial versus approximate normal), by performing pairwise comparisons where these confounding factors were constant.

Univariate versus bivariate random effects meta-analyses using the normal approximation (DerSimonian-Laird and generalized DerSimonian-Laird methods)

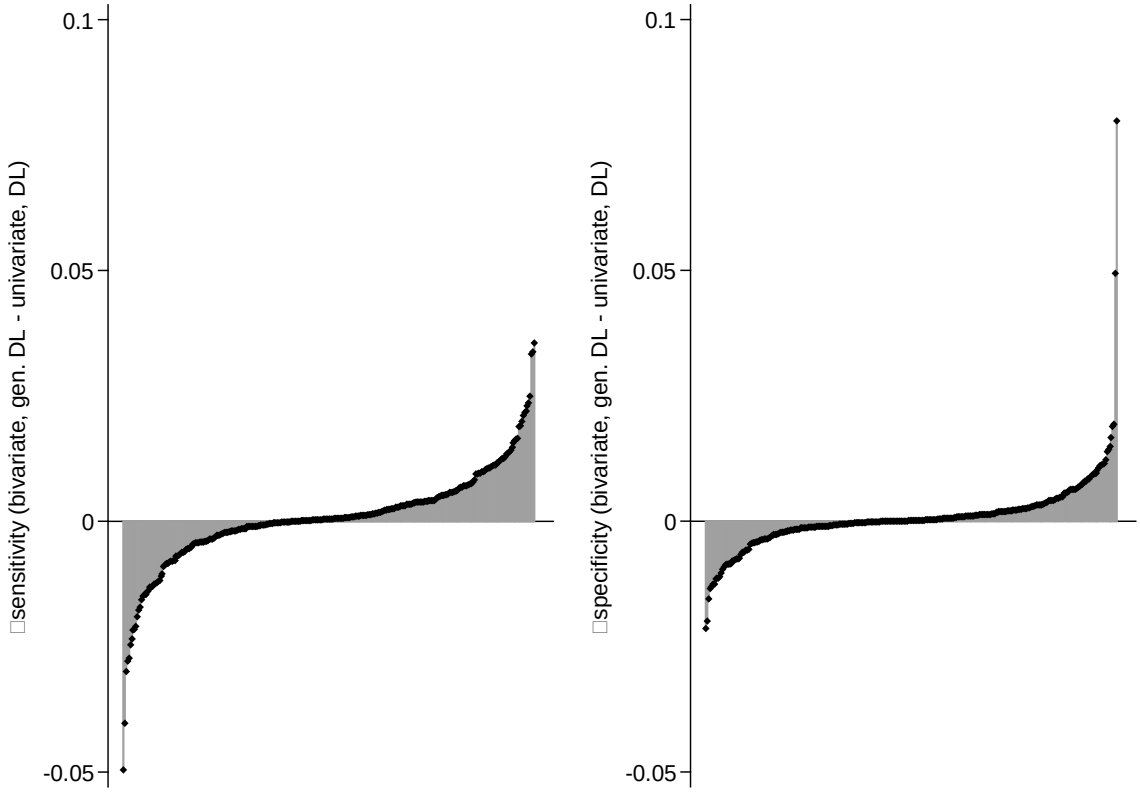
Figure 18 compares the point estimates (logit-transformed) and confidence interval widths for meta-analyses of sensitivity and specificity comparing univariate and bivariate random effects models (using a normal approximation for within-study variability). The univariate analyses used the DerSimonian-Laird method to estimate between-study heterogeneity; the bivariate analyses use a multivariate generalization of the same method. Overall, point estimates and confidence interval widths from the two methods were similar. **Figure 18** presents the absolute difference (untransformed scale) between estimated sensitivities and specificities, which emphasizes the general concordance of the point estimates of the two methods (no differences larger than 10% were observed and only occasional differences beyond 5%).

Figure 18: Comparison of point estimates and confidence interval widths of summary logit-transformed sensitivity and specificity (univariate random effects versus bivariate random effects inverse variance methods, both using a normal approximation for within-study variability)



Scatter plot of estimated logit-transformed sensitivity, specificity and their corresponding confidence interval widths from univariate random effects meta-analyses over bivariate random effects inverse variance (DerSimonian-Laird and generalized DerSimonian-Laird) meta-analysis (with an approximate normal likelihood to describe within-study variability). CI = confidence interval; DL = DerSimonian-Laird; gen. DL = generalized DerSimonian=Laird.

Figure 19: Absolute differences of estimated sensitivity and specificity in univariate versus bivariate random effects inverse variance meta-analyses (both using a normal approximation for within-study variability)

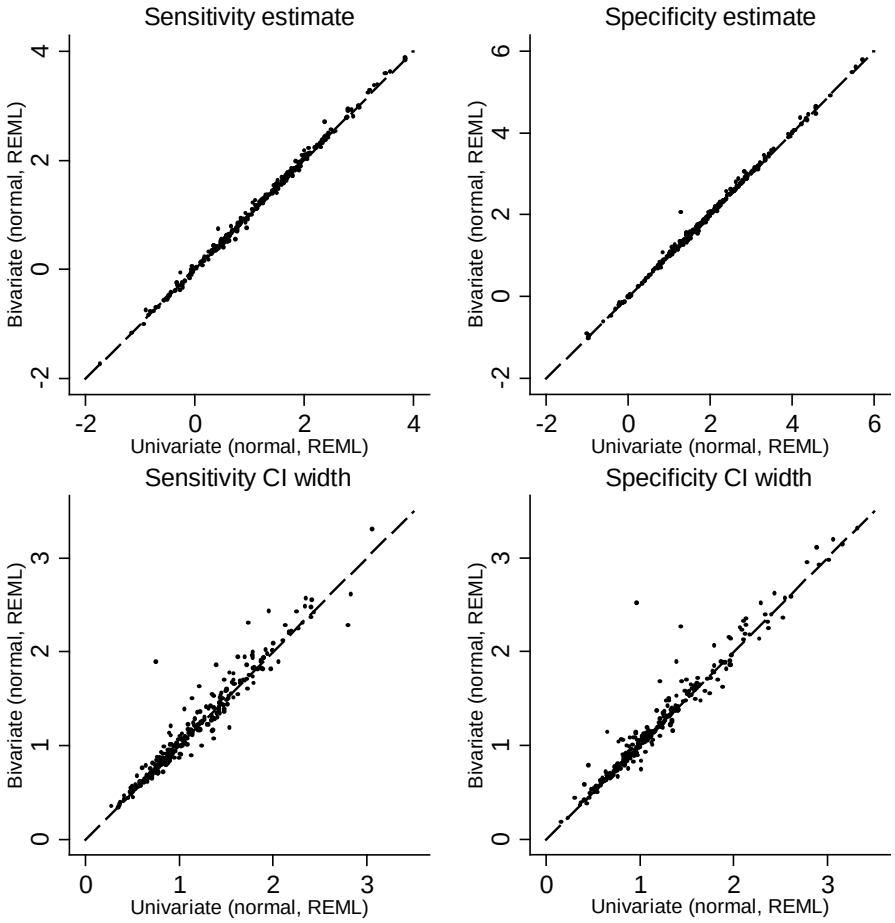


Difference plot of sensitivity (left panel) and specificity (right panel) comparing bivariate versus univariate random effects inverse variance meta-analyses (Der-Simonian-Laird vs. generalized DerSimonian-Laird; both using a normal approximation for within-study variability). DL = DerSimonian-Laird; gen.DL = generalized DerSimonian-Laird.

Univariate versus bivariate random effects meta-analyses using the normal approximation (using REML)

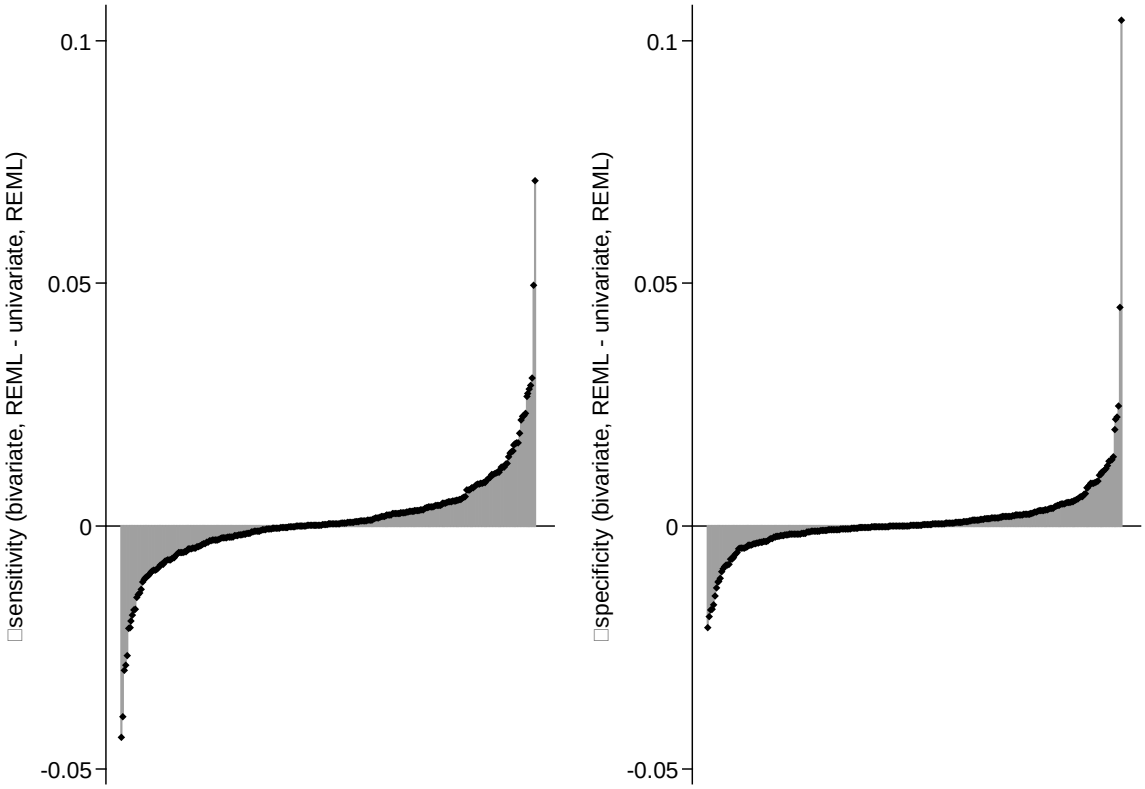
Figure 20 compares the point estimates (logit-transformed) and confidence interval widths for meta-analyses of sensitivity and specificity comparing univariate and bivariate random effects models using REML estimation (with a normal approximation for within-study variability). Overall, point estimates from the two methods are similar, however the uncertainty around the estimates is often different between methods; in many cases the bivariate model produces greater confidence interval widths, indicating larger uncertainty around estimates. **Figure 21** presents the absolute difference (untransformed scale) between estimated sensitivities and specificities from the two models; the graph emphasizes the general concordance of the point estimates of the two methods (no differences larger than 10% were observed; only occasional differences beyond 5%).

Figure 20: Comparison of point estimates and standard errors of summary logit-transformed sensitivity and specificity from univariate versus bivariate random effects meta-analyses with REML estimation (using a normal approximation for within-study variability)



Scatter plot of estimated logit-transformed sensitivity, specificity and their corresponding confidence interval widths from univariate and bivariate random effects meta-analyses fit using REML (with an approximate normal likelihood to represent within-study variability). CI = confidence interval; REML = restricted maximum likelihood.

Figure 21: Absolute differences of estimated sensitivity and specificity from univariate and bivariate random effects meta-analyses with REML estimation (using a normal approximation to represent within-study variability for both models)

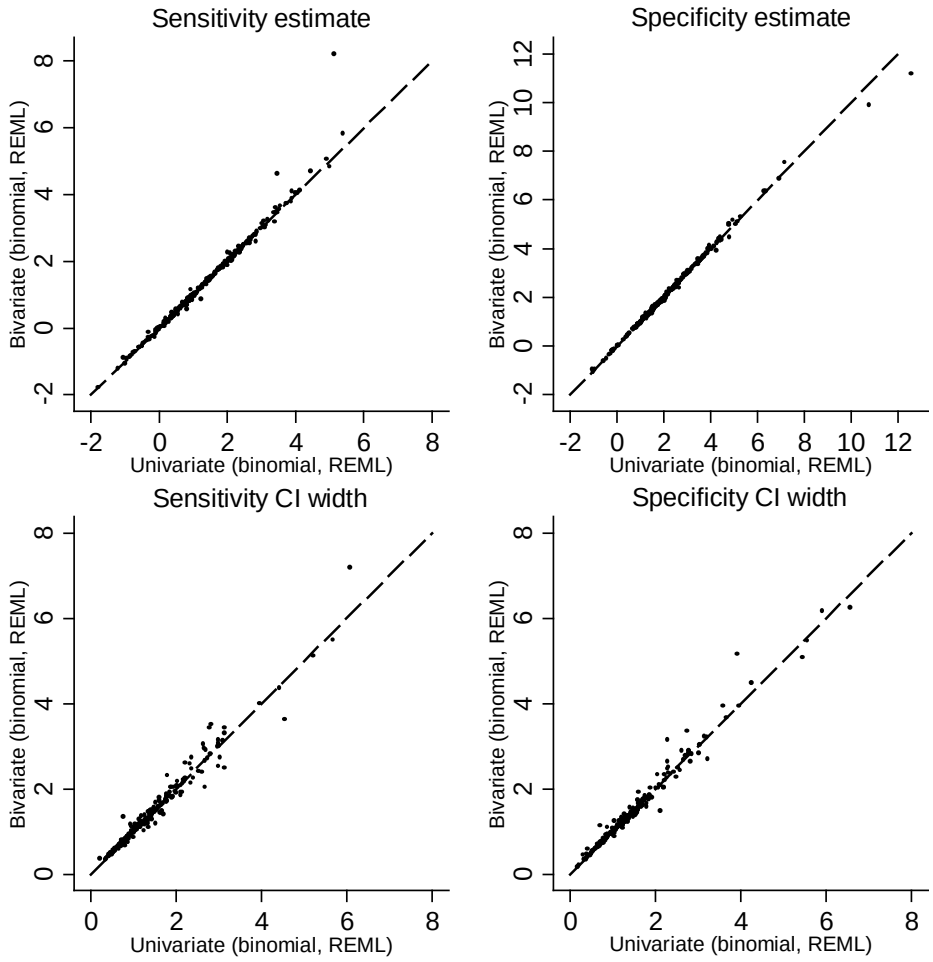


Difference plots of sensitivity (left panel) and specificity (right panel) comparing bivariate versus univariate random effects meta-analyses fit with REML (using a normal approximation to represent within-study variability).

Univariate versus bivariate random effects meta-analyses using the exact binomial likelihood (using REML)

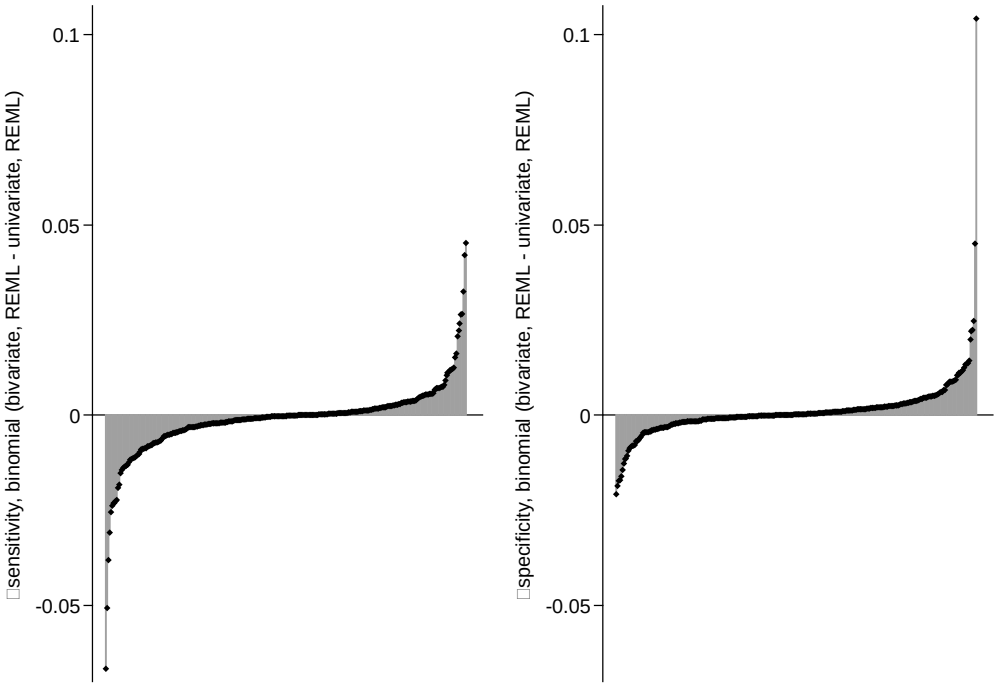
Figure 22 compares the point estimates (logit-transformed) and confidence interval widths for meta-analyses of sensitivity and specificity comparing univariate and bivariate models (using the exact binomial likelihood to represent within-study variability). Overall, point estimates and confidence interval widths from the two methods were similar. **Figure 23** presents the absolute difference (untransformed scale) between estimated sensitivities and specificities, which emphasizes the general concordance of the point estimates of the two methods (only one discrepancy beyond 5%).

Figure 22: Comparison of point estimates and confidence interval widths of summary logit-transformed sensitivity and specificity from univariate versus bivariate random effects meta-analyses using the exact binomial likelihood (both models fit using REML)



Scatter plot of estimated logit-transformed sensitivity, specificity and their corresponding confidence interval widths from univariate and bivariate random effect meta-analyses fit using REML (both models using the exact binomial likelihood to represent within-study variability). CI = confidence interval; REML = restricted maximum likelihood.

Figure 23: Absolute differences of estimated sensitivity and specificity comparing univariate versus bivariate random effects meta-analyses fit with REML (both models using the exact binomial likelihood to represent within-study variability)

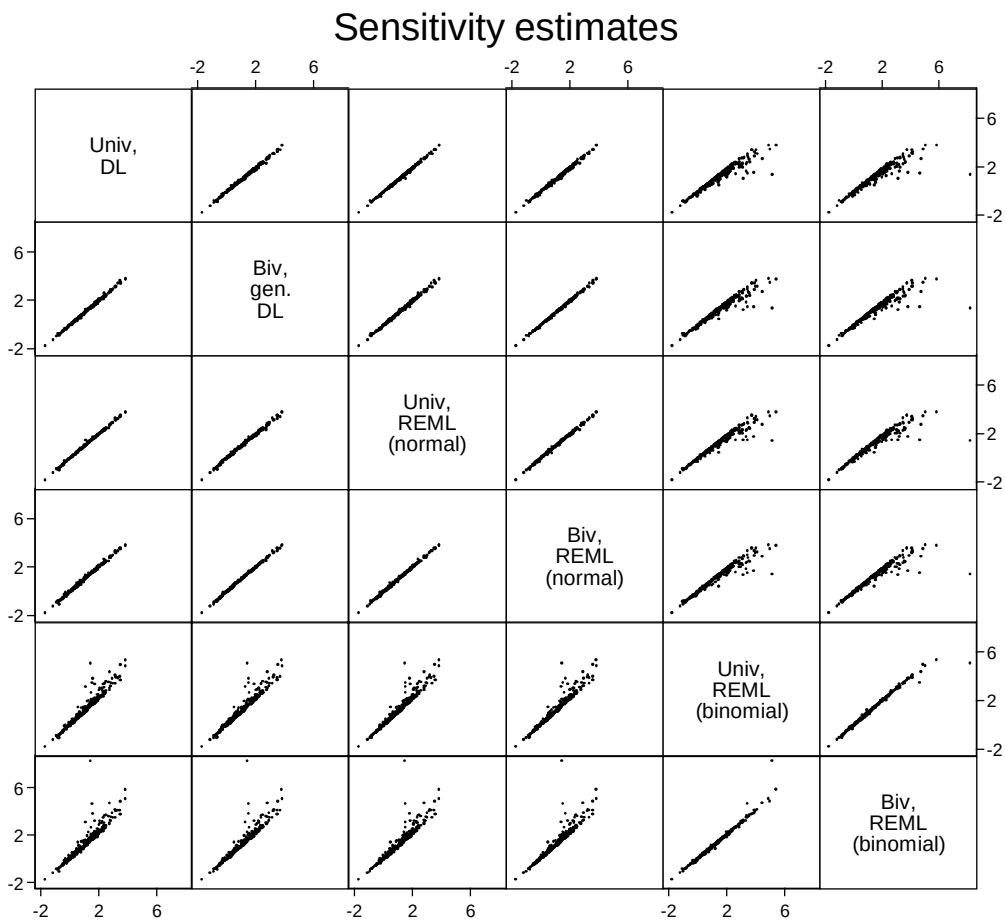


Difference plot of sensitivity (left panel) and specificity (right panel) comparing bivariate versus univariate random effects models fit using REML (both models used the exact binomial likelihood to represent within-study variability). REML = restricted maximum likelihood.

Summary for the comparison of univariate and bivariate methods

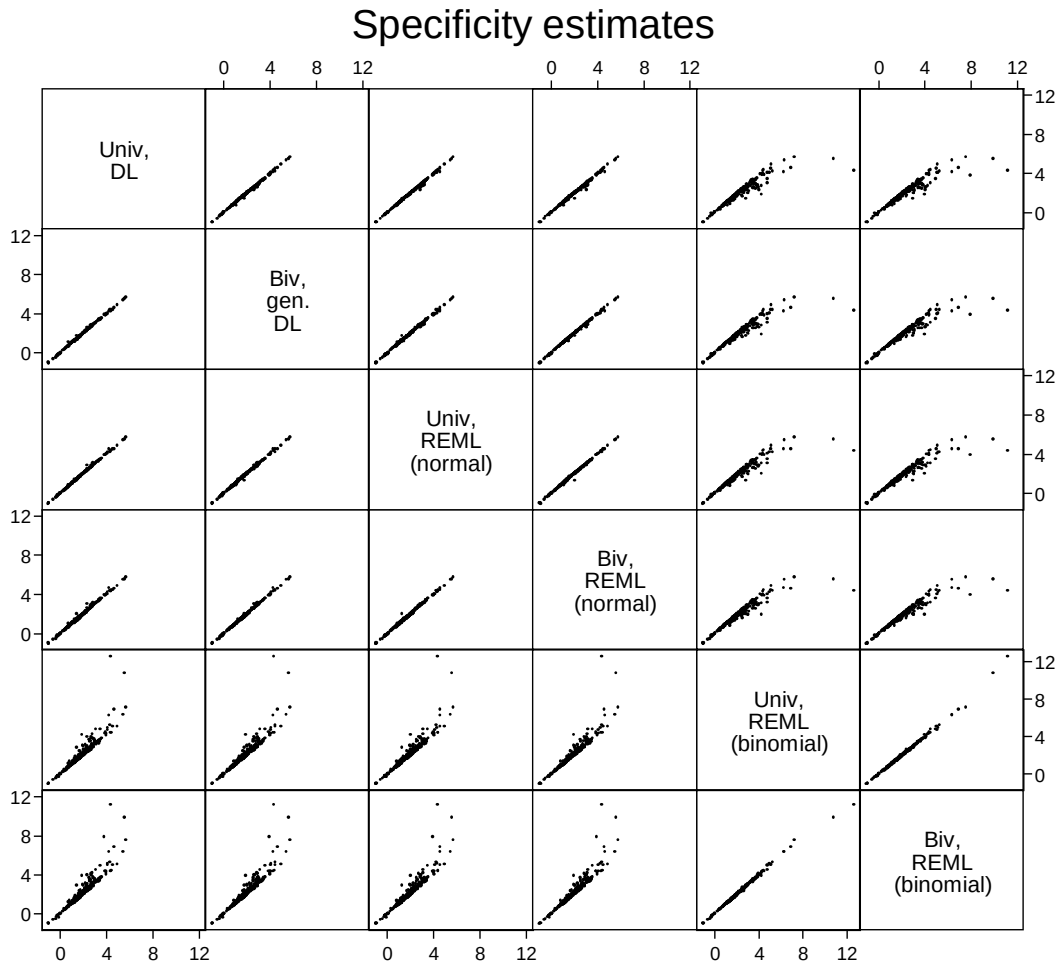
Figures 24 and **25** summarize the comparisons of effect sizes from all non-Bayesian random effects methods used in this report (univariate and bivariate; DerSimonian-Laird/generalized DerSimonian-Laird and REML; using the exact binomial likelihood or a normal approximation). Overall, the greatest discrepancies were observed between methods using the exact binomial likelihood versus those relying on the normal approximation. The point estimates from univariate and bivariate meta-analyses were similar both for methods using non-iterative (DerSimonian-Laird/generalized DerSimonian-Laird) or iterative (REML) estimators of between-study variance.

Figure 24: Summary comparison of logit-transformed sensitivity and specificity estimates from all methods considered in this report



Binomial = model using the exact binomial likelihood; biv = bivariate; gen. DL = generalized DerSimonian-Laird method; DL = DerSimonian-Laird model; normal = model using a normal approximation; REML = restricted maximum likelihood; univ = univariate.

Figure 25: Summary comparison of logit-transformed sensitivity and specificity estimates from all methods considered in this report



Binomial = model using the exact binomial likelihood; biv = bivariate; gen. DL = generalized DerSimonian-Laird method; DL = DerSimonian-Laird model; normal = model using a normal approximation; REML = restricted maximum likelihood; univ = univariate.

Meta-analysis in the receiver operating characteristic space

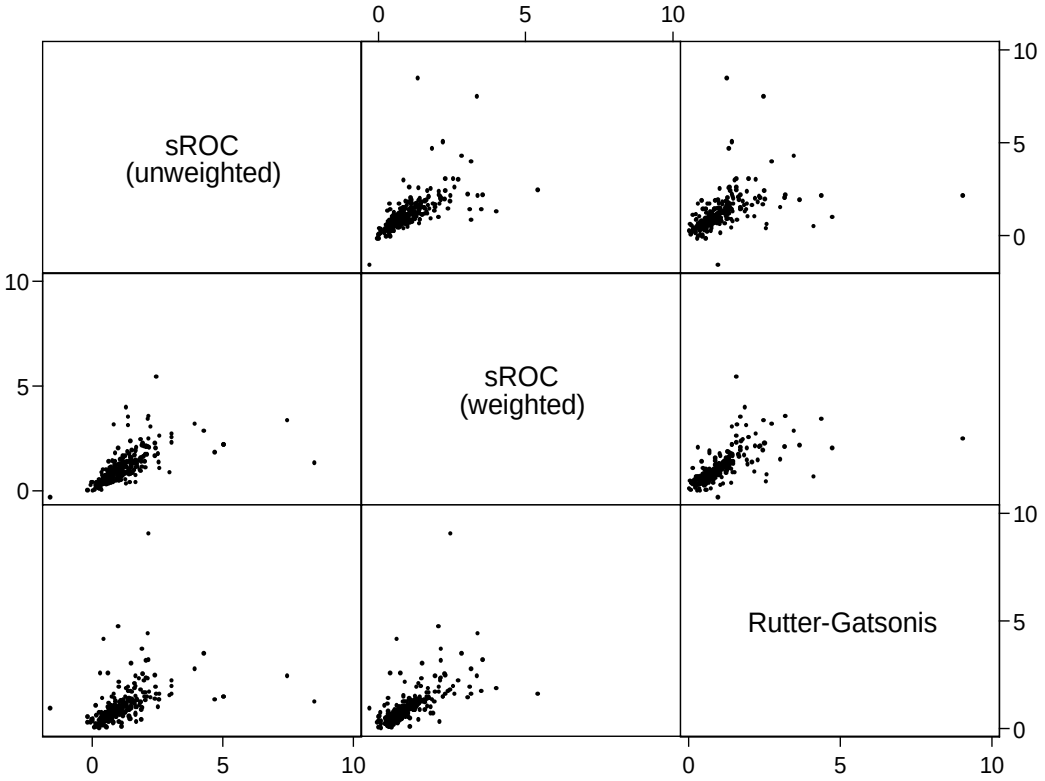
Methods for constructing ROC curves work by transforming a straight line that describes the bivariate distribution of logit-transformed sensitivity and specificity to the ROC space. The most common method used to obtain estimates of the intercept and slope of the straight line is the model proposed by Moses and Littenberg. Their approach is based on regressing the difference (D) of the logit-transformed sensitivity (η) and false positive rate (1-specificity = ξ) over their sum (S). Typically this method is implemented in a fixed effects framework (using both weighted and unweighted regression) and measurement error in S (the sum of logit-transformed sensitivity and specificity) is ignored. Hierarchical regression methods that overcome these limitations have been proposed but it is unclear whether they result in substantially different estimates of

performance. The HsROC approach proposed by Rutter and Gatsonis is the method used in most published meta-analyses that employ such hierarchical models.

Moses-Littenberg sROC vs. Rutter-Gatsonis HsROC

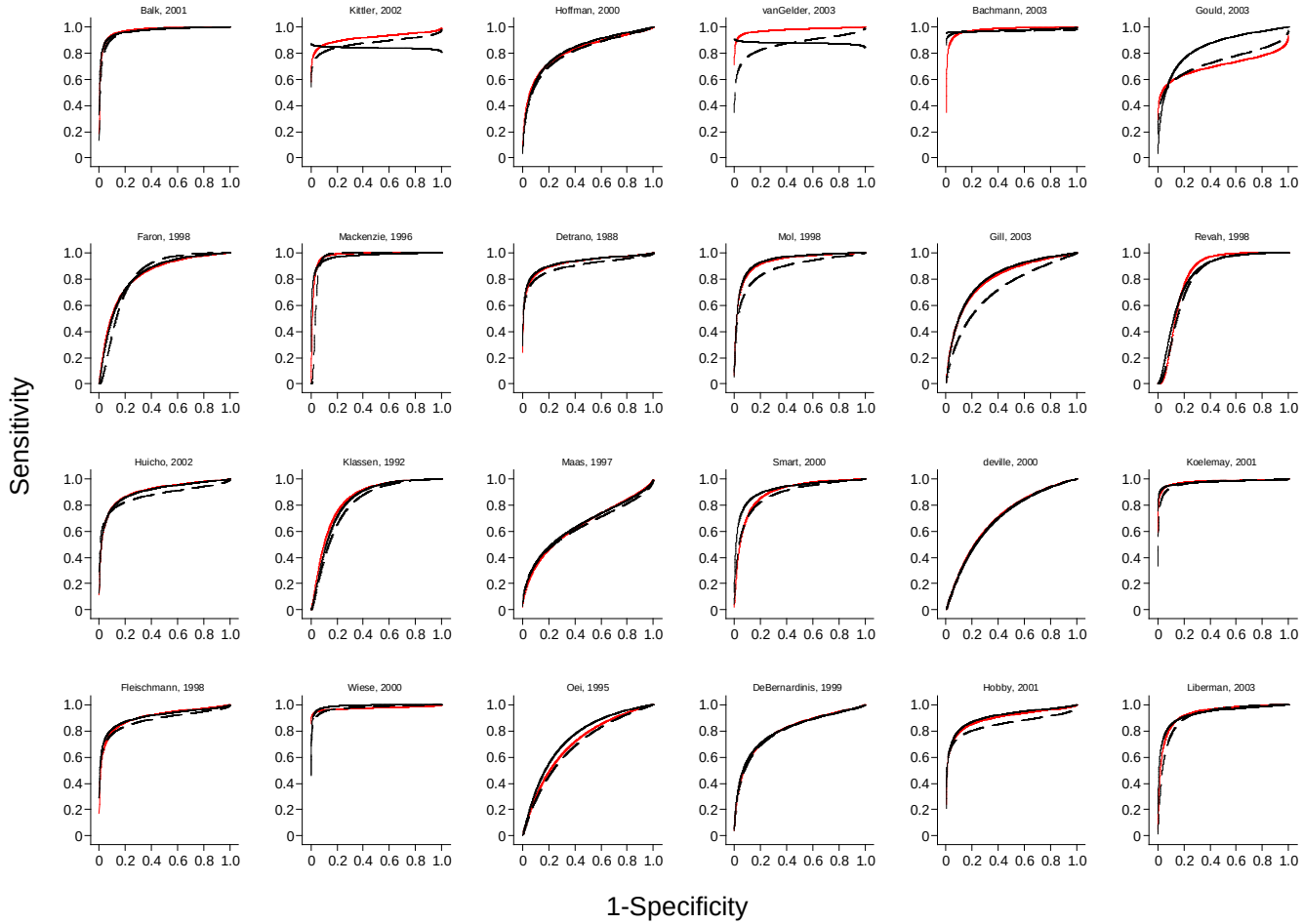
Figure 26 compares the slope in the logit space of ROC lines produced by the Moses-Littenberg (weighted and unweighted) and Rutter-Gatsonis methods. **Figure 27** presents the sROC curves produced from these 3 methods for 24 randomly selected examples.

Figure 26: Scatter plot of the slopes of alternative sROC lines (logit space)



Alternative sROC curves based on bivariate meta-analysis (fit using maximum likelihood) or the Moses-Littenberg method. Examples that resulted in extreme slope values (>10 or <-10) are not shown to avoid distortion of the graph. sROC = summary receiver operating characteristic curve.

Figure 27: sROC curves for 24 randomly selected meta-analyses (bivariate random effects versus Moses-Littenberg methods)



Meta-analytic ROC curves for 24 randomly selected meta-analyses. The title for each panel contains the first author and year of publication for the corresponding review. HsROC curves were produced with the Rutter-Gatsonis HsROC method (red lines), and the unweighted (solid black line) and weighted (dashed black line) Moses-Littenberg methods. HsROC = hierarchical summary receiver operating characteristic curve; sROC = receiver operating characteristic curve.

Different sROC curves based on the bivariate model

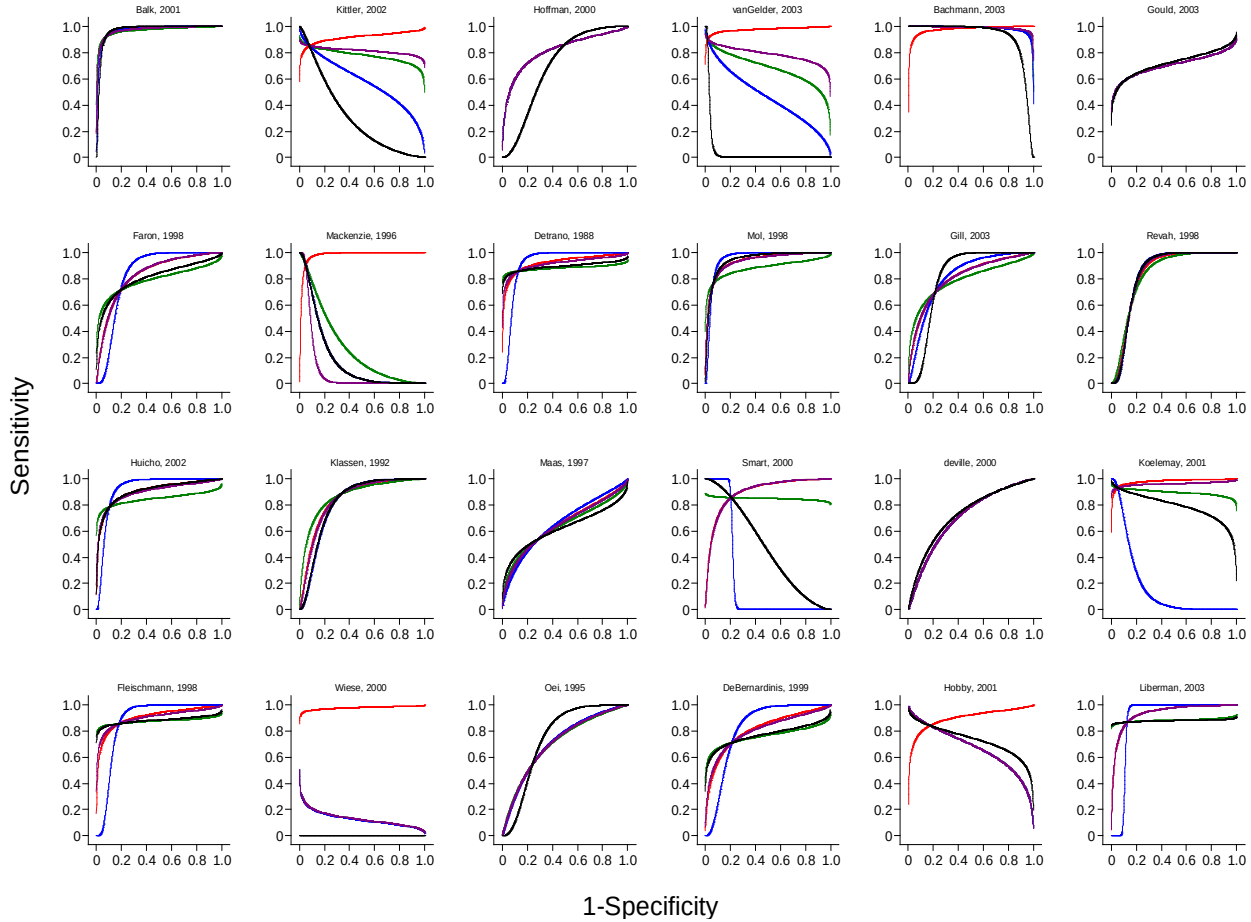
The Rutter-Gatsonis HsROC curve is only one possible parameterization of the meta-analytic ROC curve. We followed Arends 2008 (25) to obtain alternative HsROC curves based on the results of the bivariate random effects meta-analysis model using the exact binomial likelihood (fit with REML). In **Appendix C** we present the regression lines corresponding to each alternative model. Importantly, the Rutter-Gatsonis HsROC curve always has a positive slope (because the slope is equal to the ratio of the variances of the logit-transformed sensitivity over the variance of the logit-transformed false positive rate). In contrast, the slopes of other HsROC curves are not always positive: the slopes estimated by MAR of $\eta \sim \xi$, regression of $\eta \sim \xi$, and of $\xi \sim \eta$ will be negative whenever the correlation between logit-sensitivity and logit-specificity is positive (i.e. when the covariance between logit-transformed sensitivity and false positive rate is negative); the slope of the regression line corresponding to the "D on S" model may also be negative in some cases, but this will depend on the relative values of the variances of logit-transformed sensitivity and specificity and their covariance.

Figure 28 presents the alternative sROC curves discussed above for 24 randomly selected meta-analyses: the Rutter-Gatsonis slopes are always positive, whereas in some examples the slopes from the D~S, MAR of $\eta \sim \xi$, $\eta \sim \xi$, and $\xi \sim \eta$ models, are negative. The latter three models "track together" (i.e. either all have positive slopes or all have negative slopes). In contrast there are some examples where the D~S model has a positive slope when the MAR, $\eta \sim \xi$, and $\xi \sim \eta$ models have negative slopes.

Figure 29 presents the study-level data and fitted hsROC curves for one of the datasets we analyzed(31) (fourth dataset on the top row of **Figure 28**). This was a case where all regression methods (except the Rutter-Gatsonis model) resulted in a negative slope. The data points from the studies in this meta-analysis were clustered in the top left corner of the ROC space, many studies had sensitivities and specificities near 1, and the estimated correlation between sensitivity and specificity was positive. Note also that the differences between the fitted lines are pronounced outside the observed region of data.

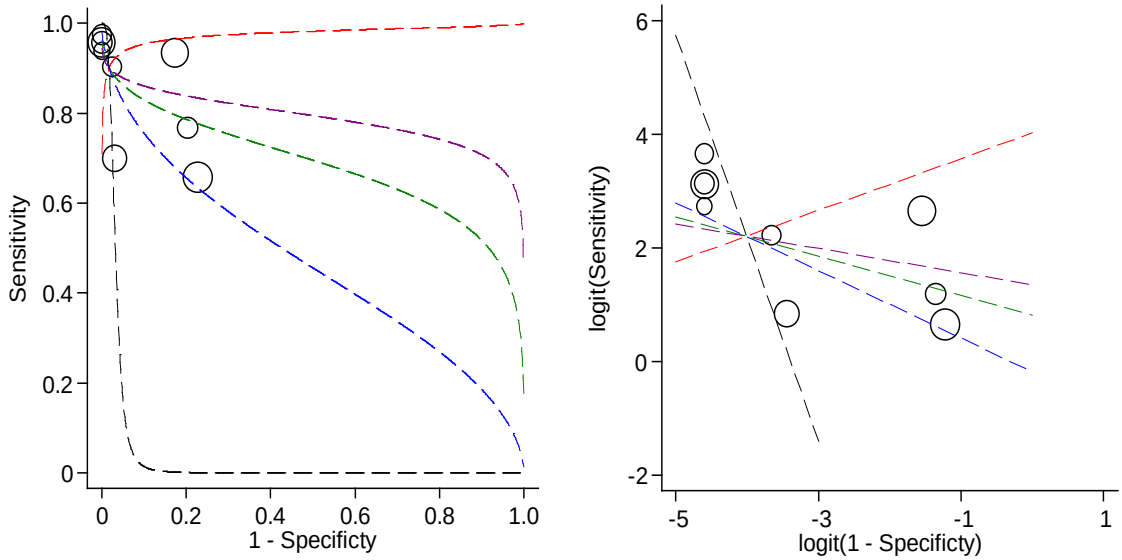
Table 4 shows that the regression of $\eta \sim \xi$, regression of $\xi \sim \eta$, and MAR of $\eta \sim \xi$ result in negative slopes in 32% of the meta-analyses we performed; the D~S model results in negative slopes in 13%. By design, the Rutter-Gatsonis model always produces a positive slope. **Figure 30** presents a matrix scatter plot of the slope values produced the different parameterizations of the HsROC curve, as well as those from the Moses-Litenberg method (weighted and unweighted).

Figure 28: HsROC curves for 24 randomly selected meta-analyses (alternative parameterizations of the HsROC curve)



Alternative ROC curves based bivariate meta-analysis (fit using maximum likelihood) for 24 randomly selected meta-analyses. The title for each panel contains the first author and year of publication for the corresponding review. Results are shown for the Rutter-Gatsonis (red lines), η~ξ (green lines), ξ~η (blue lines), D~S (purple lines), and MAR of η~ξ (black lines) parameterizations. See text and Appendix C for additional details. D = logit(sensitivity) – logit(1 – specificity); S = logit(sensitivity) + logit(1 – specificity); MAR = major axis regression of η on ξ; η = logit(sensitivity); ξ = logit(1-specificity).

Figure 29: Study results and fitted hsROC curves for an example dataset



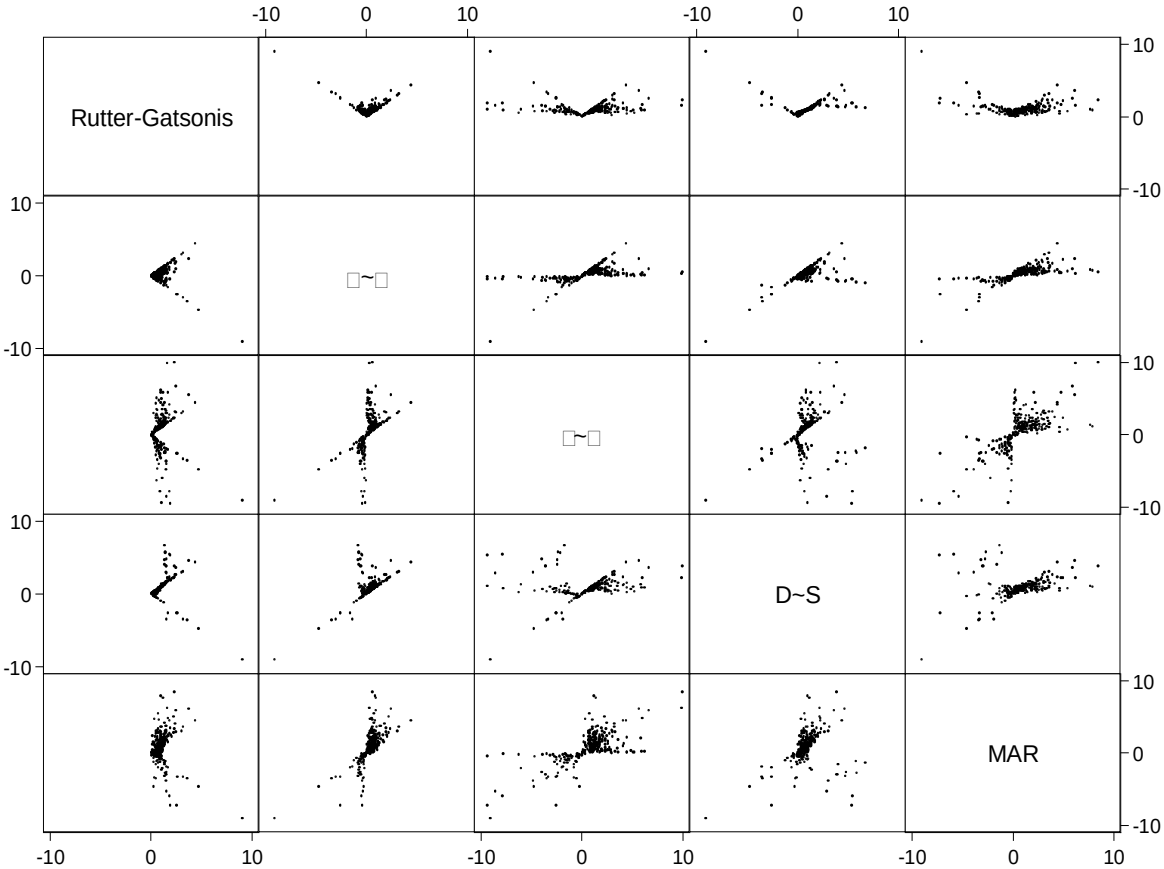
Alternative hsROC curves for the meta-analysis by vanGelder, 2003 (31) (fourth dataset on the top row of Figure 28). Results are shown for the Rutter-Gatsonis (red lines), η~ξ (green lines), ξ~η (blue lines), D~S (purple lines), and MAR of η~ξ (black lines) parameterizations, in the ROC space (left panel) and the logit-transformed ROC space (right panel). All models were estimated using REML. See text and Appendix C for additional details. For the logit space graph, in studies where the estimated false positive rate was 0 we used a value of 0.01 (because the logit of zero is undefined). D = logit(sensitivity) – logit(1 – specificity); hsROC = hierarchical receiver operating characteristic; S = logit(sensitivity) + logit(1 – specificity); MAR = major axis regression of η on ξ; REML = restricted maximum likelihood = logit(sensitivity); ξ = logit(1 – specificity).

Table 4: Slope of ROC line in the logit-space

Method	Meta-analyses with positive slope N (%)	Meta-analyses with negative slope N (%)
R & G	298 (100)	0 (0)
η ~ ξ	202 (68)	96 (32)
ξ ~ η	202 (68)	96 (32)
D ~ S	258 (87)	40 (13)
MAR	202 (68)	96 (32)

D = logit(sensitivity) – logit(1 – specificity); S = = logit(sensitivity) + logit(1 – specificity); MAR = major axis regression of logit(sensitivity) on logit(specificity); R & G = Rutter-Gatsonis model; η = logit-transformed sensitivity; ξ = logit-transformed specificity.

Figure 30: Scatter plot of the slopes of alternative HsROC lines (logit space)



D = logit(sensitivity) – logit(1 – specificity); MAR = major axis regression; S = logit(sensitivity) + logit(1 – specificity); HsROC = hierarchical summary receiver operating characteristic curve; η = logit(sensitivity); ξ = logit(1 – specificity). Examples that resulted in extreme slope values (>10 or <-10) are not show to avoid distortion of the graph.

Discussion

Key findings

We present a comprehensive empirical comparison of meta-analytic methods for studies of test accuracy, both in terms of number of meta-analyses included and in terms of the scope of the meta-analytic methods considered. Univariate and bivariate meta-analyses most often resulted in similar point estimates, regardless of the estimation method (inverse variance or REML) or the likelihood structure used to describe between-study variability (approximate normal or exact binomial). Use of a normal approximation (both in univariate and bivariate meta-analyses) resulted in summary estimates with lower values and led to too narrow confidence intervals, compared to methods that used the exact binomial likelihood. Bivariate models fit using Bayesian and maximum likelihood methods produced almost identical summary estimates of sensitivity and specificity. The credibility intervals produced by Bayesian bivariate meta-analysis methods were substantially wider compared to the confidence intervals of maximum

likelihood methods (using the exact binomial likelihood to describe within-study variability for both models). Although often not well estimated, the between-study correlation (of sensitivity and specificity) was frequently far from zero. This indicates that ignoring it is generally inappropriate for meta-analyses. Alternative meta-analytic methods to obtain sROC curves resulted in substantially different curves; differences were substantial between alternative parameterizations of the HsROC curves (particularly when the correlation between sensitivity and specificity was estimated to be positive).

Meta-analysis of sensitivity and specificity

Our findings substantially extend previous comparisons between methods for meta-analysis of test accuracy. **Table 5** summarizes selected empirical comparisons of meta-analytic methods for test accuracy, where at least one of the methods allowed for correlation of sensitivity and specificity at the between-study level. Generally, previous reports have assessed only few applied meta-analysis examples (ranging from 1 to 50 meta-analyses), whereas we analyzed a much larger database using a wide array of analytic approaches.

Previous theoretical and simulation studies have suggested that the binomial distribution may be preferable to the normal approximation for modeling within-study variability. We believe that our observations are in concordance with this position. Not unexpectedly, the differences between the two methods were more pronounced in studies of small sample size and meta-analyses where tests had high sensitivity and specificity. In such cases the normal distribution will be a poor approximation to the binomial. Furthermore, in studies where some of the counts are zero, analysis using the normal likelihood will require the use of a continuity correction (so that the variance and point estimate of the study-level logit-sensitivity or logit-specificity can be calculated). The continuity correction will bias the point estimate of individual studies; this is why the difference in the summary estimates between methods that rely on the normal approximation versus those that do not is greater when the summary sensitivity or specificity are closer to one (32). An additional reason for the systematically smaller summary sensitivity or specificity with normal approximation methods may be that in the meta-analysis, the estimate (logit-transformed sensitivity or specificity) and its variance are correlated, in that the variance is a function of the estimate and the sample size. This correlation is positive for proportions larger than 0.5, and thus estimates near one have larger variance (and receive less weight in the meta-analysis) compared to estimates near 0.5. The net effect is that summary sensitivity or specificity are biased towards 0.5 (25,30). Such a bias is not a problem for meta-analysis methods using the exact likelihood, and is not observed when variance stabilizing transformations are used for meta-analysis of proportions (such as the arcsine transformation) [reference to be provided post peer review].

Few studies have compared the results of bivariate meta-analysis using maximum likelihood versus fully Bayesian methods for the meta-analysis of sensitivity and specificity and those that did used models that were not directly comparable) (16,18,19). Many investigators have commented that Bayesian methods are less accessible to meta-analysts than the corresponding maximum likelihood methods. We provide the BUGS code we used to fit the bivariate model for the model in **Appendix B**. We found that convergence problems were not common when fitting the bivariate model; when present

they were mostly due to numerical instability in cases where the number of studies was small, sensitivity and specificity were close to 1, or the between-study variance was very low. For Bayesian analyses, we were able to obtain model convergence in most datasets by slightly modifying the non-informative prior distributions used. Bayesian analyses resulted in summary estimates of sensitivity and specificity that were very close to those obtained from the maximum likelihood estimation. However, there were substantial differences in the width of the credibility and confidence intervals produced by Bayesian and maximum likelihood analyses, respectively.

Bivariate methods provide estimates of the correlation between sensitivity and specificity at the between-study level. Alternative models (normal approximation for within-study variability versus exact binomial distribution) and estimation methods (non-iterative versus REML; frequentist versus Bayes) can yield quite different correlation estimates. This may be another symptom of the fact that the correlation parameter is generally poorly estimated. A telling observation is the following: frequentist approaches (maximum likelihood and inverse variance methods) often estimated the correlation parameter in the extremes of its domain, namely -1 (and sometimes +1). Riley 2007 (33) made a similar observation in a simulation study. In contrast, Bayesian methods rarely produced extreme correlation values, due to shrinkage toward the mean of the prior distribution (the mean is zero for the uniform(-1,1) prior distribution that we used).

Table 5: Summary of selected previous empirical comparisons of meta-analysis methods, including simulation studies

Author, year [reference]	Number of MAs	Methods compared and model fitting	Software	Key findings	Authors' recommendations/conclusions
Macaskill (16), 2004	1 (3 index tests)	HsROC [ML vs. fully Bayesian] vs. sROC (unweighted)	SAS	Estimates from ML analyses agree with Bayesian analyses; CIs from ML analysis were narrower than those from Bayesian analysis; the ROC curves for all analyses were similar for each index test.	ML formulation of HsROC model may be more accessible; Bayesian methods allow greater modeling flexibility; model checks are required for distributional assumptions of RE.
Reitsma (10), 2005	1 (3 index tests)	sROC vs. BREMA normal	SAS	Summary values of sensitivity and specificity at the Q-point for one of the index tests were very different from the pooled estimate produced by the bivariate model.	The bivariate model is an "improvement and extension" of the "traditional" sROC approach. Explanatory variables with separate effects on sensitivity and specificity can be added in the bivariate model.
Chu & Cole (20), 2006	1	BREMA normal vs. binomial	SAS	Results of the two methods were similar in the applied example; in a limited simulation study the exact binomial likelihood gave unbiased results whereas the normal likelihood produced biased results for Se, Sp and ρ, when Se and Sp are close to 1.	In sparse datasets the exact likelihood may be preferable.
Harbord (13), 2007	1	Bivariate binomial vs. HsROC	SAS	Parameter estimates of each model can be used to calculate those of the other; results are nearly identical. In the single example, the correlation between sensitivity and specificity was positive.	The bivariate and HsROC models are closely related and in common situations identical. The HsROC model generally allows additional modeling flexibility.
Riley (33), 2007	1	Bivariate normal vs. univariate normal; univariate binomial vs. bivariate binomial (the latter failed to converge)	SAS	Univariate and bivariate analyses with a normal approximation produced similar point estimates; the correlation was estimated as -1, leading to slightly larger between-study variances in the bivariate analyses. The univariate analysis using the exact binomial likelihood produced slightly higher estimates of sensitivity and specificity; the corresponding bivariate analysis failed to converge. Simulation results indicated that the between-study correlation is often estimated to be -1 (or, less commonly, +1) when the number of studies is small or the within-study variance is large (compared to the between-study variance); this leads to an upward bias in the estimates of between-study variance in the bivariate case. In simulated bivariate meta-analyses of 10 studies the model using the binomial likelihood failed to converge in 397/1000 iterations.	The BREMA using a normal likelihood is preferable to two normal UREMAs. For meta-analyses of proportions (e.g., sensitivity and specificity) the exact binomial likelihood is preferable compared to the BREMA using a normal likelihood or two separate UREMAs using the exact likelihood. The bivariate model with the exact likelihood may occasionally fail to converge, often due to difficulties in estimating the correlation between sensitivity and specificity.
Arends (25), 2008	2	BREMA normal vs. binomial; alternative HsROC curves derived from the bivariate model vs. the sROC method	SAS	Results using the approximate normal likelihood were similar to those using the exact binomial likelihood in one example (however the exact likelihood produced higher values of sensitivity and specificity). The alternative HsROC curves generated by the bivariate model differ	Multiple HsROC curves can be derived from the bivariate model. BREMA extends the sROC approach and provides a unifying framework for other approaches to the meta-analysis of diagnostic

		(Moses-Littenberg)		substantially between them and may differ from the Moses-Littenberg sROC.	tests.
Hamza (32), 2008	1	UREMA, normal vs. binomial	SAS	In separate UREMA, estimates from the approximate normal method were lower compared to those produced from the exact binomial method, both for sensitivity and specificity. In simulation, the exact likelihood always performed better than the approximate approach and gave unbiased estimates; the approximate method had large bias and poor coverage.	The exact binomial likelihood is the method of preference and should be used whenever feasible.
Hamza (34), 2008	1	Random intercept sROC vs. BREMA (normal and binomial)	SAS	In the data example, the parameter estimates from the three methods differed substantially and resulted in substantially different sROC curves. In simulations, the BREMA with the exact binomial likelihood gave unbiased estimates of the intercept and slope parameter of the sROC; coverage was also acceptable except when the number of studies was low. The random intercept sROC method and the BREMA with a normal approximation produced biased results and had poor coverage probabilities. Bivariate methods may fail to converge when the correlation of sensitivity and specificity is close to -1 or +1.	The BREMA with exact binomial likelihood for the within-study model performed better than other methods.
Harbord (21), 2008	8	Pooling vs. univariate RE MA vs. separate MA of LRs (not considered here) vs. sROC (weighted and unweighted) vs. BREMA/HsROC	SAS	In 6/8 examples, all methods gave similar point estimates; in 2/8 cases pooling gave different point estimates; CIs from pooling were "too narrow". In 5/8 examples, sROC curves from all methods (pooling not assessed) were similar; in 1 example the sROC results (weighted and unweighted) differed from other methods; in 1 example BREMA/HsROC, weighted sROC and separate RE MA produced different results from the unweighted sROC; in 1 example all methods produced similar results within the range of data but diverged outside that range.	HsROC or BREMA methods should be used as the standard; simple pooling should not be used to derive summary values of sensitivity and specificity; univariate RE MA may be used to give a valid estimate of the summary point alone; HsROC/BREMA appears to be the only way to obtain a valid sROC curve.
Chappell (26), 2009	4 examples of an algorithm for deciding the optimal analysis method	Alternative HsROC curves from the bivariate model; univariate vs. bivariate methods (fixed and random effects). BREMA estimates obtained from ML methods; interval estimates from the posterior distribution with seemingly non-informative priors for the hyperparameters through MCMC.	R	Multiple HsROC curves can be derived from the bivariate meta-analysis model; these curves can have substantially different shapes. Application of a proposed algorithm to guide the selection of the optimal meta-analysis model (bivariate versus univariate; fixed versus random) resulted in different choices (i.e. univariate meta-analyses were considered appropriate in some examples; bivariate in others). Parameters of the bivariate model were sometimes poorly estimated (particularly the between-study variances and the correlation of logit-sensitivity and specificity).	A zero or positive correlation between sensitivity and specificity "does not invalidate the bivariate model, as such", but means that "the data should not be summarized by an sROC curve". In some situations the sROC model may be inappropriate or there may not be enough data to estimate model parameters reliably. The authors propose an algorithm for the determination of the optimal analysis method for diagnostic test data.
Simel (22), 2009	2 (one with 3 index tests) "selected for highlighting the merits" of BREMA	BREMA normal vs. BREMA binomial vs. univariate analysis (fixed effects and RE)	SAS; Meta-Disc and CMA	In analyses selected for highlighting the merits" of BREMA: among pairwise comparisons between all methods the median difference was 1.5% (25th-75th perc.=1.0-2.2%, maximum=6.0%); in 5 unselected analyses differences in sensitivity were 0-6%; differences in specificity were 0-2%; across all 7 examples the median difference in posterior	The two approaches lead to relatively small differences in posterior probabilities; premature to insist that BREMA is the only way to get clinically useful results; more work needs to be done to solve the problem of non-convergence.

	+ 5 unselected			probability (assuming a prior probability of 0.5) between methods was 2.5% (25th-75th perc.=2.2–3.2% , maximum=11%).	
Paul (18), 2010	1	BREMA binomial [ML vs. Bayesian MCMC vs. Bayesian INLA vs. empirical Bayes INLA	R	In the applied example, point estimates and intervals around sensitivity, specificity and their variances were similar across methods. Bayesian methods and INLA with Empirical Bayes modeling produced similar point estimates and intervals around the correlation of sensitivity and specificity; the ML approach produced a different estimate of the correlation (-1) with very wide confidence intervals (-1 to 1). In a simulation study, INLA and ML methods produced similar results in terms of bias and mean squared error. INLA had better coverage; ML modeling underestimated the variance parameters whereas INLA produced less downwardly biased variance estimates and more, reliable estimates of the correlation.	INLA is more stable and gives generally better coverage probabilities for the pooled estimates and less biased estimates of variance parameters compared to ML modeling. INLA may be more user-friendly compared to full MCMC Bayesian modeling.
Menke (35), 2010 19936437	50	BREMA binomial vs. HsROC, both fixed and random effects	SAS	Estimates of BREMA and HsROC analyses were nearly identical; correlations between point estimates or SEs produced by the two methods were >0.99. Comparisons were not reported between random end fixed effects method; convergence was fast (1.4 seconds per MA).	Generalized linear random effects models are an alternative to the HsROC approach.
Verde (19), 2010	2	BREMA binomial using different specifications of the random effects distribution (binomial-normal vs. binomial-normal based on scale mixtures vs. binomial-t based on scale mixtures) and link functions (logit vs. c-log-log). Comparisons with the exact binomial BREMA model (from Chu & Cole 2006); the Rutter-Gatsonis HsROC model; and a Bayesian model accounting for disease prevalence were also reported for 1 of the examples.	R; Winbugs	The logit and c-log-log link functions produced similar results and comparable model fit. The binomial-normal and binomial-t (based on scale mixture of normals) models produced different results for the summary sensitivity and specificity (wider intervals for the bivariate-normal model) and their respective predictive distributions (wider intervals for the bivariate-t model with scale mixtures). The Bayesian bivariate-normal model produced similar results to the ML modeling. The Bayesian bivariate-t model with scale mixtures may offer better fit compared to the Bayesian bivariate-normal model with scale mixtures or the Bayesian bivariate-normal model.	Inference regarding random effects should be based on distributions more flexible than the normal. The predictive distribution of meta-analysis results reflects their future use. Model checking for meta-analysis results should not be ignored.

Unless otherwise stated analyses used ML methods. The studies by Macaskill (16) 2004; Reitsma (10) 2005; Chu & Cole (20) 2006; Harbord (13) 2007; Chappell 2009 (26); Simel (22) 2009; and Verde 2010 (19), considered a common meta-analysis example, based on a systematic review by Scheidler 1997 (36).

BREMA = bivariate random effects meta-analysis; CI = confidence interval; CMA = comprehensive meta-analysis; HsROC = hierarchical sROC; INLA = integrated nested Laplace approximations; LR = likelihood ratio; MA = meta-analysis; MCMC = Markov Chain Monte Carlo; ML = maximum likelihood; perc. = percentile; RE = random effect; Se = sensitivity; Sp = specificity; sROC = summary receiver operating characteristic curve; UREMA = univariate random effects meta-analyses.

Constructing meta-analytic ROC curves

Arguably, ROC curves provide additional information compared to meta-analytic estimates of sensitivity and specificity, because they illustrate the relationship between average sensitivity and average specificity. Based on our previous survey [citation to be provided post peer review] the most commonly used method for constructing sROC curves is the approach proposed by Moses and Littenberg. Despite its popularity this model has several shortcomings, including its failure to account for underlying binomial distribution of data, between-study heterogeneity, and measurement error on its independent variable. These shortcomings of the Moses-Littenberg sROC model are overcome by the hierarchical modeling approaches, including the increasingly used model proposed by Rutter and Gatsonis (12). It can been shown that the Rutter-Gatsonis HsROC model is equivalent to the bivariate meta-analysis of sensitivity and specificity, in the absence of covariates in the regression (13,25). Thus, the parameters of the HsROC model can be "back-calculated" using estimates from the bivariate meta-analysis model (an approach we followed in this report).

The Rutter-Gatsonis HsROC model is one of several possible parameterization of the HsROC curve. Arends 2008 (25) discuss alternative parameterizations, which we implemented for all meta-analyses we performed (plots available from the authors upon request). These parameterizations often result in substantially different curves compared to the one produced by the Rutter-Gatsonis HsROC model (12,26). Importantly, in some cases the slope of the ROC curve is not always positive (in contrast to the Rutter-Gatsonis method). This means that the relationship between the average sensitivity and specificity cannot be explained by threshold effects across studies. Based on this Chappell 2009 (26) discuss that sROC curves are not always a helpful summary of the data, and propose a stepwise algorithm for determining the most appropriate approach to summarize accuracy studies.

Limitations

Some limitations need to be considered when interpreting our results. Because of the way we constructed the database of systematic reviews of test accuracy, all included meta-analyses were conducted prior to 2003 and were published in English-language journals. Although this may limit the clinical applicability of their actual findings, it does not substantially affect the conclusions of our empirical comparison of methods because the datasets included are very diverse in terms of number of included studies, sample size, and reported test accuracy (**Table 1**). In a recent comprehensive review of reporting and design characteristics of systematic reviews of test accuracy that gave quantitative synthesis results (covering 1983 to 2009), we found no substantial change over time in the number of included studies or the number of meta-analyses conducted per review article [citation to be provided post peer review].

Another limitation of our work is that many systematic reviews contributed multiple datasets to the empirical comparison (approximately 2 datasets per review, on average). We believe that the effect of this clustering is probably minor because in most cases when multiple meta-analyses are presented in the same systematic review, they typically address different index or reference standard tests (often based on non-overlapping sets of primary studies).

Finally we have focused on meta-analysis of sensitivity, specificity and meta-analytic ROC curves, but did not consider other metrics such as likelihood ratios or areas under the ROC curve. We note that these metrics can be derived from the methods we assess (for example, likelihood ratios can be estimated from the output of the bivariate model) and are generally less commonly used in the diagnostic literature.

Conclusions

This work represents the most comprehensive empirical comparison of meta-analytic methods for studies of test accuracy, both in terms of the included number of meta-analyses and the scope of the meta-analytic methods considered. Based on our empirical observations and a review of the relevant literature, we summarize key findings relevant to meta-analytic practice in **Box 1**.

Box 1: Summary of findings relevant to meta-analytic practice

Bivariate versus univariate analyses

In our empirical comparison, bivariate meta-analyses produced point estimates that were largely similar to those of separate univariate analyses (also observed elsewhere(34)). Because bivariate methods account for the correlation between the sensitivity and specificity across studies, the confidence region around the summary point is different from the univariate analyses, and the same is true for predictive distributions for future studies. Our findings suggest that this correlation is generally poorly estimated; however, bivariate models have stronger theoretical motivation for most common diagnostic test meta-analysis scenarios.

Approximate normal versus exact binomial distribution for modeling within-study variability

Based on large sample theory, the normal approximation is inadequate when the sample size of included studies is small and the sensitivity or specificity of tests is extreme (very high or very low). We found that continuity corrections (required for the normal approximation) introduced bias in meta-analytic estimates. This is consistent with simulation studies suggesting that meta-analysis using the exact binomial likelihood outperforms methods relying on the normal approximation (32). The normal approximation could be reserved for cases where the model using the exact likelihood cannot be fit (e.g., inability to converge), or there is no access to statistical software able to fit generalized linear mixed models.

Maximum likelihood versus Bayesian methods

Bayesian methods are theoretically appealing and allow for more flexible modeling, particularly when complex data structures arise. Further, they allow use of external information in the form of informative prior distributions. In our empirical assessment point estimates of sensitivity and specificity produced by the two methods were very similar; however, Bayesian methods often resulted in credibility intervals that were wider compared to the confidence intervals of maximum likelihood methods. This reflects the Bayesian models' ability to model the uncertainty in the estimation of variance parameters more completely.

Bivariate and HsROC models (summary point versus summary line to synthesize data)

Meta-analyses of sensitivity and specificity aim to provide helpful summaries of the findings of individual studies. Sometimes a helpful way to summarize individual studies is to provide one "summary point" of combined average sensitivity and specificity For example, a summary point is helpful when the results of the studies are relatively similar, and when the studied tests do not have different explicit thresholds for positive results. Other times, it is more helpful to synthesize data using a "summary line" that describes how the average sensitivity changes with the average specificity. For example, a summary line may be a more helpful way to synthesize data when studies have different explicit thresholds and their results range widely. Choosing the most helpful summary is subjective and case dependent, and both summaries can be reasonably employed as they provide complementary information.

Choosing between alternative sROC curves

We found that alternative parameterizations of the sROC curve derived from the bivariate model can occasionally result in curves of different shape. Specifically, some parameterizations can result in negative estimated slopes when the correlation between sensitivity and specificity is positive (i.e., when the correlation between sensitivity and false positive rate is negative) (25,26). In such cases the relationship between the average sensitivity and specificity cannot be explained by varying thresholds for positive test results across studies. Some authors argue that such sROC curves are not a helpful summary of the data (34).

Standard models are not always appropriate

The standard bivariate/HsROC models will not be appropriate for all diagnostic settings, for example when test results are reported for multiple thresholds within each study or when the classification problem is not binary. In such cases more complex modeling approaches are necessary to obtain correct estimates of test accuracy (37-39).

Abbreviations

AHRQ	Agency for Healthcare Research and Quality
BREMA	bivariate random effects meta-analysis
EPC	Evidence-based Practice Center
HsROC	hierarchical summary receiver operating characteristic
MAR	major axis regression
REML	restricted maximum likelihood
ROC	receiver operating characteristic
sROC	summary receiver operating characteristic

References

(1) Miettinen OS. The modern scientific physician: 4. The useful property of a diagnostic. CMAJ. 2001;165:910-911.

(2) 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;144:850-855.

(3) 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;142:1048-55.

(4) Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol. 2003;56:1129-35.

(5) Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. AJR Am J Roentgenol. 2006;187:271-81.

(6) Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. Ann Intern Med. 1997;127:820-826.

(7) Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. Lancet. 1998;351:123-27.

(8) 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.

(9) Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making. 1993;13:313-21.

(10) Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58:982-90.

(11) Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. Health Technol Assess. 2005;9:1-113, iii.

(12) Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001;20:2865-84.

(13) Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007;8:239-51.

(14) van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med. 2002;21:589-624.

(15)

(16) Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. J Clin Epidemiol. 2004;57:925-32.

(17) Willis BH, Quigley M. Uptake of newer methodological developments and the deployment of meta-analysis in diagnostic test research: a systematic review. BMC Med Res Methodol. 2011;11:27.

(18) Paul M, Riebler A, Bachmann LM, Rue H, Held L. Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested Laplace approximations. Stat Med. 2010;29:1325-39.

(19) Verde PE. Meta-analysis of diagnostic test data: a bivariate Bayesian modeling approach. Stat Med. 2010;29:3088-102.

(20) Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. J Clin Epidemiol. 2006;59:1331-32.

(21) Harbord RM, Whiting P, Sterne JA, Egger M, Deeks JJ, Shang A et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. J Clin Epidemiol. 2008;61:1095-103.

(22) Simel DL, Bossuyt PM. Differences between univariate and bivariate models for summarizing diagnostic accuracy may not be large. J Clin Epidemiol. 2009;62:1292-300.

(23) DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7:177-88.

(24) Jackson D, White IR, Thompson SG. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. Stat Med. 2010;29:1282-97.

(25) Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MG, Stijnen T. Bivariate random effects meta-analysis of ROC curves. Med Decis Making. 2008;28:621-38.

(26) Chappell FM, Raab GM, Wardlaw JM. When are summary ROC curves appropriate for diagnostic meta-analyses? Stat Med. 2009;28:2653-68.

(27) Scheffe H. A Method for Judging all Contrasts in the Analysis of Variance. Biometrika. 1953;40:87-104.

(28) Cox NJ. Graphing agreement and disagreement. The Stata Journal. 2004;4:329-49.

(29) Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing. 2011;10:325-37.

(30) Chang BH, Waternaux C, Lipsitz S. Meta-analysis of binary data: which within study variance estimate to use? Stat Med. 2001;20:1947-56.

(31) van Gelder JM. Computed tomographic angiography for detecting cerebral aneurysms: implications of aneurysm size distribution for the sensitivity, specificity, and likelihood ratios. Neurosurgery. 2003;53:597-605.

(32) Hamza TH, van Houwelingen HC, Stijnen T. The binomial distribution of meta-analysis was preferred to model within-study variability. J Clin Epidemiol. 2008;61:41-51.

(33) Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. BMC Med Res Methodol. 2007;7:3.:3.

(34) Hamza TH, Reitsma JB, Stijnen T. Meta-analysis of diagnostic studies: a comparison of random intercept, normal-normal, and binomial-normal bivariate summary ROC approaches. Med Decis Making. 2008;28:639-49.

(35) Menke J. Bivariate random-effects meta-analysis of sensitivity and specificity with SAS PROC GLIMMIX. Methods Inf Med. 2010;49:54.

(36) Scheidler J, Hricak H, Yu KK, Subak L, Segal MR. Radiological evaluation of lymph node metastases in patients with cervical cancer. A meta-analysis. JAMA. 1997;278:1096-101.

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

(38) Bipat S, Zwinderman AH, Bossuyt PM, Stoker J. Multivariate random-effects approach: for meta-analysis of cancer staging studies. Acad Radiol. 2007;14:974-84.

(39) Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. Biometrics. 2003;59:936-46.