Methods Research Report

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

Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • **www.ahrq.gov**

Methods Research Report

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

Contract No. 290-2007-10055-I

Prepared by:
Tufts Evidence-based Practice Center, Tufts Medical Center
Boston, MA

Investigators:
Issa J. Dahabreh, M.D., M.S.[*]
Thomas A. Trikalinos, M.D.[*]
Joseph Lau, M.D.
Christopher Schmid, Ph.D.[*]

[*] Currently at: Center for Evidence-based Medicine, Program in Public Health, Brown University, Providence, RI

AHRQ Publication No. 12(13)-EHC136-EF
November 2012

This report is based on research conducted by the Tufts Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10055-I). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, 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 health care decisionmakers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for 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.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact EffectiveHealthCare@ahrq.hhs.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Suggested citation: Dahabreh IJ, Trikalinos TA, Lau J, Schmid C. An Empirical Assessment of Bivariate Methods for Meta-Analysis of Test Accuracy. Methods Research Report. (Prepared by Tufts Evidence-based Practice Center under Contract No. 290-2007-10055-I.) AHRQ Publication No 12(13)-EHC136-EF. Rockville, MD: Agency for Healthcare Research and Quality. November 2012. www.effectivehealthcare.ahrq.gov/reports/final/cfm.

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 email 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

Figures

Appendixes

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

Structured Abstract

Background. Meta-analyses of sensitivity and specificity pairs reported from diagnostic test accuracy studies employ a variety of statistical models for estimating mean performance and performance across different test thresholds. The impact of these alternative models on conclusions in applied settings has not been studied systematically.

Methods. We constructed a database of PubMed-indexed meta-analyses (1987–2003) 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 maximum likelihood (ML; 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 summary receiver operating characteristic (SROC) curves using the Moses-Littenberg fixed effects method (weighted and unweighted) and the Rutter-Gatsonis hierarchical SROC (HSROC) method. We also obtained alternative SROC 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 reanalyzed 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 ML) 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, ML with normal approximation, or ML with the exact binomial likelihood for estimation). Fitting the bivariate model using ML 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 imprecisely 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. Our results are generally in agreement with published simulation studies and the theoretically expected behavior of meta-analytic estimators.

Conclusion. 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 may be particularly

useful for parameters that are poorly estimated in the bivariate model. Alternative SROC curves provide useful global summaries of test performance.

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 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 individuals with disease as "test positives" (sensitivity) and individuals with no disease as "test negatives" (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,[12] 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 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.

---

[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.

Although it allows for a between-study dependency between 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.[10,13-15] 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[16,17] and are becoming increasingly popular in meta-analytic practice.[18] These implementations are based on maximization of likelihoods, but Bayesian methods have also been proposed.[13,19,20] Limited empirical work suggests that these approaches yield similar conclusions in applied meta-analysis examples.[17] 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 comparisons (typically based on a few meta-analysis examples).[10,21,22] Some methodologists have suggested that "hierarchical models are necessary,"[22] and others have conjectured that "differences between univariate and bivariate models […] may not be large."[23]

This report is the second in a series of three on meta-analysis of test accuracy, conducted by the Tufts 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 addresses the following aims by using a previously established database of meta-analytic datasets:

- Compare univariate (one outcome at a time) and bivariate (joint analysis of two outcomes) methods for meta-analysis of sensitivity and specificity.
- Compare inverse variance (DerSimonian-Laird), maximum likelihood (ML) and Bayesian methods for random effects meta-analysis of sensitivity and specificity.
- Compare methods using a normal approximation versus those using the exact binomial likelihood for meta-analysis of sensitivity and specificity.
- 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 database of PubMed-indexed English-language meta-analyses of test accuracy (published between 1987 and 2003) to identify those that reported adequate information to reconstruct the 2×2 cross-classification of diagnoses and test results of included primary studies.[12] This increased the efficiency of the data extraction process by avoiding the need 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 a previous AHRQ report produced by the Tufts EPC.[12]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. The point of average sensitivity and specificity is a useful summary when the results of the studies are relatively similar and the studied tests do not have different explicit thresholds for positive results. When studies have different explicit thresholds and their results range widely, though, a "summary line" that describes how sensitivity changes with specificity across studies may be a more informative description. In many cases, 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 meta-analyses of sensitivity and specificity to synthesize logit-transformed sensitivity and specificity values separately (i.e., ignoring their correlation) using fixed and random effects models weighting studies by the inverse of their sampling variance on the logit scale and assuming a normal sampling distribution. Between-study heterogeneity in random effects models was estimated using a non-iterative method (the DerSimonian-Laird moments-based estimator)[24] and with an iterative method (restricted maximum likelihood; REML). We used a continuity correction of 0.5 for studies where the observed count was zero for any of the cells of the 2×2 table; the correction was applied to all four cells of such studies. 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 maximum likelihood (ML) maximization.

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 to describe variation at the within-study level. 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) based on cell counts.

At the between-study level, both variants assumed that the (true) logit-transformed sensitivities and specificities followed a bivariate normal distribution centered at their summary estimates and with a covariance matrix that represented 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[e] (a multivariate generalization of the DerSimonian-Laird model),[25] as well as iteratively by REML.[15] For the analysis using the exact binomial likelihood we used random effects logistic regression fit with ML.[10,21] Because the log likelihood for this model has no closed form, it was approximated by numerical methods (adaptive Gaussian quadrature).[26,27]

The bivariate model with the exact binomial likelihood for within-study variability was also fit using a fully Bayesian approach, using non-informative prior distributions for parameters unspecified in the model. The model structure used for our analyses is presented in Appendix B. For each meta-analysis we ran 20,000 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 20,000). 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% central 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. For parsimony, in the remainder of the report we refer to methods for model fitting using restricted (REML) or unrestricted maximum likelihood (ML) as "maximum likelihood estimation" (MLE) methods.

[e] We refer to this method as "multivariate DerSimonian-Laird."

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 (ML)	• Random effects (ML); likelihood maximized by adaptive Gaussian quadrature
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; ML = maximum likelihood; REML = restricted maximum likelihood.

Meta-Analytic SROC Curves (Summary Lines)

SROC curves depict graphically the relationship between sensitivity and specificity and provide a visual summary of overall test performance. These SROC curves summarize the trade-off between sensitivity and specificity *across* studies; they are distinct from ROC curves obtained in individual studies by varying the threshold over a continuous measurement or a predicted probability. 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 (by inverse of the variance of the diagnostic odds ratio) regressions for this method and compared results to the more theoretically motivated hierarchical regression methods (see below).

Hierarchical SROC (HSROC)

Rutter and Gatsonis proposed HSROC meta-analysis methods to address the limitations of the Moses-Littenberg SROC approach.[13] As noted by Arends 2008[28] several alternative parameterizations of the HSROC curve, in addition to the Rutter-Gatsonis model, can be produced from the bivariate meta-analysis model.[13] These models 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, using methods described in the literature.[28,29] Briefly, we estimated the intercept and slope of the ROC line (in logit space) based on: (a) regression of logit-sensitivity on logit-false positive rate; (b) regression of logit-false positive rate on logit-sensitivity; (c) regression of the difference of logit-sensitivity and logit-false positive rate on their sum); and (d) 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[f] with scatter plots and histograms.[31] 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). Most meta-analyses aim to generalize beyond the observed studies, and thus random effects models are appropriate. For this reason, we focus on random effects models in this report.

Factors Associated With the Magnitude of Differences Between Methods

We hypothesized that the following meta-analysis level factors might be associated with differences in the summary estimates obtained from different methods: (1) the total number of included studies; (2) the median number of affected (for sensitivity) or unaffected (for specificity) participants (across studies in the meta-analysis); (3) the total number of studies in a meta-analysis where a cell of the 2×2 table (true positive or false negative for sensitivity; false positive or true negative for specificity) was zero.

We assessed whether these factors were associated with the magnitude of the difference between the following pairs of random effects meta-analyses: (1) univariate meta-analysis using REML with the normal within-study likelihood versus univariate meta-analysis using ML with the exact binomial likelihood; (2) bivariate meta-analysis using REML with normal within-study likelihood versus bivariate meta-analysis using ML with binomial within-study likelihood; (3) univariate meta-analysis using ML versus bivariate meta-analysis using ML, both with binomial within-study likelihood; (4) bivariate meta-analysis using ML versus bivariate meta-analysis using Bayesian methods, both with binomial within-study likelihood.

All comparisons were made on the logit scale, separately for estimates of sensitivity and specificity. The association between the factors of interest and the difference in estimates were evaluated using pairwise scatter plots and Spearman correlations (with associated p-values).

Software

All non-Bayesian meta-analyses were performed in Stata/IC (version 12; Stata Corp., College Station, TX). Bayesian analyses were implemented in WinBugs[32] (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.

[f] For all bivariate analyses we obtained non-simultaneous confidence intervals for estimates of sensitivity and specificity. One could opt to obtain simultaneous confidence intervals[30] 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.

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 percent 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 (25^{th}–75^{th} 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 unaffected individuals across meta-analyses = 62). Additional characteristics of the included meta-analyses and their component studies are presented in Table 2. The data summarized in the table highlight the diversity of sensitivity and specificity values encountered in our dataset.

Table 2. Descriptive characteristics of test accuracy meta-analyses

Characteristics	Median	25^{th} Percentile	75^{th} Percentile	Minimum	Maximum
Number of included primary studies	11	8	18	6	61
Median number of affected individuals (cases)*	30	17	47	2	468
Median number of unaffected individuals (controls)*	61	29	106	8	9979
Median ratio of affected (cases) to unaffected (control) individuals	0.46	0.21	1.04	0.002	15.14
Median number of "positive" test results	32	19	61	3	549
Median number of "negative" test results	59	30	99	7	9415
Median ratio of "positive" to "negative" test results	0.51	0.30	1	0.01	7.17
Median true positive count	21	12	35	2	464
Median false positive count	7	3	19	0	491
Median false negative count	6	2	12	0	157
Median true negative count	48	24	85	5	9,411
"Crude"† sensitivity	0.77	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).

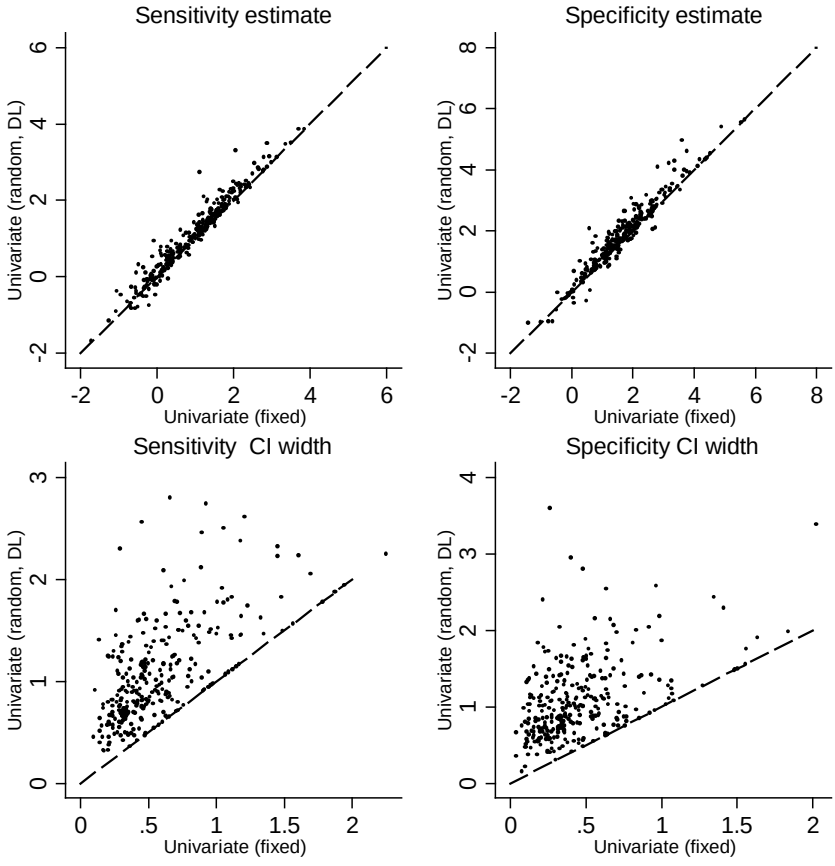
Some investigators have suggested that problems with convergence are a major concern when fitting the bivariate models for meta-analysis of test accuracy.[23] In total, for 10 of 308 meta-analyses (3%) we could not obtain estimates from all methods listed in Table 1 and they have been excluded from the comparisons presented in the following sections.[g]

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 greater for random effect analyses because they incorporate 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 sensitivity and specificity (logit scale; univariate DL random effects vs. 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.

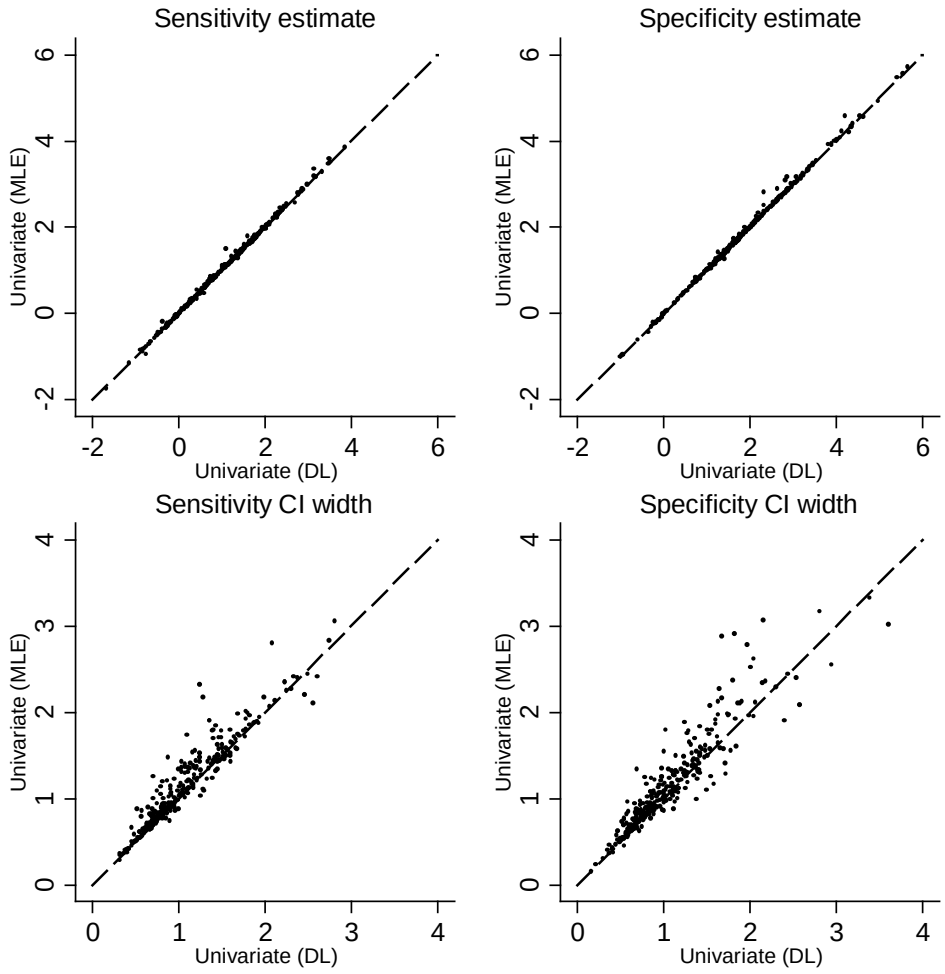
Univariate Random Effects Meta-Analysis Methods

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

Inverse Variance Versus ML Univariate Random Effects Meta-Analyses (Normal Approximation)

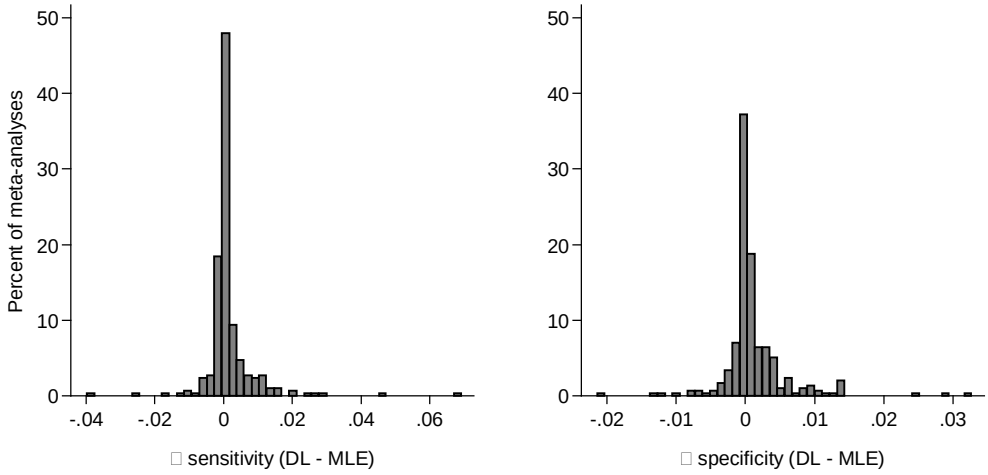
Figure 2 compares the point estimates (logit-transformed) and confidence interval widths for univariate meta-analyses of sensitivity and specificity using inverse variance versus ML random effects models based on the assumption of normal distributions for the logit-transformed probabilities. The point 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 a histogram of the differences in estimated sensitivities and specificities (untransformed scale) between methods. The absolute differences are rarely greater than 2.5 percent.

Figure 2. Comparison of point estimates and standard errors of summary sensitivity and specificity (logit scale) from random effects meta-analyses using a normal approximation (estimation of heterogeneity with DL vs. MLE)



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 MLE. CI = confidence interval; DL = DerSimonian-Laird; MLE = maximum likelihood estimation.

Figure 3. Histograms of differences in estimated summary sensitivity and specificity from univariate random effects meta-analyses (DerSimonian-Laird vs. MLE)



Histograms of differences in estimated summary sensitivity (left panel) and specificity (right panel) comparing univariate random effects meta-analyses using the DerSimonian-Laird inverse variance method versus MLE (both using a normal approximation for within-study variability). DL = DerSimonian-Laird; MLE = maximum likelihood estimation.

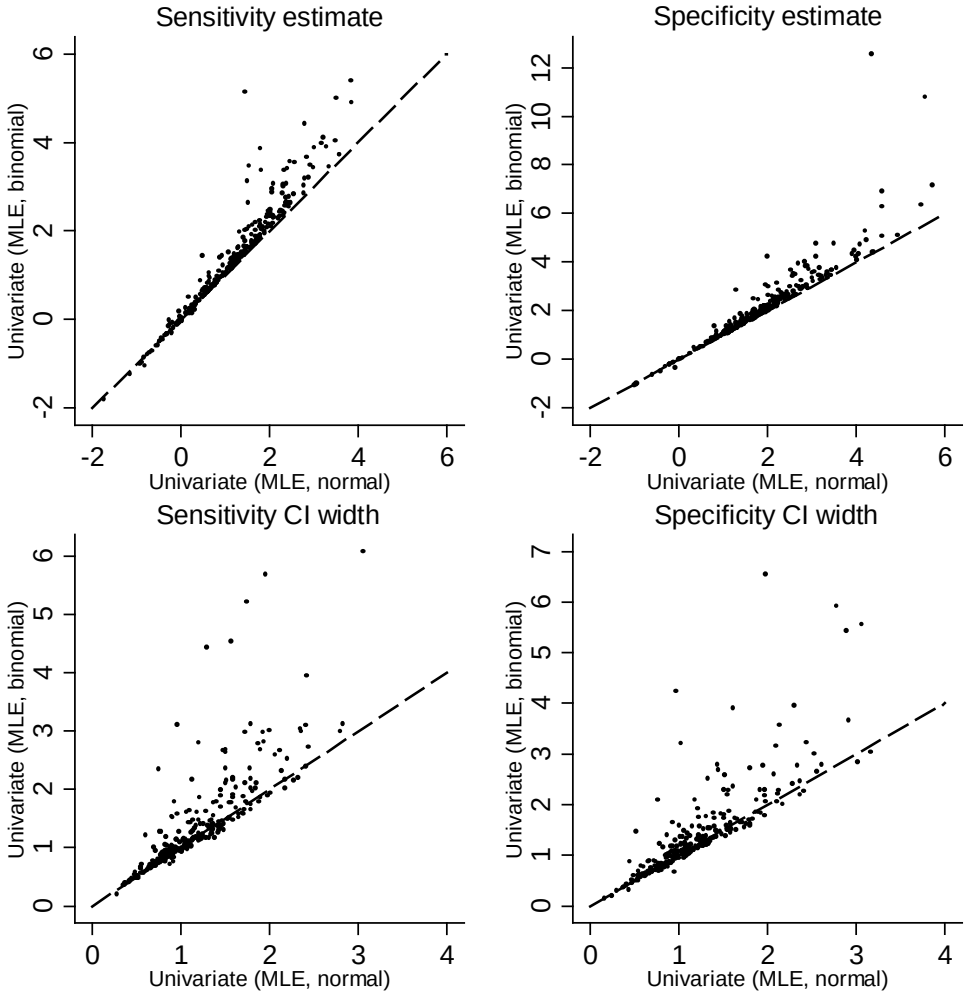
Approximate Normal Versus Exact Binomial Univariate Random-Effects Meta-Analyses (MLE)

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 MLE). 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. The need to add 0.5 (or any other constant) to each zero cell biases the estimated proportions towards lower values (towards 0.5). An additional reason may be that the estimates and variances are correlated;[33] the variance is a function of the estimate and the sample size. Indeed, the variance is a function of the proportion in such a way that it gets larger as the proportion approaches the extremes (zero or one). Thus, proportion estimates near the extremes receive less weight in the meta-analysis compared to estimates near 0.5. The net effect is that summary sensitivity or specificity is 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

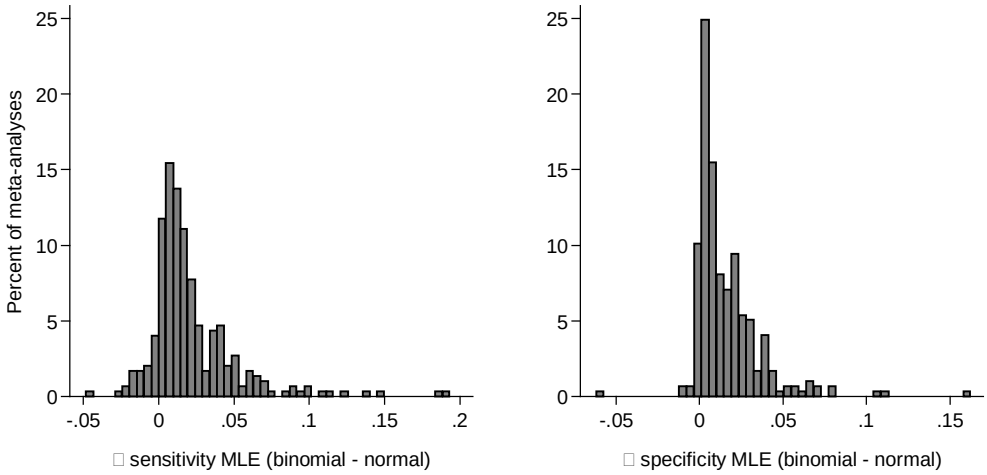
summary estimates. (This phenomenon has been described in several simulation analyses that we summarize in our Discussion section). Figure 5 presents histograms of differences in estimated sensitivities and specificities (untransformed scale) between methods. It is not uncommon to have differences in summary sensitivity and specificity of 5 percent or higher. As mentioned, summary sensitivity and specificity obtained with the normal approximation tend to be smaller than those obtained from exact methods.

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



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; MLE = maximum likelihood estimation.

Figure 5. Histograms of differences in estimated summary sensitivity and specificity (univariate random effects meta-analyses using the exact binomial likelihood vs. a normal approximation; both models fit with MLE)

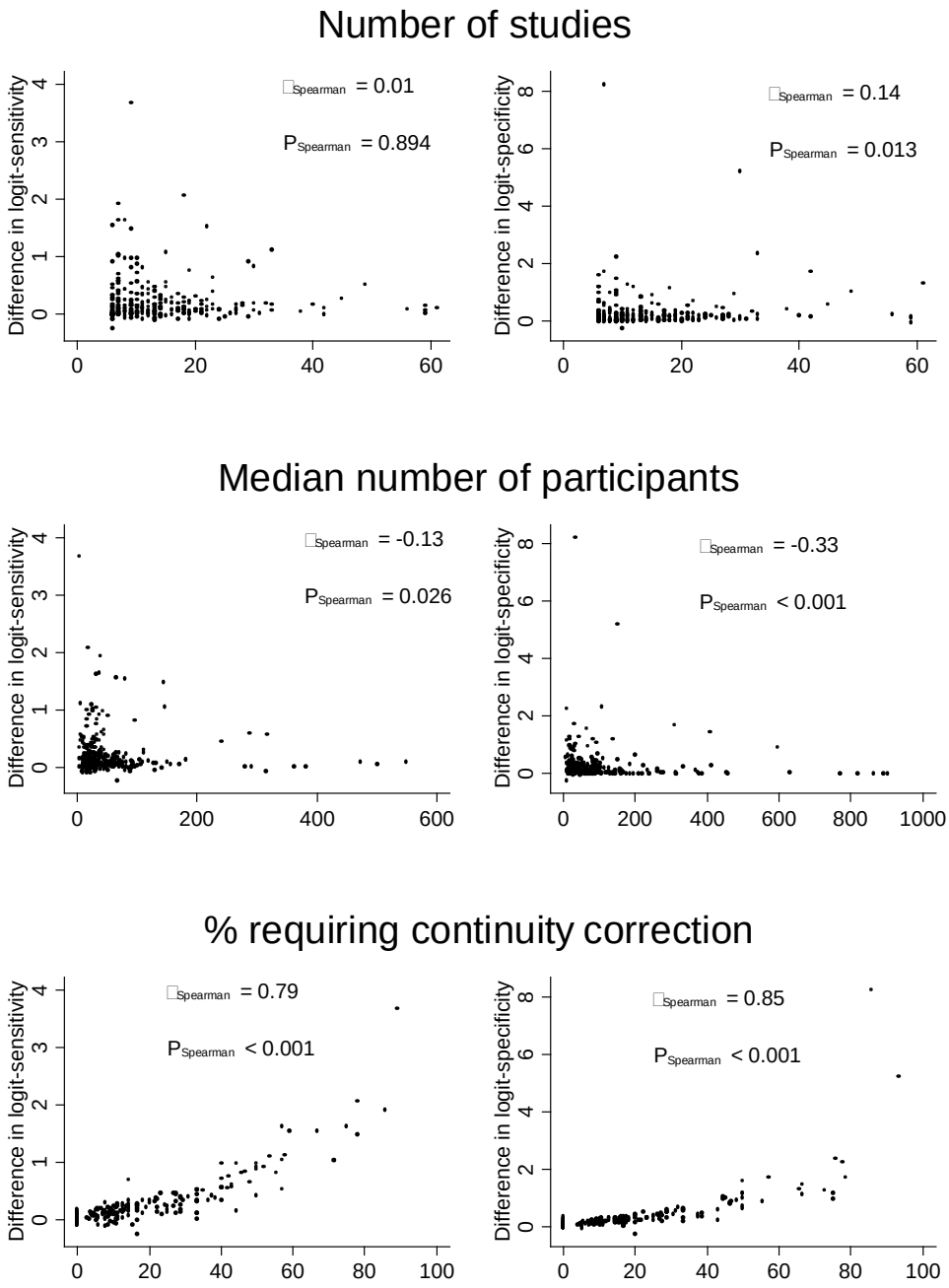


Histograms of differences in estimated summary sensitivity (left panel) and specificity (right panel) comparing univariate random effects meta-analyses using the exact binomial likelihood versus using a normal approximation for within-study variability (both models fit with MLE). MLE = maximum likelihood estimation.

Factors That Influence the Difference Between Univariate Random Effects Meta-Analyses Using the Normal Versus Binomial Likelihood (both with MLE)

Figure 6 presents scatter plots of the differences (logit-transformed) of estimated sensitivity and specificity over factors that we hypothesized could affect results with various methods. Differences between methods were larger in meta-analyses where a large proportion of the available studies required a continuity correction (for the normal approximation) and in meta-analyses where studies were generally small. The total number of included studies in the meta-analysis generally had a smaller effect.

Figure 6. Differences of estimated sensitivity and specificity (logit scale, univariate random effects meta-analyses using the exact binomial likelihood vs. a normal approximation; both models fit with MLE) over meta-analysis characteristics

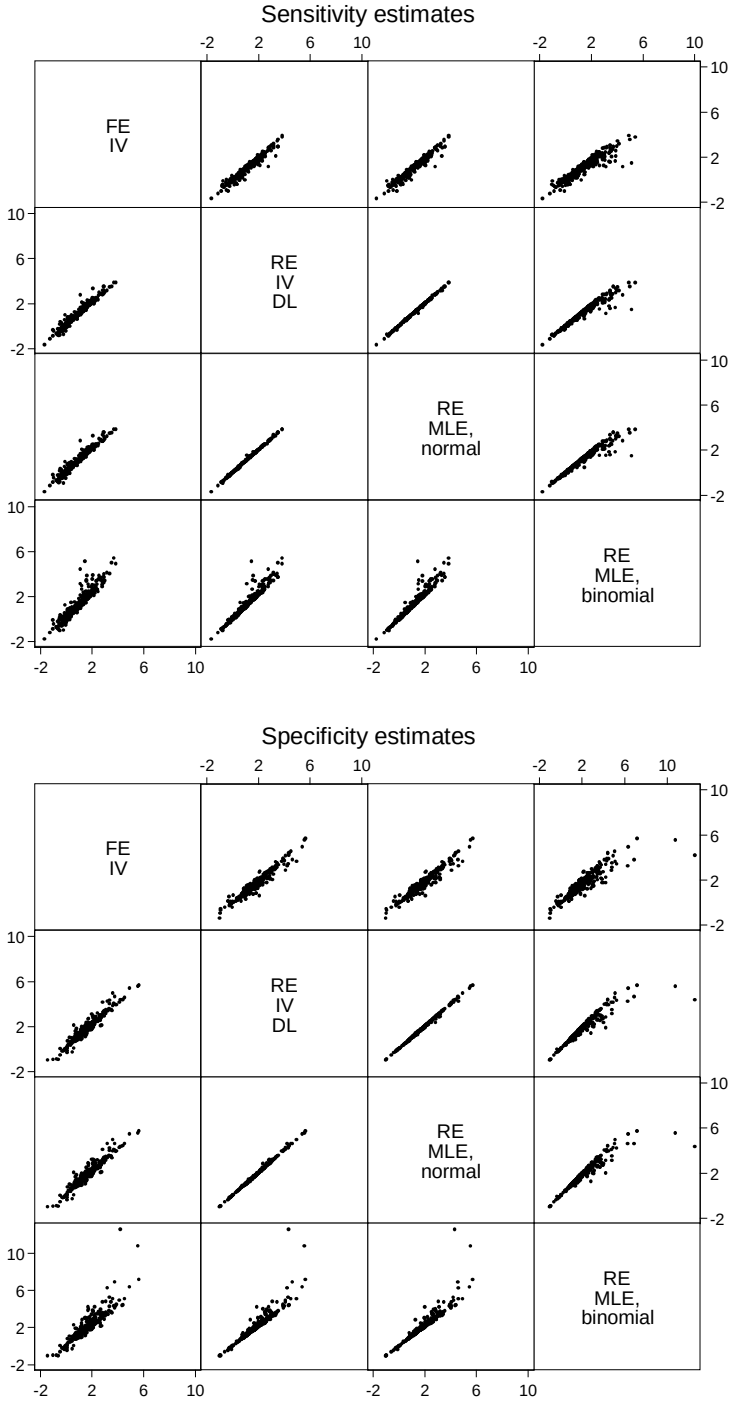


Note: Positive differences indicate that estimates from analyses using the binomial likelihood are larger than those from analyses using the normal approximation. Eight meta-analyses with a median number of unaffected individuals >1000 have not been plotted (in the middle right panel) to avoid distortion of the graph.

Summary for Univariate Meta-Analysis Methods

Figure 7 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 MLE with a normal approximation; random effects using ML with the 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 7. Summary comparison of sensitivity and specificity estimates (logit scale) from all univariate methods considered in this report



Note: 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; MLE = maximum likelihood estimation.

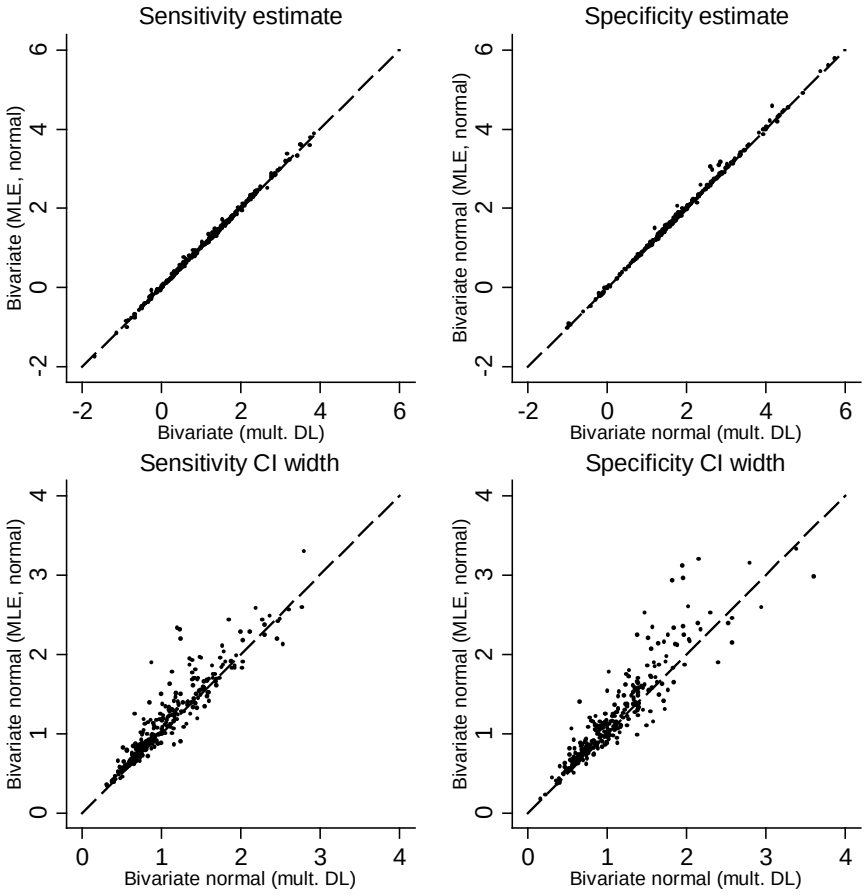
Bivariate Meta-Analysis Methods

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

Noniterative (Inverse Variance) Versus Iterative (MLE) Estimation of Between-Study Variability in Bivariate Random Effects Meta-Analyses (Normal Approximation)

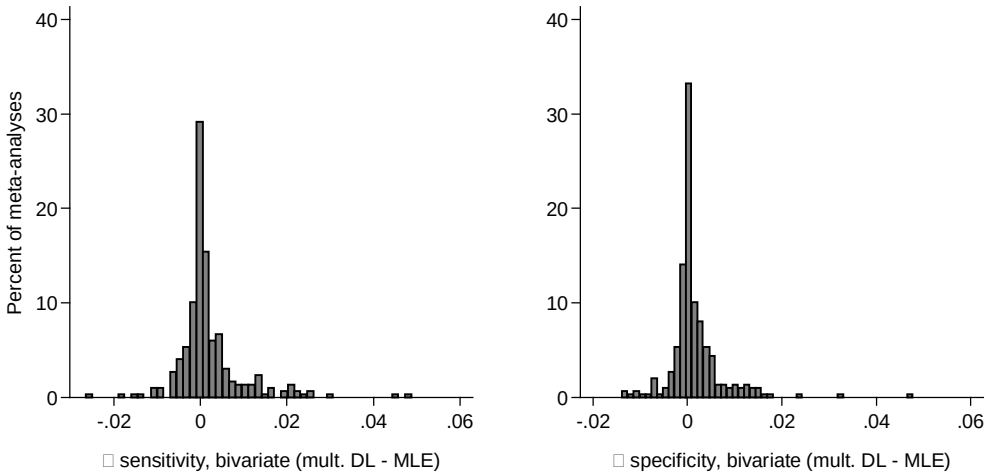
Figure 8 compares the point estimates (logit-transformed) and confidence interval widths from bivariate random effects meta-analyses of sensitivity and specificity using a normal approximation to model within-study variability. We compare models that use a non-iterative estimator of between-study variance (a generalization of the DerSimonian-Laird method) versus an iterative estimator obtained with REML.[25] Point estimates from the two methods were almost identical. However, MLE 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 9 presents histograms of the differences in estimated sensitivities and specificities (untransformed scale) between the two methods. Differences in summary estimates higher than 2.5 percent are uncommon.

Figure 8. Comparison of point estimates and confidence interval widths of summary sensitivity and specificity (logit scale) from bivariate random effects methods using a normal approximation (multivariate DerSimonian-Laird inverse variance vs. MLE)



Note: Scatter plot of estimated logit-transformed sensitivity, specificity and their corresponding confidence interval widths from bivariate random effects meta-analyses using the multivariate DerSimonian-Laird and MLE methods (both with a normal approximation to represent within-study variability).
CI = confidence interval width; mult. DL = multivariate DerSimonian-Laird; MLE = maximum likelihood estimation.

Figure 9. Histograms of differences in estimated summary sensitivity and specificity from bivariate random effects meta-analyses with multivariate DerSimonian-Laird inverse variance versus MLE (both models using a normal approximation to represent within-study variability)

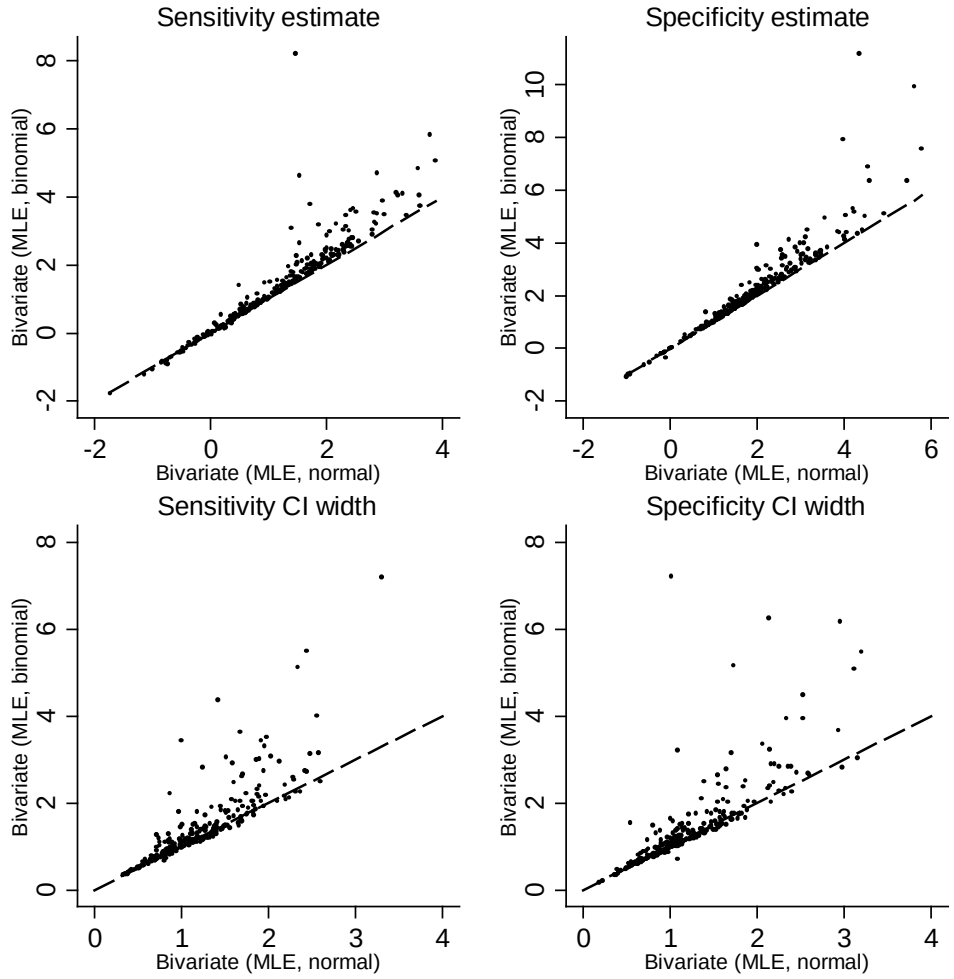


Note: Histograms of differences in estimated summary sensitivity (left panel) and specificity (right panel) comparing bivariate random effects meta-analysis using a normal approximation for 2 estimation methods: multivariate DerSimonian-Laird and MLE. Mult. DL = multivariate DerSimonian-Laird; MLE = maximum likelihood estimation.

Approximate Normal Versus Exact Binomial Bivariate Random-Effects Meta-Analyses (MLE)

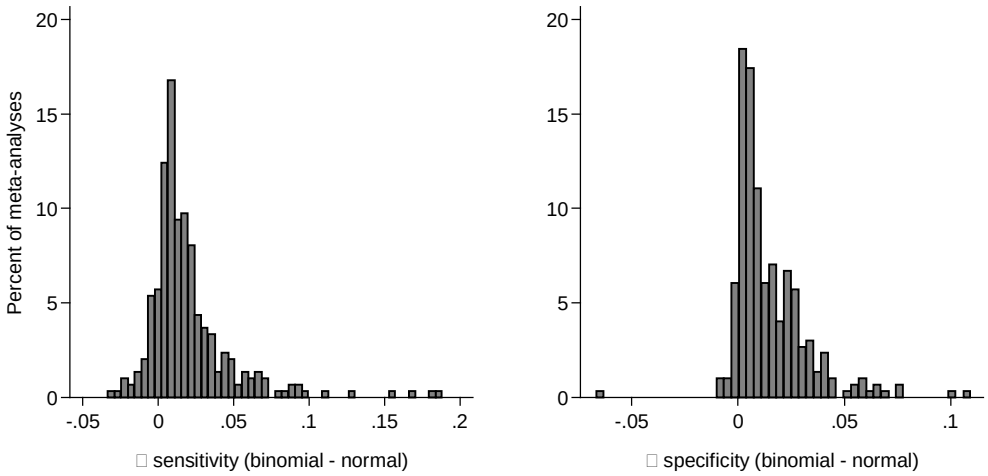
Figure 10 compares the point estimates (logit-transformed) and confidence interval widths for bivariate random effects meta-analyses of sensitivity and specificity using the exact binomial likelihood versus using a normal approximation for within-study variability. It is evident that point estimates 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 11 presents histograms of the differences in estimated sensitivities and specificities (untransformed scale) between methods. The results indicate that differences higher than 5% are not uncommon and typically suggest underestimation of sensitivity and specificity in meta-analyses using a normal approximation (compared to those using the exact binomial likelihood).

Figure 10. Comparison of point estimates and confidence interval widths of summary sensitivity and specificity (logit scale) from bivariate random effects meta-analyses (approximate normal vs. exact binomial; both models fit with MLE)



Note: 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; MLE = maximum likelihood estimation.

Figure 11. Histograms of differences in estimated summary sensitivity and specificity from bivariate random effects meta-analyses (approximate normal vs. exact binomial; both models fit with MLE)

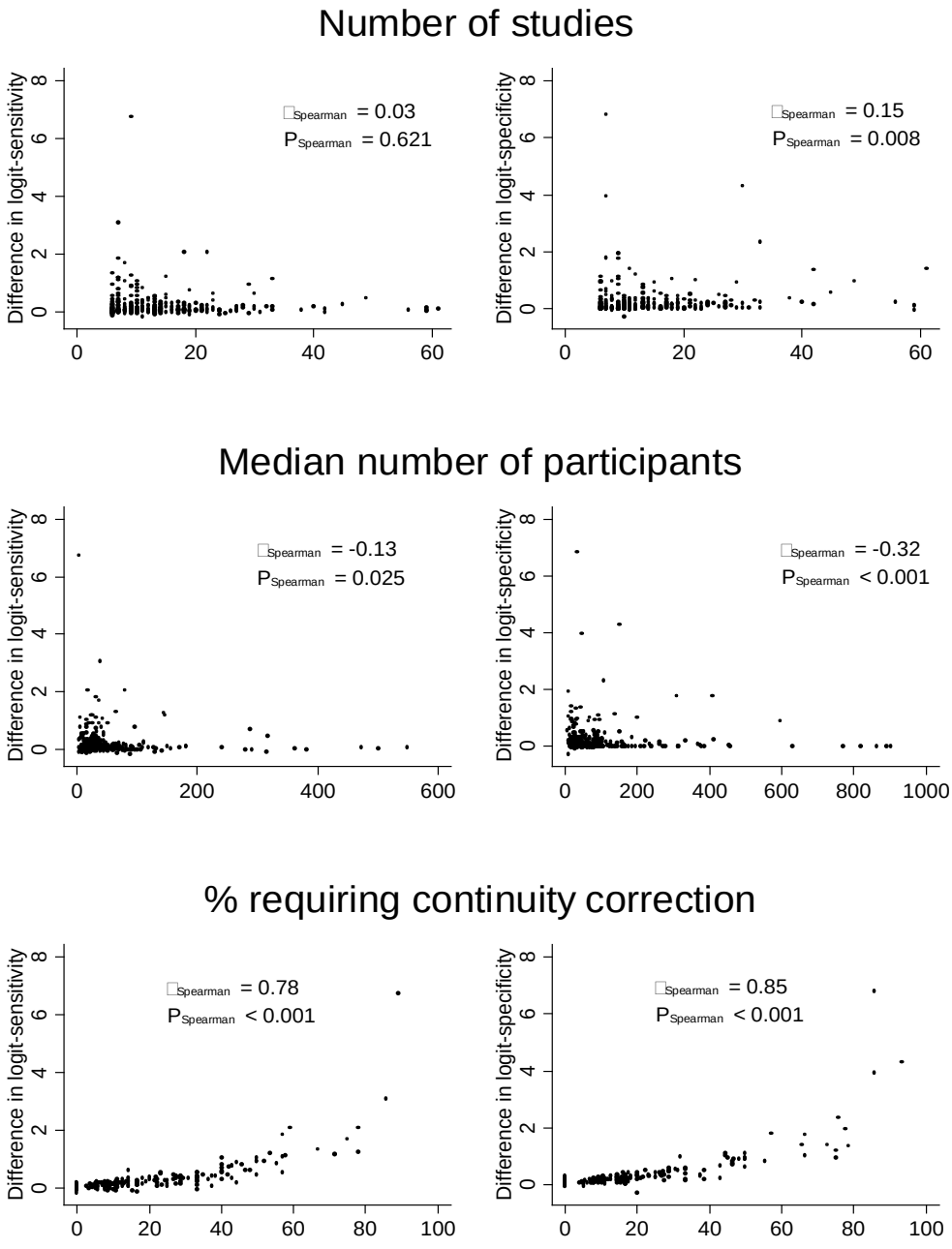


Note: Histograms of differences in estimated summary sensitivity (left panel) and specificity (right panel) comparing bivariate random effects meta-analysis using the exact binomial likelihood versus using a normal approximation for within-study variability.
MLE = maximum likelihood estimation.

Factors Influencing the Difference Between Bivariate Random Effects Meta-Analyses Using the Normal Versus Binomial Likelihood (Both With MLE)

Figure 12 presents scatter plots of the differences (logit-transformed) of estimated sensitivity and specificity versus factors that we hypothesized could affect estimation. The results demonstrate that differences between methods were larger in meta-analyses in which a large proportion of the available studies required a continuity correction (for the normal approximation) and in meta-analyses where studies were generally small. The total number of included studies in the meta-analysis generally had a smaller effect.

Figure 12. Differences of estimated sensitivity and specificity (logit scale, bivariate random effects meta-analyses using the exact binomial likelihood vs. a normal approximation; both models fit with MLE) over meta-analysis characteristics

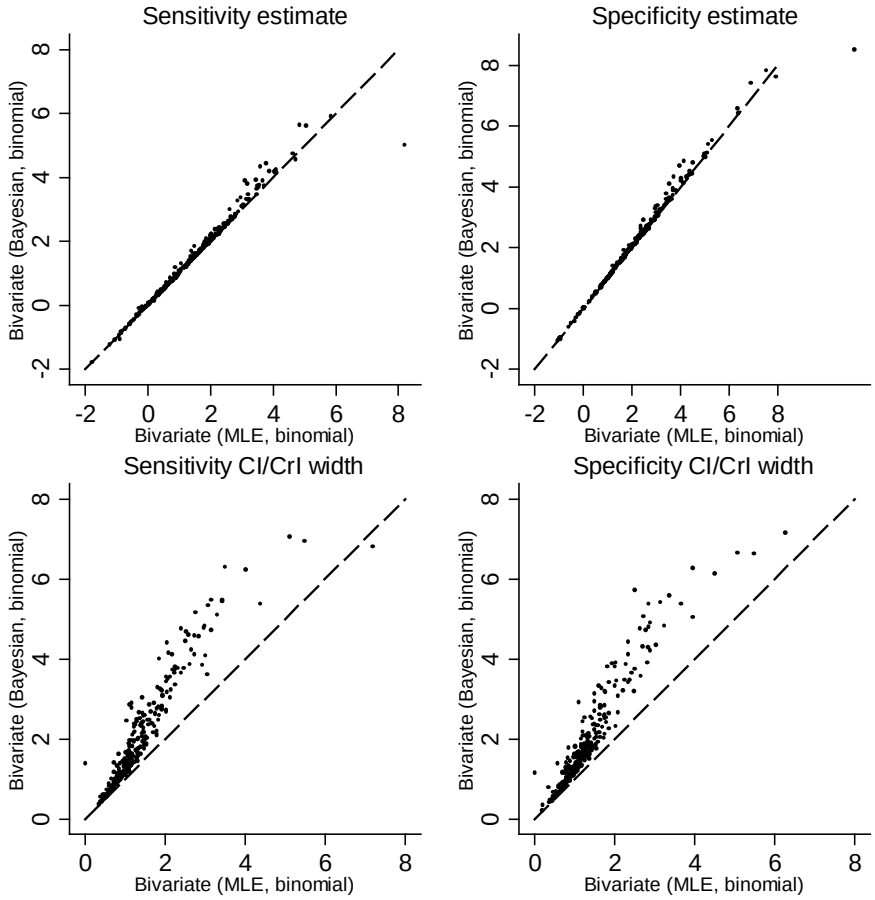


Note: Positive differences indicate that estimates using the binomial likelihood are larger than those using the normal approximation. Eight meta-analyses with a median number of unaffected individuals >1000 have not been plotted (in the middle right panel) to avoid distortion of the graph.

Bayesian Methods Versus MLE for the Binomial Bivariate Random Effects Model (Exact Likelihood)

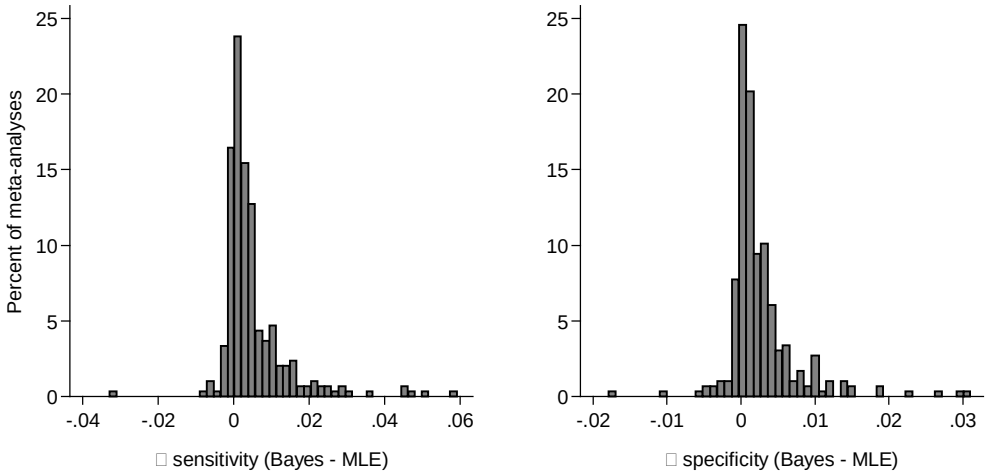
Figure 13 compares the point estimates (logit-transformed) and confidence or credibility interval widths for bivariate random effects meta-analyses of sensitivity and specificity using fully Bayesian methods and MLE (both based on the exact binomial likelihood for within-study variability). Point estimates 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 MLE. Figure 14 presents histograms of the differences in estimated sensitivities and specificities (untransformed scale) between the two methods, confirming that point estimate differences higher than 5 percent were very uncommon.

Figure 13. Comparison of point estimates and confidence/credibility interval widths of summary sensitivity and specificity (logit scale) from bivariate random effects meta-analyses (Bayesian versus MLE; both models using the exact binomial likelihood to represent within-study variability)



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

Figure 14. Histograms of differences in estimated summary sensitivity and specificity from bivariate random effects meta-analyses fit using fully Bayesian versus MLE estimation (using the exact binomial likelihood to represent within-study variability)

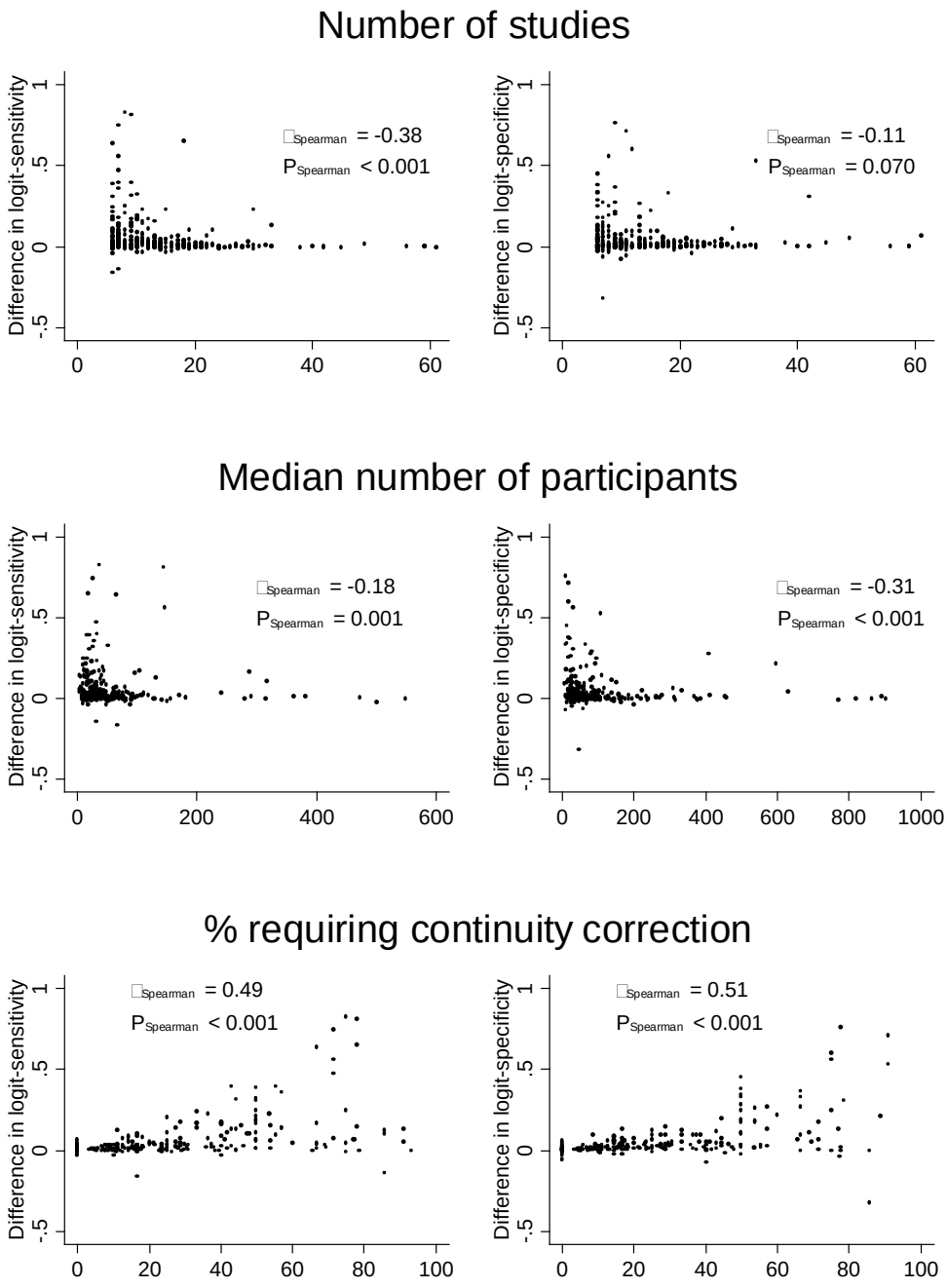


Note: Histograms of differences in estimated summary sensitivity (left panel) and specificity (right panel) comparing bivariate random effects meta-analysis models fit using fully Bayesian versus MLE (both models used the exact binomial likelihood to represent within-study variability).
MLE = maximum likelihood estimation.

Factors Influencing the Difference Between Bivariate Random Effects Meta-Analyses Using the Binomial Likelihood (MLE vs. Bayesian)

Figure 15 presents scatter plots of the differences (logit-transformed) of estimated sensitivity and specificity over factors that we hypothesized could affect estimation. Generally differences between methods were small. However, differences were smaller with increasing number of included studies and increasing median number of participants per study. Differences were larger with increasing proportion of studies requiring a continuity correction, possibly reflecting the larger impact of the priors in such cases.

Figure 15. Differences of estimated logit sensitivity and specificity between models fit with Bayesian methods versus MLE (both from bivariate random effects meta-analyses using the exact binomial likelihood) over meta-analysis characteristics

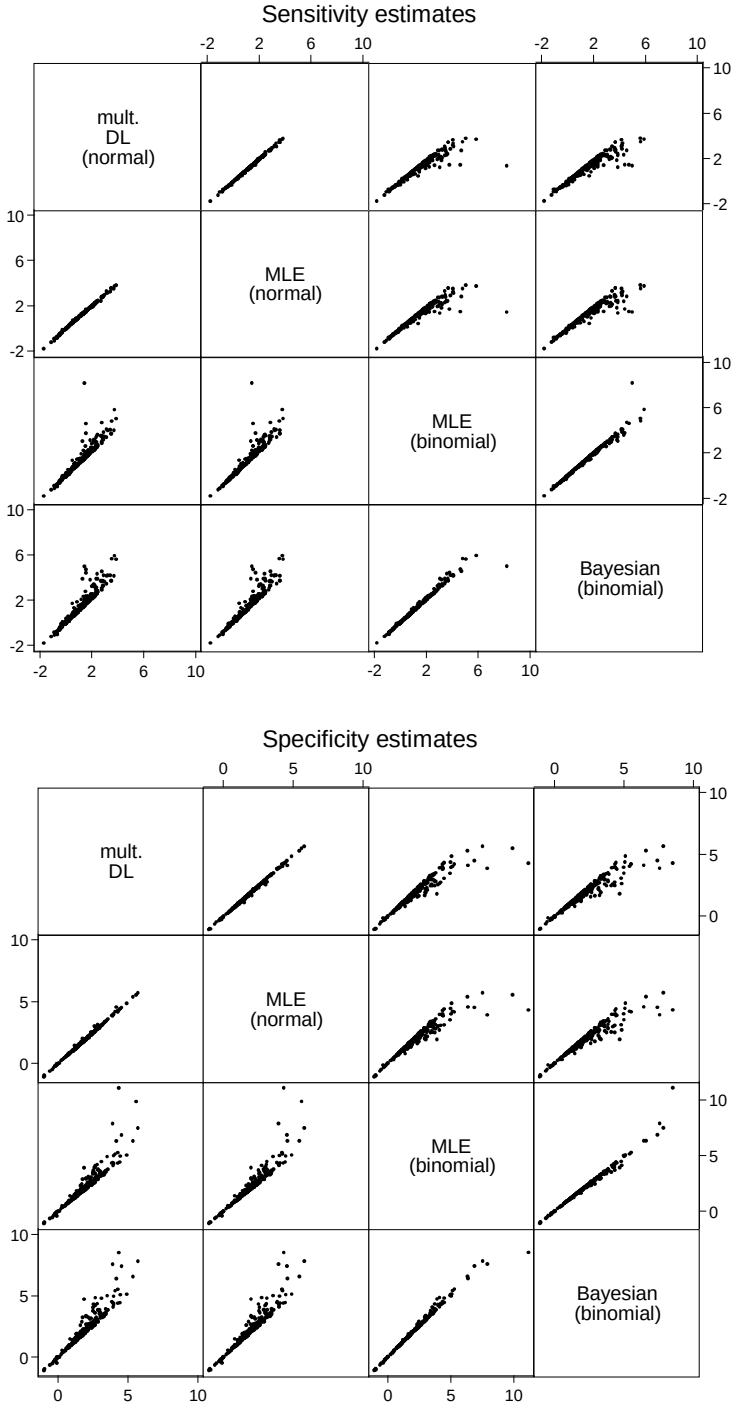


Note: Positive differences indicate that estimates from Bayesian analyses are larger than those from MLE. Three datapoints (1 for sensitivity and 2 for specificity) with absolute differences between methods larger than 1 have not been plotted to avoid distortion in the graph. An additional eight meta-analyses with a median number of unaffected individuals >1000 have not been plotted (in the middle right panel) for the same reason.

Summary for Bivariate Meta-Analysis Methods

Figure 16 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 ML [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 16. Summary comparison of sensitivity and specificity estimates (logit scale) from all univariate methods considered in this report



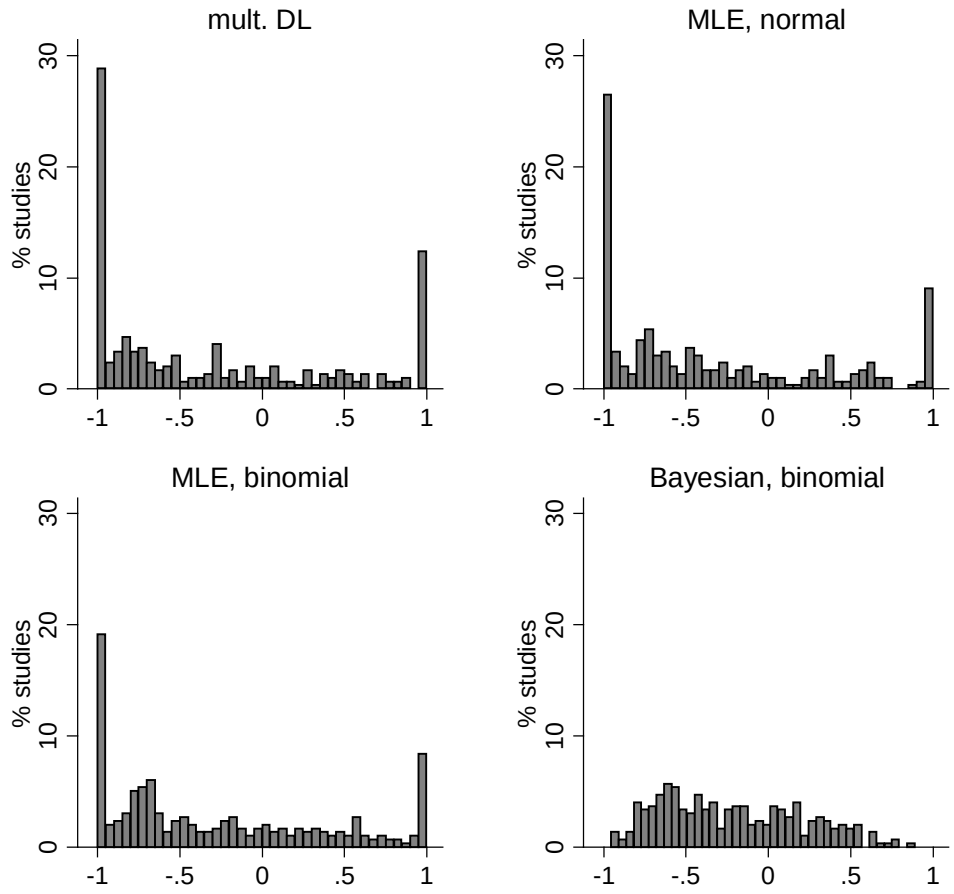
Binomial = model using the exact binomial likelihood for within-study variability; mult. DL= multivariate DerSimonian-Laird; normal = model using a normal approximation for within-study variability; MLE = maximum likelihood estimation.

Estimation of the Between-Study Correlation (Inverse Variance, MLE, 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 17 presents scatter plots of the estimated correlation from four bivariate models: the two models with a normal approximation (fit using either the multivariate DerSimonian-Laird method or REML), and two models using the exact binomial likelihood to describe within-study variability (fit using ML 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 18, which is a matrix scatter plot of the results from the same 4 methods. Figure 19 presents histograms of the differences in 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 20 shows the correlation point estimates and corresponding 95% confidence intervals from 228 studies for which they could be calculated (for studies where the correlation was estimated to be very close to -1 or +1, confidence intervals could not be calculated for numerical reasons). 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 154 cases where the point estimate of the correlation was negative (Fisher exact $p<0.001$ for the difference in statistical significance).

Figure 17. Histograms of the estimated correlation between sensitivity and specificity from the three bivariate methods compared in this report



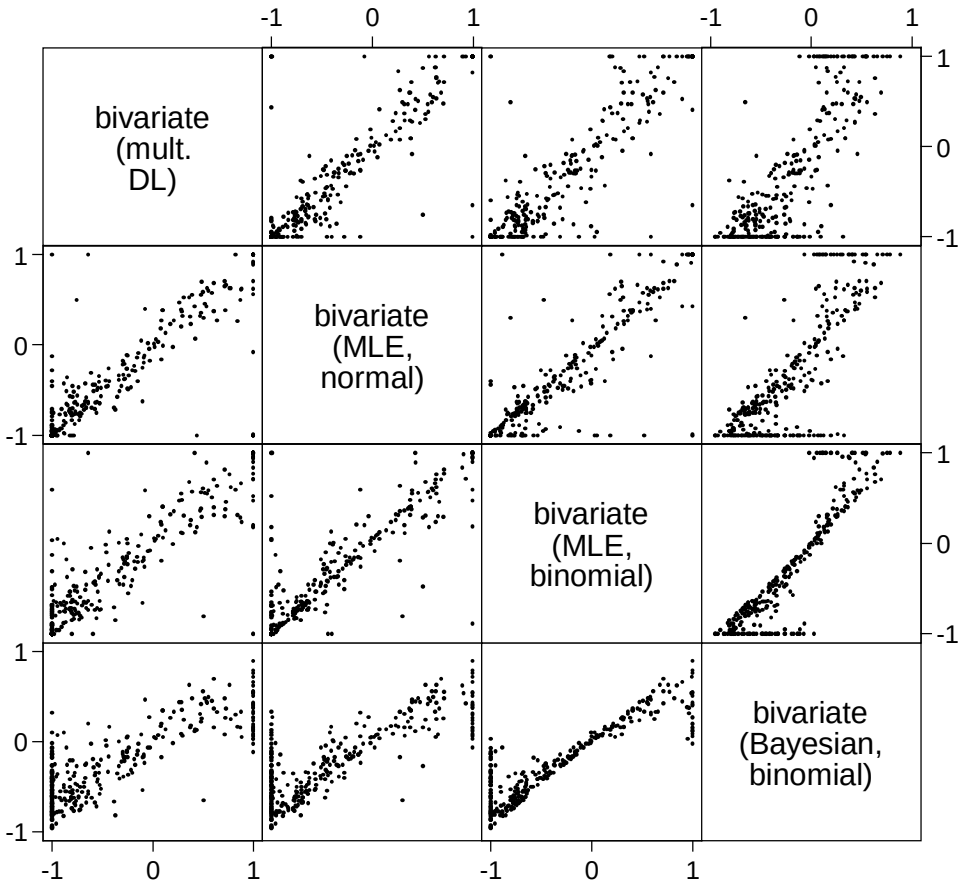
Note: Histograms of the estimated correlation between sensitivity and specificity from bivariate random effects meta-analysis models using: (top left) normal approximation with multivariate DerSimonian-Laird method; (top right) normal approximation with MLE; (bottom left) exact binomial likelihood with MLE; (bottom right) exact binomial likelihood with Bayesian model. Mult. DL = multivariate DerSimonian-Laird; MLE = maximum likelihood estimation.

Table 3. Estimated correlation by different bivariate methods

Method (Model /Approach for Within-Study Variability/ Model Fitting Method)	Estimated Correlation <0 N (%)
BREMA, normal, mult. DL*	208 (70)
BREMA, normal, MLE	214 (72)
BREMA, binomial, MLE	203 (68)
BREMA, binomial, Bayesian	200 (67)

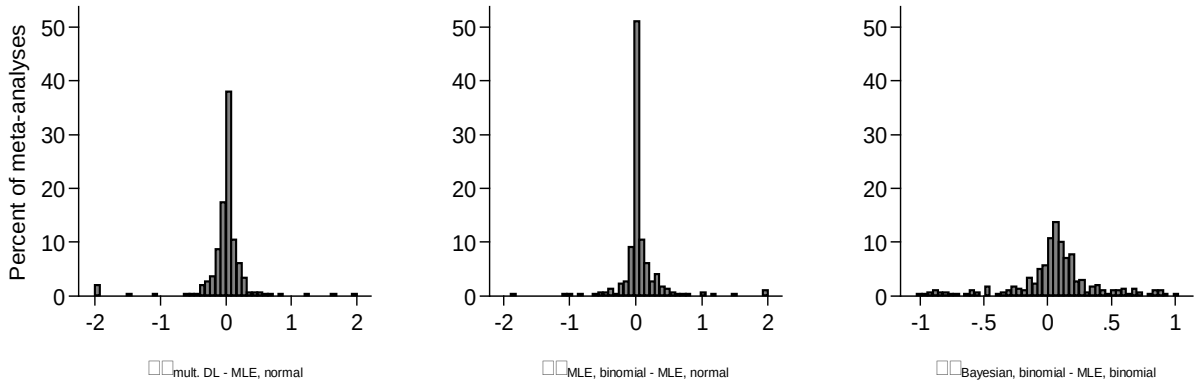
* In one meta-analysis the correlation could not be estimated.
BREMA = bivariate random effects meta-analysis; mult. DL = multivariate DerSimonian-Laird; MLE = maximum likelihood estimation.

Figure 18. Matrix scatter plot of correlation estimates from the four bivariate methods considered in this report



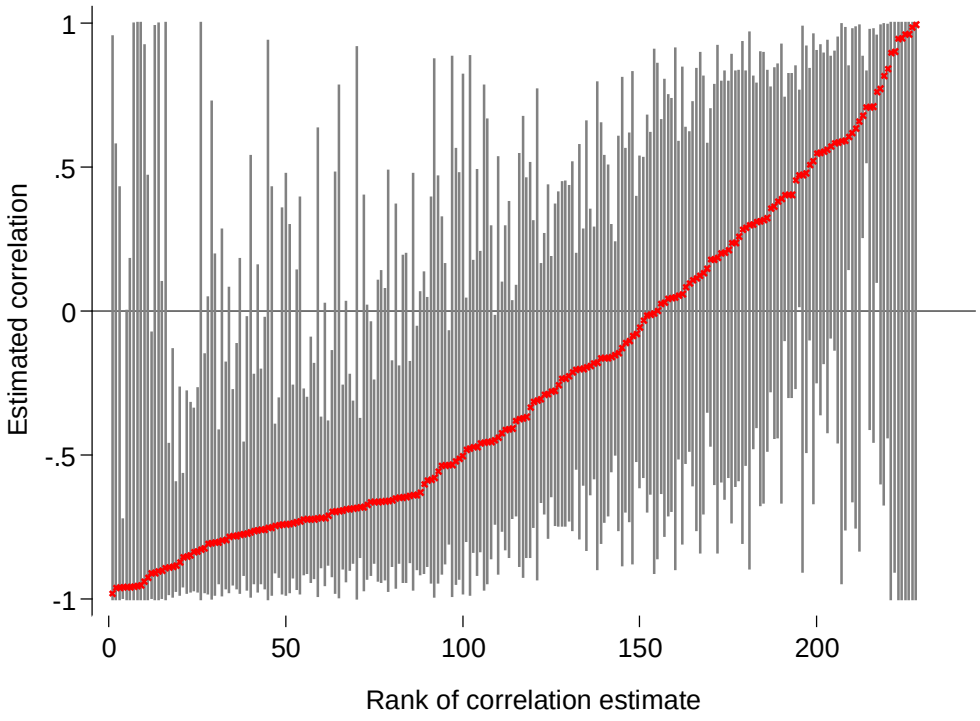
Binomial = model using the exact binomial likelihood to describe within-study variability; mult. DL = multivariate DerSimonian-Laird; normal = model using an approximate normal likelihood to describe within-study variability; MLE = maximum likelihood estimation.

Figure 19. Histograms of differences in correlation estimates from bivariate random effects meta-analyses (multivariate DerSimonian-Laird model using a normal approximation; MLE using a normal approximation; MLE using the exact binomial likelihood; and fully Bayesian model using the exact binomial likelihood)



Note: Histograms of differences in correlation estimates between models: top panel, multivariate DerSimonian-Laird method (with a normal approximation) versus MLE (with a normal approximation); middle panel, MLE with the exact binomial likelihood versus MLE with a normal approximation; bottom panel, Bayesian model (exact binomial likelihood) versus MLE (exact binomial likelihood).
ρ = correlation; bin. = model using the exact binomial likelihood; MLE = maximum likelihood estimation; norm. = model using a normal approximation.

Figure 20. Point estimates and 95% confidence intervals for the correlation of sensitivity and specificity, as estimated by the bivariate model using the exact binomial likelihood (MLE estimation)



Note: 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.
MLE = maximum likelihood estimation.

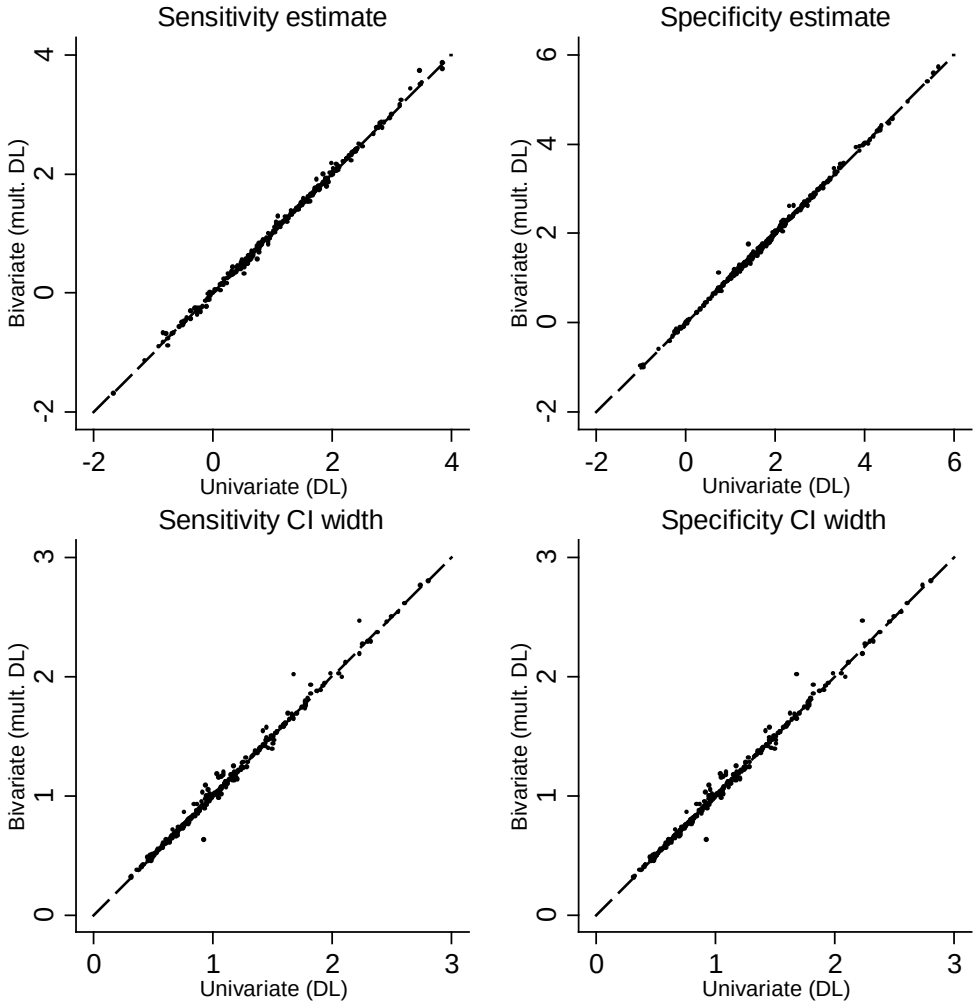
Comparison of Univariate and Bivariate Methods

This section compares the results of univariate and bivariate methods for meta-analysis of sensitivity and specificity. We stratify comparisons by estimation method (non-iterative versus iterative estimation of heterogeneity) and by use of normal distribution versus the binomial likelihood to model within-study variation.

Univariate Versus Bivariate Random Effects Meta-Analyses Using the Normal Approximation and Noniterative Heterogeneity Estimators (DerSimonian-Laird and Multivariate DerSimonian-Laird Methods)

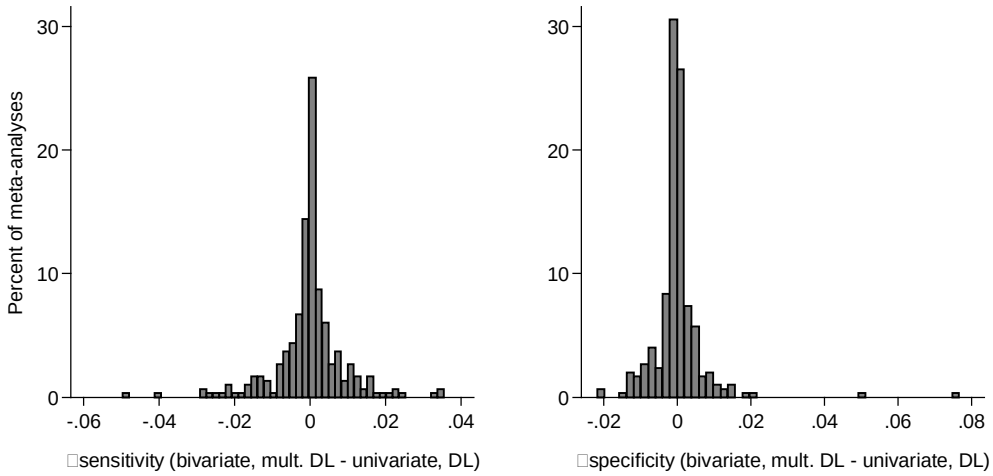
Figure 21 compares the point estimates (logit-transformed) and confidence interval widths for meta-analyses of sensitivity and specificity comparing univariate and bivariate random effects models (using a normal approximation for within-study variability). The univariate analyses used the DerSimonian-Laird 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 22 presents histograms of the differences in estimated sensitivities and specificities (untransformed scale) between methods, which emphasizes the general concordance of the point estimates (no differences larger than 10% were observed and only occasional differences beyond 5%).

Figure 21. Comparison of point estimates and confidence interval widths of summary sensitivity and specificity (logit scale, univariate random effects vs. bivariate random effects inverse variance methods, both using a normal approximation for within-study variability and a noniterative estimator for heterogeneity)



Note: 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 multivariate DerSimonian-Laird) meta-analysis (with an approximate normal likelihood to describe within-study variability).
CI = confidence interval; DL = DerSimonian-Laird; mult. DL = multivariate DerSimonian-Laird.

Figure 22. Histograms of differences in estimated summary sensitivity and specificity in univariate versus bivariate random effects inverse variance meta-analyses (both using a normal approximation for within-study variability)

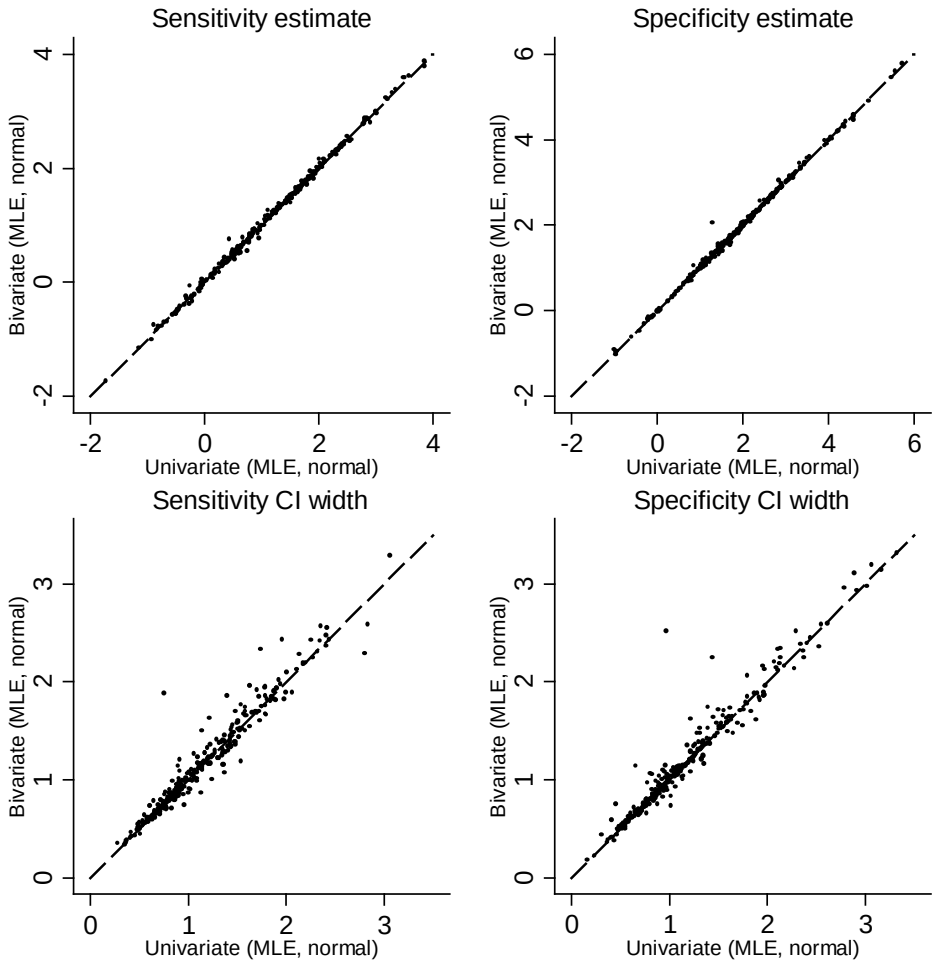


Note: Histograms of differences in estimated summary sensitivity (left panel) and specificity (right panel) comparing bivariate versus univariate random effects inverse variance meta-analyses (Der-Simonian-Laird vs. multivariate DerSimonian-Laird; both using a normal approximation for within-study variability).
DL = DerSimonian-Laird; mult. DL = multivariate DerSimonian-Laird.

Univariate Versus Bivariate Random Effects Meta-Analyses Using the Normal Approximation (Using MLE)

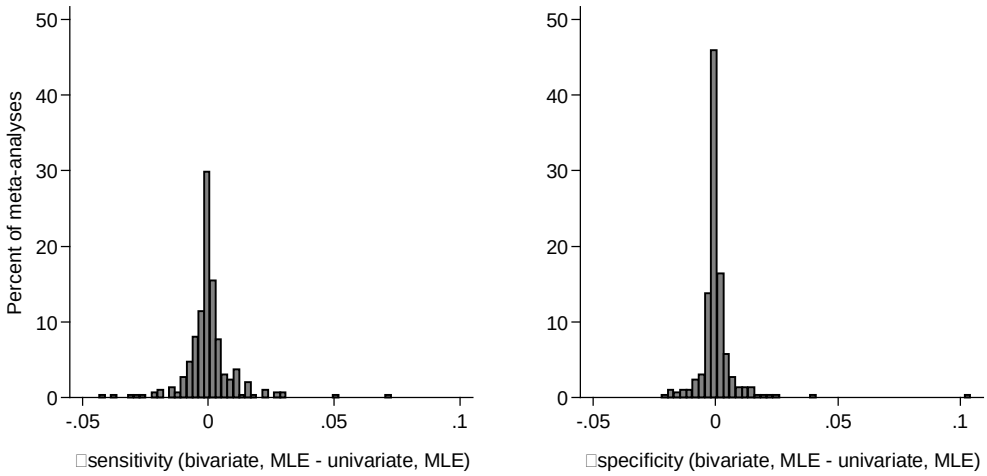
Figure 23 compares the point estimates (logit-transformed) and confidence interval widths for meta-analyses of sensitivity and specificity comparing univariate and bivariate random effects models using ML estimation (with a normal approximation for within-study variability). Overall, point estimates 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 24 presents histograms of the differences in estimated sensitivities and specificities (untransformed scale) between the two models; the graph emphasizes the general concordance of the point estimates of the two methods (only occasional differences beyond 2.5%).

Figure 23. Comparison of point estimates and standard errors of summary sensitivity and specificity (logit scale) from univariate versus bivariate random effects meta-analyses with MLE (using a normal approximation for within-study variability)



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

Figure 24. Histograms of differences in estimated summary sensitivity and specificity (logit scale) from univariate and bivariate random effects meta-analyses with MLE (using a normal approximation to represent within-study variability for both models)

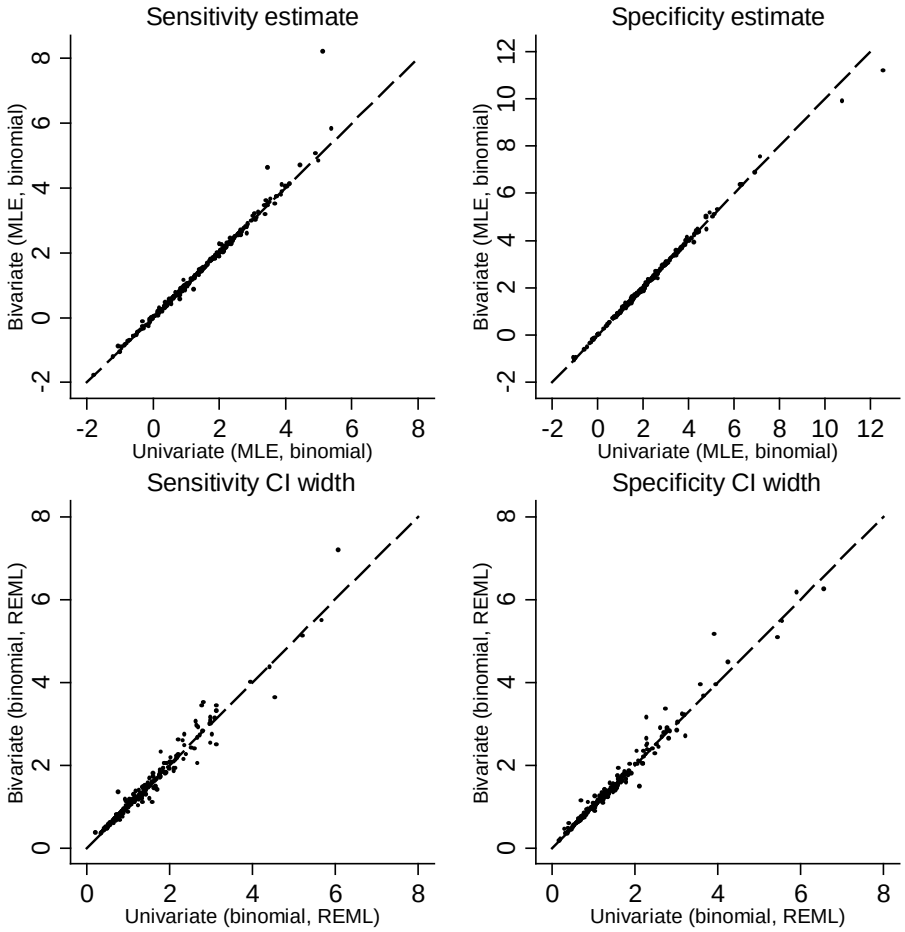


Note: Histograms of differences in estimated summary sensitivity (left panel) and specificity (right panel) comparing bivariate versus univariate random effects meta-analyses fit with MLE (using a normal approximation to represent within-study variability).
MLE = maximum likelihood estimation.

Univariate Versus Bivariate Random Effects Meta-Analyses Using the Exact Binomial Likelihood (MLE)

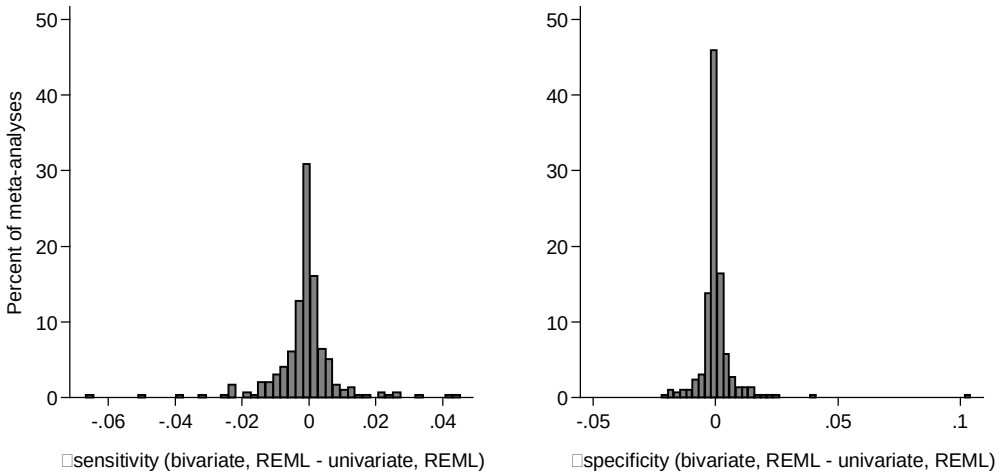
Figure 25 compares the point estimates (logit-transformed) and confidence interval widths for meta-analyses of sensitivity and specificity comparing univariate and bivariate models (using the exact binomial likelihood to represent within-study variability). Overall, point estimates and confidence interval widths from the two methods were similar. Figure 26 presents histograms of the differences in estimated sensitivities and specificities (untransformed scale) between methods and emphasizes the general concordance of the point estimates (only two discrepancies beyond 5%).

Figure 25. Comparison of point estimates and confidence interval widths of summary sensitivity and specificity (logit scale) from univariate versus bivariate random effects meta-analyses using the exact binomial likelihood (both models fit using MLE)



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

Figure 26. Histograms of differences in estimated summary sensitivity and specificity comparing univariate versus bivariate random effects meta-analyses fit with MLE (both models using the exact binomial likelihood to represent within-study variability)

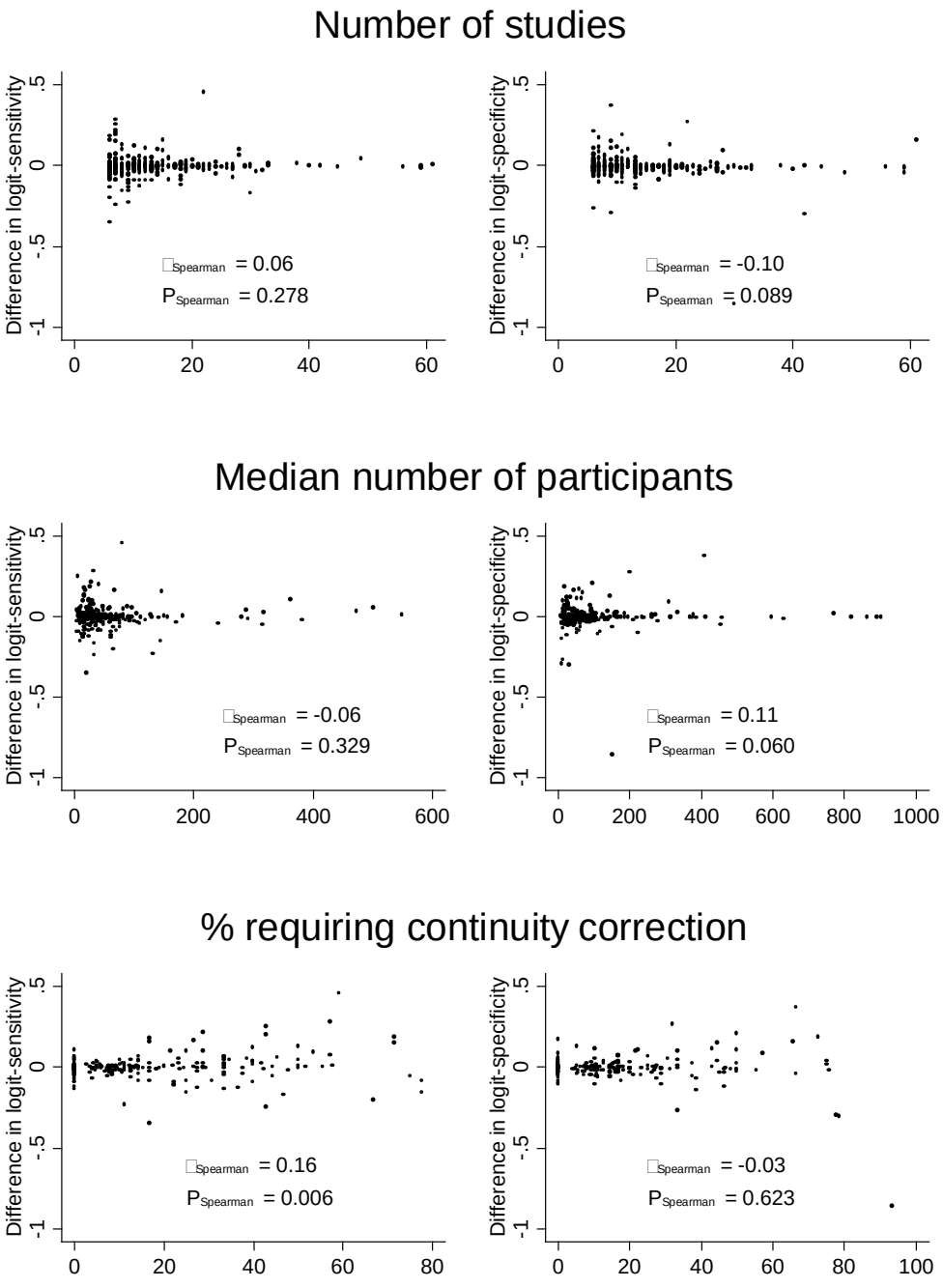


Note: Histograms of differences in estimated summary sensitivity (left panel) and specificity (right panel) comparing bivariate versus univariate random effects models fit using MLE (both models used the exact binomial likelihood to represent within-study variability).
MLE = maximum likelihood estimation.

Factors Influencing the Difference Between Univariate and Bivariate Random Effects Meta-Analyses Using the Binomial Likelihood (MLE)

Figure 27 presents scatter plots of the differences (logit-transformed) of estimated sensitivity and specificity over factors that we hypothesized could affect estimation. The results demonstrate that differences between methods were generally small and that no meta-analysis characteristic had a large effect on the differences between methods.

Figure 27. Differences of estimated sensitivity and specificity (logit scale) comparing univariate versus bivariate random effects meta-analyses (both using the exact binomial likelihood and fit using MLE) over meta-analysis characteristics

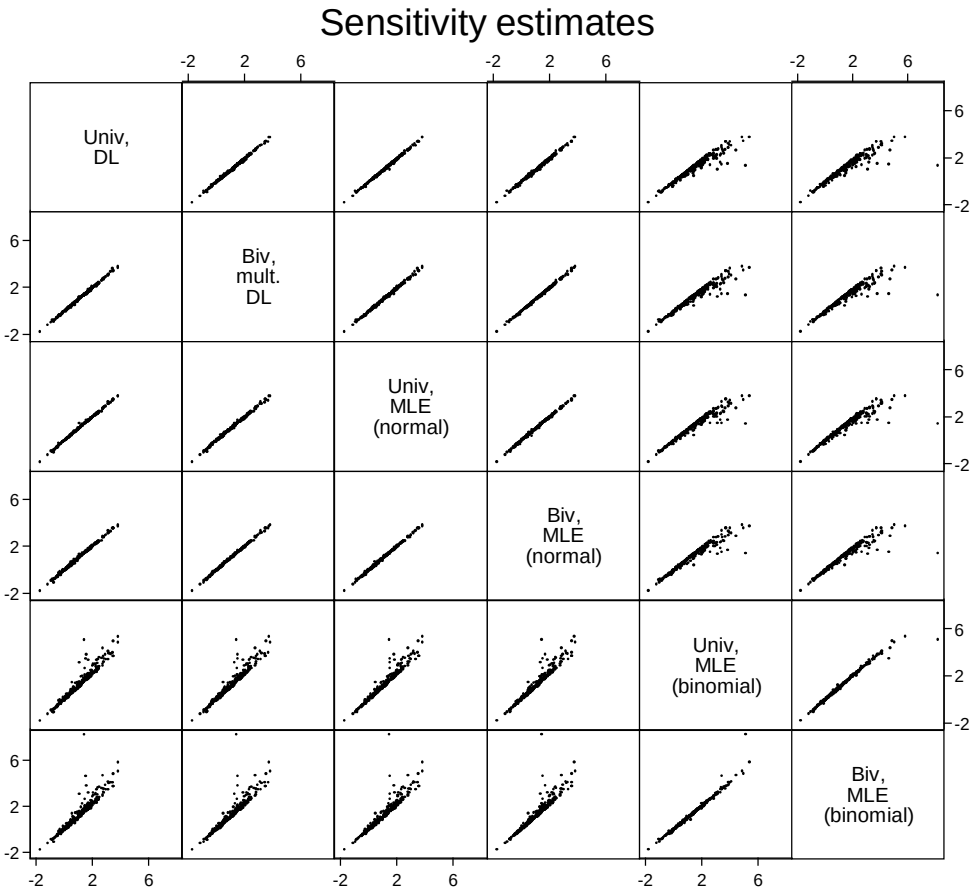


Positive differences indicate that estimates from univariate meta-analysis are larger than those from bivariate meta-analysis. Four datapoints (2 for sensitivity and 2 for specificity) with absolute differences in estimated values between methods larger than 1 have not been plotted to avoid distortion of the graph. An additional eight meta-analyses with a median number of unaffected individuals >1000 have not been plotted (in the middle right panel) for the same reason.

Summary for the Comparison of Univariate and Bivariate Methods

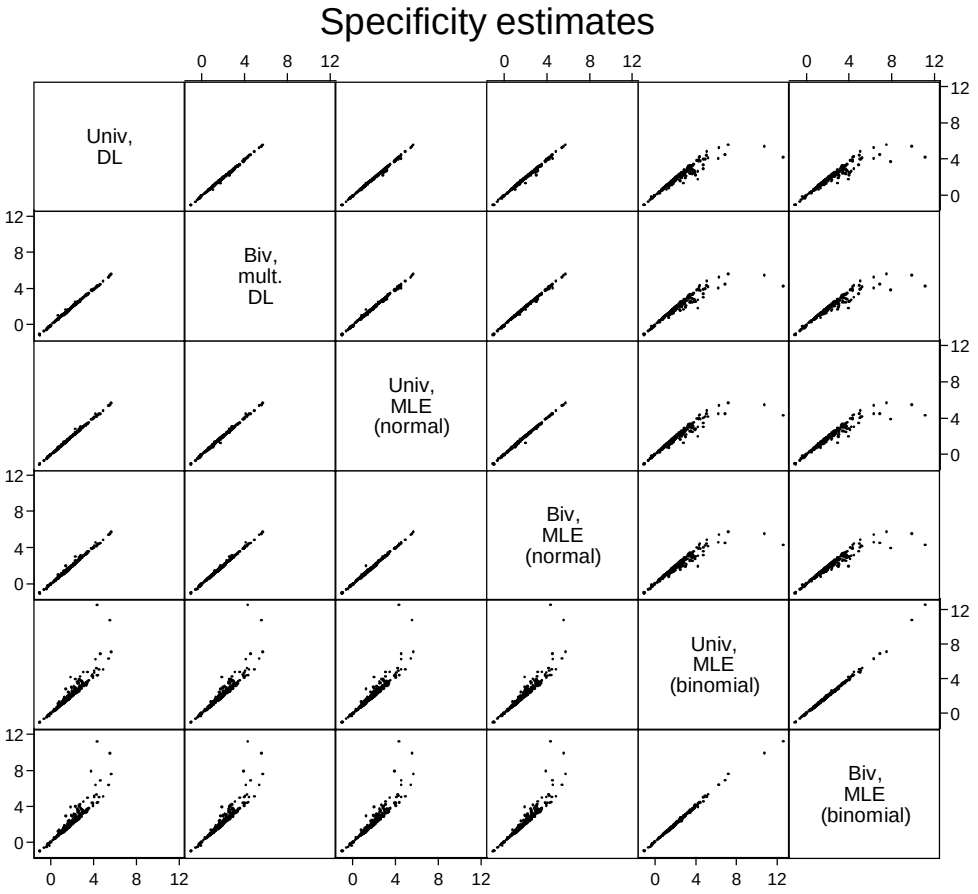
Figures 28 and 29 summarize the comparisons of effect sizes from all non-Bayesian random effects methods used in this report (univariate or bivariate; DerSimonian-Laird- or MLE-based; using the exact binomial likelihood or a normal approximation). Overall, the greatest discrepancies were observed between 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 (univariate or multivariate DerSimonian-Laird) or iterative (ML or REML) estimators of between-study variance.

Figure 28. Summary comparison of sensitivity estimates (logit scale) from all methods considered in this report



Binomial = model using the exact binomial likelihood; biv = bivariate; mult. DL = multivariate DerSimonian-Laird method; DL = DerSimonian-Laird model; normal = model using a normal approximation; MLE = maximum likelihood estimation; univ = univariate.

Figure 29. Summary comparison of specificity estimates (logit scale) from all methods considered in this report



Binomial = model using the exact binomial likelihood; biv = bivariate; mult. DL = multivariate DerSimonian-Laird method; DL = DerSimonian-Laird model; normal = model using a normal approximation; MLE = maximum likelihood estimation; univ = univariate.

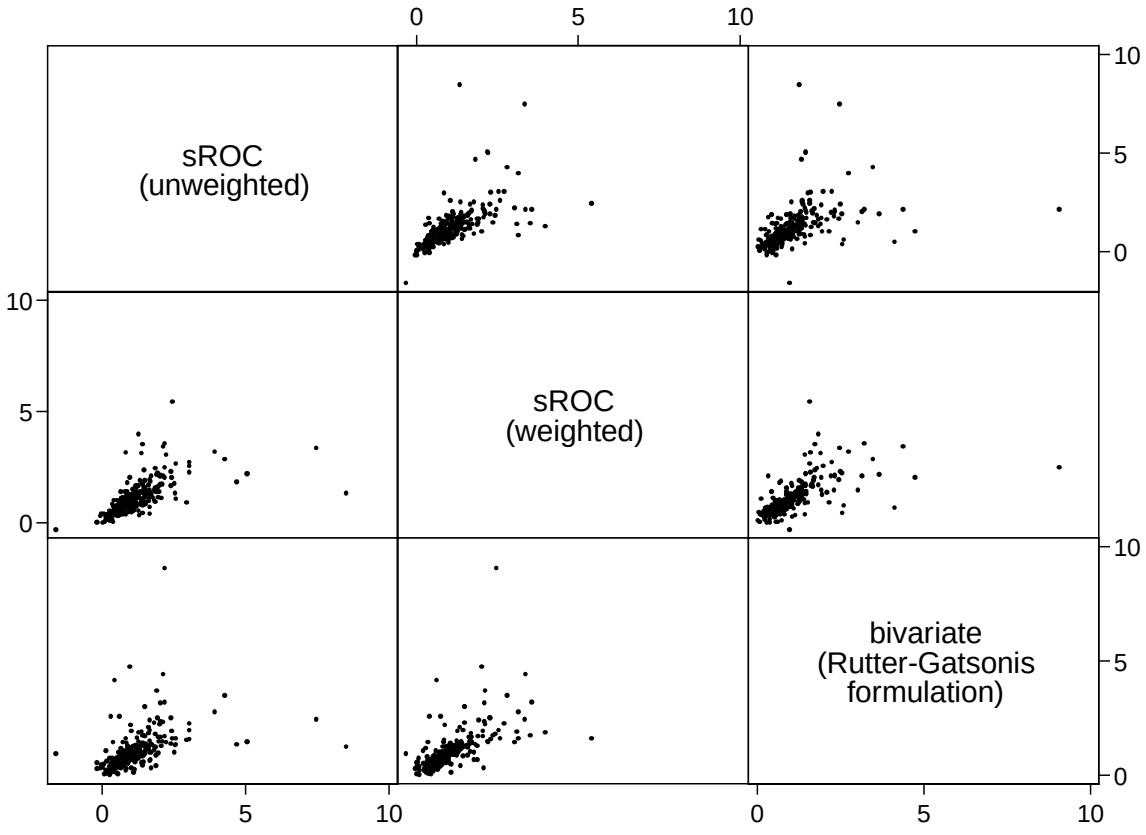
Meta-Analysis in the Receiver Operating Characteristic Space

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) on 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 Versus Rutter-Gatsonis HSROC

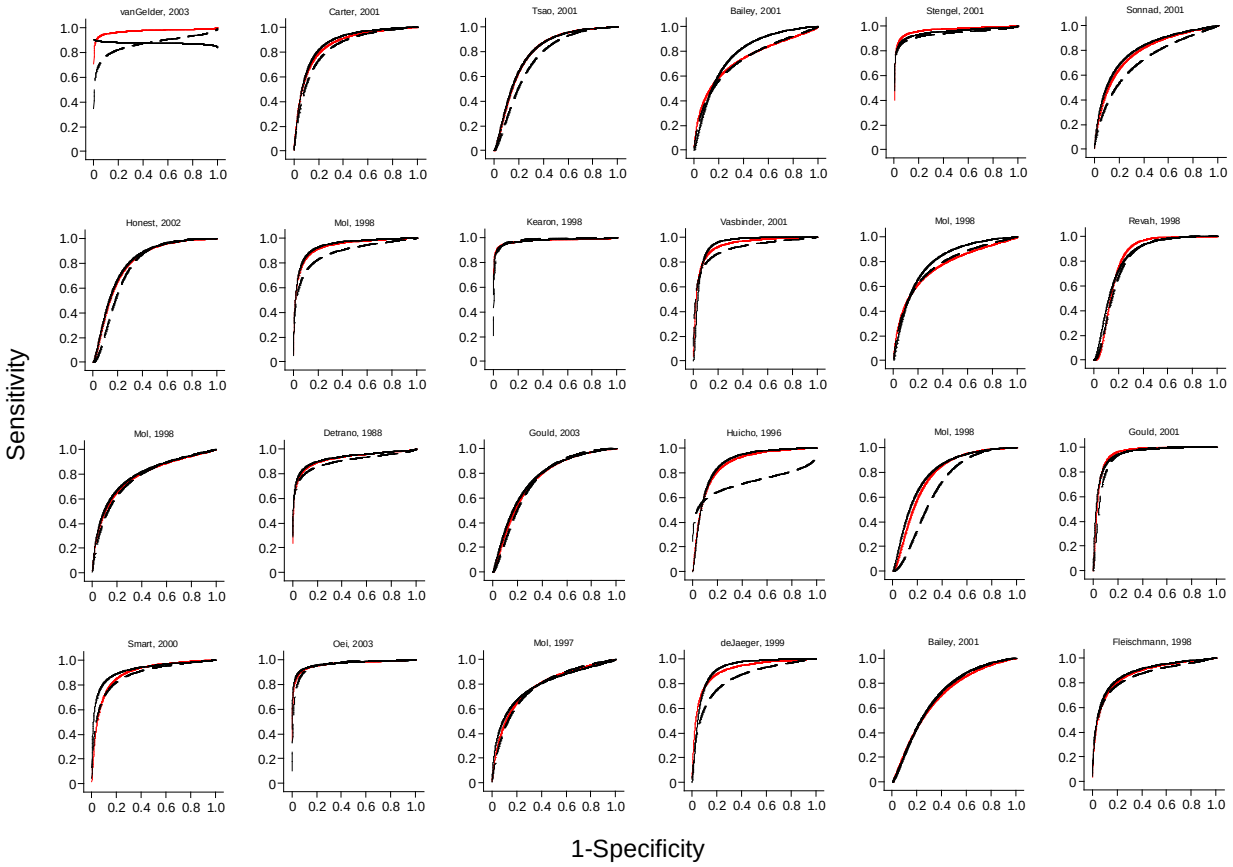
Figure 30 compares the slope in the logit space of ROC lines produced by the Moses-Littenberg (weighted and unweighted) and Rutter-Gatsonis methods. Figure 31 presents the SROC curves produced from these 3 methods for 24 randomly selected examples.

Figure 30. Scatter plot of the slopes of alternative SROC lines (logit space)



Note: Alternative SROC curves based on bivariate meta-analysis (to obtain estimates for the Rutter-Gatsonis HSROC; 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 31. SROC curves for 24 randomly selected meta-analyses (bivariate random effects model vs. Moses-Littenberg methods)



Note: 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. Estimates for SROC curves were produced by bivariate meta-analysis (red lines; using the Rutter-Gatsonis formulation), 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.

Alternative 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[28] to obtain alternative SROC curves based on the results of the bivariate random effects meta-analysis model using the exact binomial likelihood (fit with MLE). 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 square root 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 true and false positive rates 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.

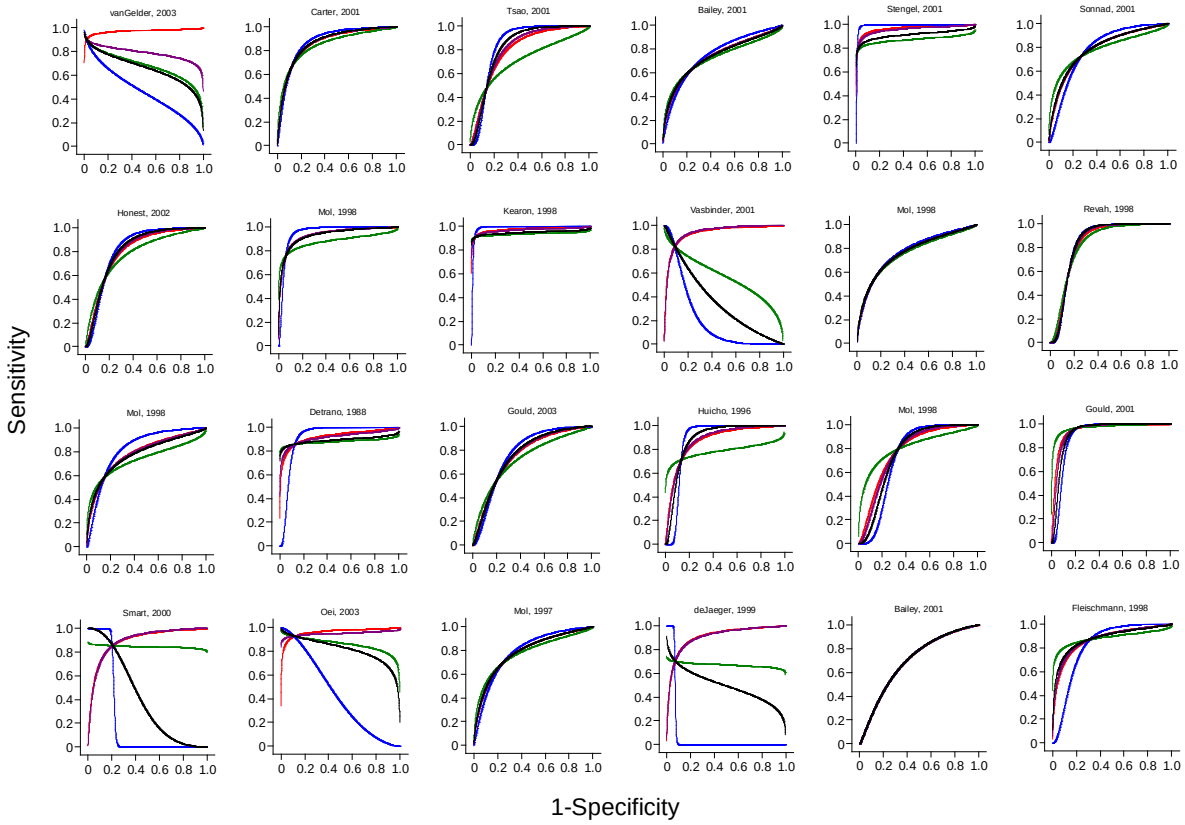
Figure 32 presents the alternative SROC curves discussed above for 24 randomly selected meta-analyses: the Rutter-Gatsonis slopes are always positive, whereas in some examples the slopes from the D~S, MAR of $\eta \sim \xi$, $\eta \sim \xi$, and $\xi \sim \eta$ models, are negative. The latter three 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 33 presents the study-level data and fitted HSROC curves for one of the datasets we analyzed[34] (fourth dataset on the top row of Figure 32. 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 less pronounced within 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 percent of the meta-analyses we performed; the D~S model results in negative slopes in 13 percent. By design, the Rutter-Gatsonis model always produces a positive slope. Figure 34 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).

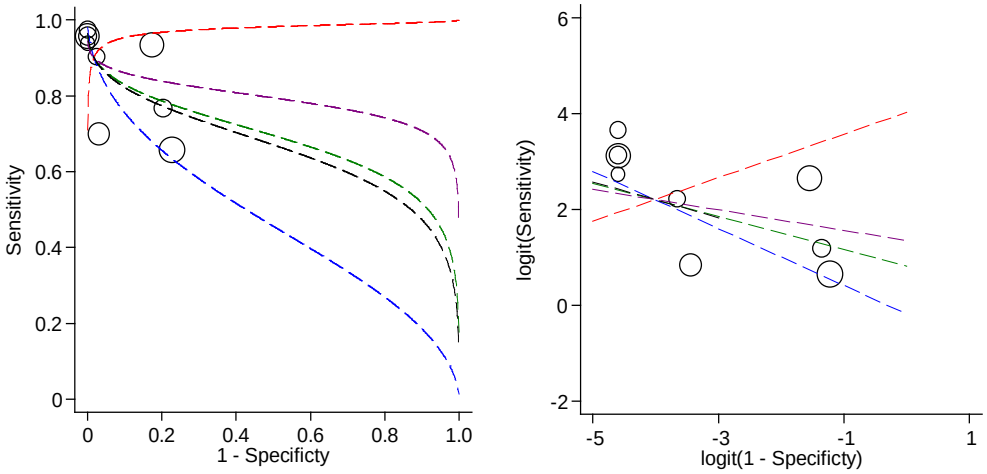
Appendix D presents a worked meta-analysis example applying all methods used in this report.

Figure 32. HSROC curves for 24 randomly selected meta-analyses (alternative parameterizations of the HSROC curve)



Note: 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), $\eta \sim \xi$ (green lines), $\xi \sim \eta$ (blue lines), D~S (purple lines), and MAR of $\eta \sim \xi$ (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 33. Study results and fitted HSROC curves for an example dataset



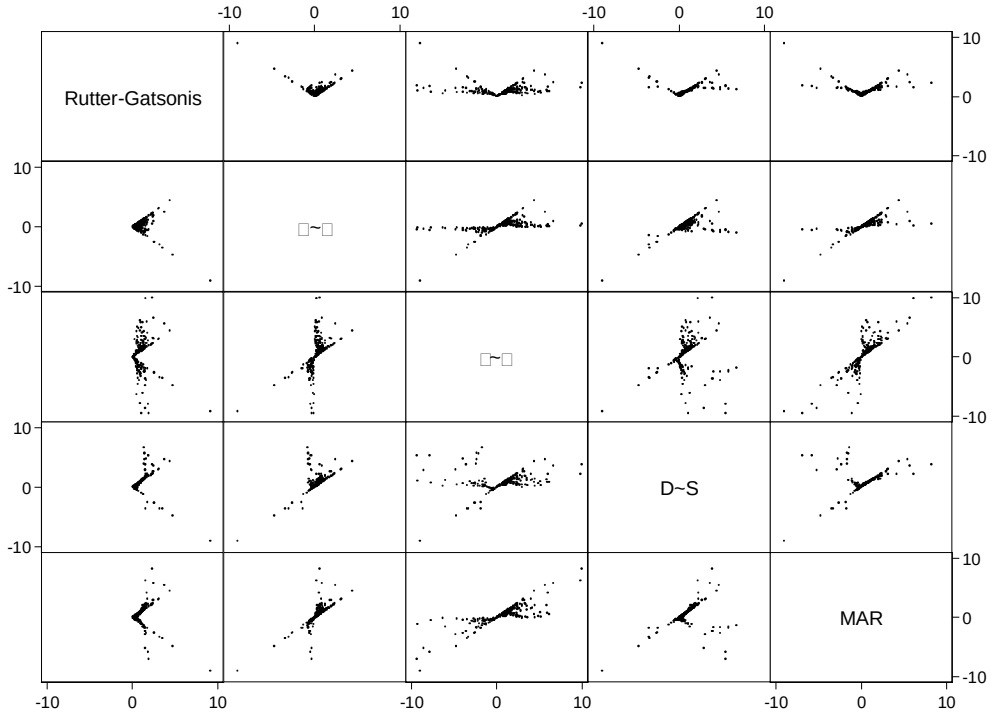
Note: Alternative SROC curves for the meta-analysis by van Gelder, 2003[34] (first dataset on the top row of Figure 32). Results are shown for the Rutter-Gatsonis (red lines), η~ξ (green lines), ξ~η (blue lines), D~S (purple lines), and MAR of η~ξ (black lines) parameterizations, in the ROC space (left panel) and the logit-transformed ROC space (right panel). All models were estimated using MLE. 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); this was done for illustration purposes only and does not affect the results of our analyses.
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; MLE = maximum likelihood estimation; η = 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	203 (68)	95 (32)
ξ ~ η	203 (68)	95 (32)
D ~ S	259 (87)	39 (13)
MAR	203 (68)	95 (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 34. Scatter plot of the slopes of alternative SROC 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 MLE) or the distribution used to model within-study variability (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 narrower confidence intervals, compared to methods that used the exact binomial likelihood. Although some of the differences between estimates were numerically large, their clinical importance is entirely context-specific. As expected, differences were larger in meta-analyses of small studies where continuity corrections (for the normal approximation) were needed for a large proportion of analyzed studies. Bivariate models fit using Bayesian and maximum likelihood methods produced almost identical summary estimates of sensitivity and specificity. The methods gave practically identical results in meta-analyses with moderate to large numbers of studies and when included studies had large sample sizes. 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[35]. 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.[28,33] 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).

We found that univariate and bivariate meta-analysis methods produced generally similar summary estimates and marginal confidence intervals for sensitivity and specificity. Differences are likely to be more pronounced when evaluating linear combinations of the estimates (e.g., the sum of sensitivity and specificity) particularly in problems of higher dimensionality (e.g., multiple index tests applied to the same patients and compared against a common reference standard). This issue is addressed in detail in a separate report of diagnostic tests in preparation by the EPC.

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).[17,19,20] 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 MLE; 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 made a similar observation in a simulation study.[36] 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	Number of MAs	Methods Compared and Model Fitting	Software	Key Findings	Authors' Recommendations/ Conclusions
Macaskill[17], 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[21], 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[14], 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.

Table 5. Summary of selected previous empirical comparisons of meta-analysis methods, including simulation studies (continued)

Author, Year	Number of MAs	Methods Compared and Model Fitting	Software	Key Findings	Authors' Recommendations/ Conclusions
Riley[36], 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. Multivariate meta-analysis methods allow borrowing strength across outcomes, particularly when data on some outcomes are missing at random.	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[28], 2008	2	BREMA normal vs. binomial; alternative SROC curves derived from the bivariate model vs. the SROC method (Moses-Littenberg)	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 SROC curves generated by the bivariate model differ substantially between them and may differ from the Moses-Littenberg SROC.	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 tests.
Hamza[35], 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[37], 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	The BREMA with exact binomial likelihood for the within-study model performed better than other methods.

Table 5. Summary of selected previous empirical comparisons of meta-analysis methods, including simulation studies (continued)

Author, Year	Number of MAs	Methods Compared and Model Fitting	Software	Key Findings	Authors' Recommendations/ Conclusions
				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.	
Harbord[22], 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[29], 2009	4 examples of an algorithm for deciding the optimal analysis method	Alternative SROC 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[23], 2009	2 (one with 3 index tests) "selected for highlighting the merits" of BREMA + 5 unselected	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% (25^{th}-75^{th} 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 probability (assuming a prior probability of 0.5) between methods was 2.5% (25^{th}-75^{th} perc.=2.2–3.2%, maximum=11%).	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.

Table 5. Summary of selected previous empirical comparisons of meta-analysis methods, including simulation studies (continued)

Author, Year	Number of MAs	Methods Compared and Model Fitting	Software	Key Findings	Authors' Recommendations/ Conclusions
Paul[19], 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[38], 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[20], 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.

Note: Unless otherwise stated analyses used ML methods. The studies by Macaskill [17] 2004; Reitsma[10] 2005; Chu & Cole[21] 2006; Harbord [14] 2007; Chappell 2009[29]; Simel [23] 2009; and Verde 2010 [20], considered a common meta-analysis example, based on a systematic review by Scheidler 1997 [39].
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 sensitivity and specificity. Based on our previous survey the most commonly used method for constructing SROC curves is the approach proposed by Moses and Littenberg.[12] 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.[13] 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.[14,28] 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[28] 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.[13,29] Importantly, in some cases the slope of the ROC curve is not always positive (in contrast to the Rutter-Gatsonis method) and, therefore, the relationship between sensitivity and specificity cannot be explained by threshold effects across studies. Based on this, Chappell et al. determine 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.[29]

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 meta-analyses published up to 2010), we found no substantial change over time in the number of included studies or the number of meta-analyses conducted per review article.[12]

Another limitation of our work is that many systematic reviews contributed multiple datasets to the empirical comparison (approximately two 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 nonoverlapping sets of primary studies). Unfortunately, data to explore the potential effect of such clustering are typically not available in meta-analyses or primary diagnostic test studies. Nonetheless, our approach can be considered representative of current practice in applied meta-analyses, where pairs of tests and diagnostic outcomes are almost always evaluated one at a time.

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, odds 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[37]). 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.[35] 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 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 sensitivity changes with 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).[28,29] In such cases the relationship between 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.[37]

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.[40-42]

Abbreviations

AHRQ	Agency for Healthcare Research and Quality
BREMA	bivariate random effects meta-analysis
EPC	Evidence-based Practice Center
FPR	false positive rate
HSROC	hierarchical summary receiver operating characteristic
MAR	major axis regression
ML	maximum likelihood
MLE	maximum likelihood estimation
REML	restricted maximum likelihood
ROC	receiver operating characteristic
SROC	summary receiver operating characteristic
TPR	true positive rate

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. Dahabreh IJ, Chung M, Kitsios GD, Terasawa T, Raman G, Tatsioni A et al. Evaluating Practices and Developing Tools for Comparative Effectiveness Reviews of Diagnostic Test Accuracy. Task 1: Comprehensive Overview of Methods and Reporting of Meta- Analyses of Test Accuracy. AHRQ Methods Research Report. AHRQ Publication No 12-EHC044-EF. 2012.
13. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001;20:2865-84.
14. 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.
15. Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med. 2002;21:589-624.
16. Harbord RM, Whiting P. metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression. The Stata Journal. 2009;9:211-29.
17. 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.
18. 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.
19. 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.

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

21. 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.

22. 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.

23. 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.

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

25. 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.

26. Rabe-Hesketh S, Skrondal A. Multilevel and longitudinal modeling using Stata. Stata Corp; 2008.

27. Rabe-Hesketh S, Skrondal A, Pickles A. Reliable estimation of generalized linear mixed models using adaptive quadrature. The Stata Journal. 2002;2:1-21.

28. 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.

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

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

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

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

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

34. 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.

35. 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.

36. 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.

37. 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.

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

39. 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.

40. 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.

41. 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.

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

Appendix A. Included Studies

1. Arbyn M, Schenck U, Ellison E, et al. Metaanalysis of the accuracy of rapid prescreening relative to full screening of pap smears. Cancer 2003 Feb 25;99(1):9-16.
2. Bachmann LM, Kolb E, Koller MT, et al. Accuracy of Ottawa ankle rules to exclude fractures of the ankle and mid-foot: systematic review. BMJ 2003 Feb 22;326(7386):417.
3. Badgett RG, Lucey CR, Mulrow CD. Can the clinical examination diagnose left-sided heart failure in adults? JAMA 1997 Jun 4;277(21):1712-9.
4. Bafounta ML, Beauchet A, Aegerter P, et al. Is dermoscopy (epiluminescence microscopy) useful for the diagnosis of melanoma? Results of a meta-analysis using techniques adapted to the evaluation of diagnostic tests. Arch Dermatol 2001 Oct;137(10):1343-50.
5. Bailey JJ, Berson AS, Handelsman H, et al. Utility of current risk stratification tests for predicting major arrhythmic events after myocardial infarction. J Am Coll Cardiol 2001 Dec;38(7):1902-11.
6. Balk EM, Ioannidis JP, Salem D, et al. Accuracy of biomarkers to diagnose acute cardiac ischemia in the emergency department: a meta-analysis. Ann Emerg Med 2001 May;37(5):478-94.
7. Barnes CJ, Pietrobon R, Higgins LD. Does the pulse examination in patients with traumatic knee dislocation predict a surgical arterial injury? A meta-analysis. J Trauma 2002 Dec;53(6):1109-14.
8. Bax JJ, Wijns W, Cornel JH, et al. Accuracy of currently available techniques for prediction of functional recovery after revascularization in patients with left ventricular dysfunction due to chronic coronary artery disease: comparison of pooled data. J Am Coll Cardiol 1997 Nov 15;30(6):1451-60.
9. Bax JJ, Poldermans D, Elhendy A, et al. Sensitivity, specificity, and predictive accuracies of various noninvasive techniques for detecting hibernating myocardium. Curr Probl Cardiol 2001 Feb;26(2):147-86.
10. Berger MY, van d, V, Lijmer JG, et al. Abdominal symptoms: do they predict gallstones? A systematic review. Scand J Gastroenterol 2000 Jan;35(1):70-6.
11. Bonis PA, Ioannidis JP, Cappelleri JC, et al. Correlation of biochemical response to interferon alfa with histological improvement in hepatitis C: a meta-analysis of diagnostic test characteristics. Hepatology 1997 Oct;26(4):1035-44.
12. Brown DL, Doubilet PM. Transvaginal sonography for diagnosing ectopic pregnancy: positivity criteria and performance characteristics. J Ultrasound Med 1994 Apr;13(4):259-66.
13. Brown MD, Rowe BH, Reeves MJ, et al. The accuracy of the enzyme-linked immunosorbent assay D-dimer test in the diagnosis of pulmonary embolism: a meta-analysis. Ann Emerg Med 2002 Aug;40(2):133-44.

14. Brown MD, Lau J, Nelson RD, et al. Turbidimetric D-dimer test in the diagnosis of pulmonary embolism: a metaanalysis. Clin Chem 2003 Nov;49(11):1846-53.

15. Carlson KJ, Skates SJ, Singer DE. Screening for ovarian cancer. Ann Intern Med 1994 Jul 15;121(2):124-32.

16. Carter BG, Butt W. Review of the use of somatosensory evoked potentials in the prediction of outcome after severe brain injury. Crit Care Med 2001 Jan;29(1):178-86.

17. Chappell ET, Moure FC, Good MC. Comparison of computed tomographic angiography with digital subtraction angiography in the diagnosis of cerebral aneurysms: a meta-analysis. Neurosurgery 2003 Mar;52(3):624-31.

18. Cher DJ, Conwell JA, Mandel JS. MRI for detecting silicone breast implant rupture: meta-analysis and implications. Ann Plast Surg 2001 Oct;47(4):367-80.

19. Chien PF, Khan KS, Ogston S, et al. The diagnostic accuracy of cervico-vaginal fetal fibronectin in predicting preterm delivery: an overview. Br J Obstet Gynaecol 1997 Apr;104(4):436-44.

20. Chien PF, Arnott N, Gordon A, et al. How useful is uterine artery Doppler flow velocimetry in the prediction of pre-eclampsia, intrauterine growth retardation and perinatal death? An overview. BJOG 2000 Feb;107(2):196-208.

21. Chin AS, Goldman LE, Eisenberg MJ. Functional testing after coronary artery bypass graft surgery: a meta-analysis. Can J Cardiol 2003 Jun;19(7):802-8.

22. Clark TJ, Mann CH, Shah N, et al. Accuracy of outpatient endometrial biopsy in the diagnosis of endometrial hyperplasia. Acta Obstet Gynecol Scand 2001 Sep;80(9):784-93.

23. Clarke CE, Davies P. Systematic review of acute levodopa and apomorphine challenge tests in the diagnosis of idiopathic Parkinson's disease. J Neurol Neurosurg Psychiatry 2000 Nov;69(5):590-4.

24. Colin C, Lanoir D, Touzet S, et al. Sensitivity and specificity of third-generation hepatitis C virus antibody detection assays: an analysis of the literature. J Viral Hepat 2001 Mar;8(2):87-95.

25. Conde-Agudelo A, Kafury-Goeta AC. Triple-marker test as screening for Down syndrome: a meta-analysis. Obstet Gynecol Surv 1998 Jun;53(6):369-76.

26. Contopoulos-Ioannidis DG, Ioannidis JP. Maternal cell-free viremia in the natural history of perinatal HIV-1 transmission: a meta-analysis. J Acquir Immune Defic Syndr Hum Retrovirol 1998 Jun 1;18(2):126-35.

27. Cook DJ, Fitzgerald JM, Guyatt GH, et al. Evaluation of the protected brush catheter and bronchoalveolar lavage in the diagnosis of nosocomial pneumonia. J Intensive Care Med 1991 Jul;6(4):196-205.

28. Dales RE, Stark RM, Raman S. Computed tomography to stage lung cancer. Approaching a controversy using meta-analysis. Am Rev Respir Dis 1990 May;141(5 Pt 1):1096-101.

29. de Albuquerque FL, Picano E. Comparison of dipyridamole and exercise stress echocardiography for detection of coronary artery disease (a meta-analysis). Am J Cardiol 2001 May 15;87(10):1193-6.

30. de Bernardinis, Violi V, Roncoroni L, et al. Discriminant power and information content of Ranson's prognostic signs in acute pancreatitis: a meta-analytic study. Crit Care Med 1999 Oct;27(10):2272-83.

31. de Jaeger, Litalien C, Lacroix J, et al. Protected specimen brush or bronchoalveolar lavage to diagnose bacterial nosocomial pneumonia in ventilated adults: a meta-analysis. Crit Care Med 1999 Nov;27(11):2548-60.

32. de Vries SO, Hunink MG, Polak JF. Summary receiver operating characteristic curves as a technique for meta-analysis of the diagnostic performance of duplex ultrasonography in peripheral arterial disease. Acad Radiol 1996 Apr;3(4):361-9.

33. Delgado-Bolton RC, Fernandez-Perez C, Gonzalez-Mate A, et al. Meta-analysis of the performance of 18F-FDG PET in primary tumor detection in unknown primary tumors. J Nucl Med 2003 Aug;44(8):1301-14.

34. Detrano R, Janosi A, Lyons KP, et al. Factors affecting sensitivity and specificity of a diagnostic test: the exercise thallium scintigram. Am J Med 1988 Apr;84(4):699-710.

35. Deville WL, van der Windt DA, Dzaferagic A, et al. The test of Lasegue: systematic review of the accuracy in diagnosing herniated discs. Spine 2000 May 1;25(9):1140-7.

36. Dijkhuizen FP, Mol BW, Brolmann HA, et al. The accuracy of endometrial sampling in the diagnosis of patients with endometrial carcinoma and hyperplasia: a meta-analysis. Cancer 2000 Oct 15;89(8):1765-72.

37. Dove SB. Radiographic diagnosis of dental caries. J Dent Educ 2001 Oct;65(10):985-90.

38. Dwamena BA, Sonnad SS, Angobaldo JO, et al. Metastases from non-small cell lung cancer: mediastinal staging in the 1990s--meta-analytic comparison of PET and CT. Radiology 1999 Nov;213(2):530-6.

39. Ebell MH, Flewelling D, Flynn CA. A systematic review of troponin T and I for diagnosing acute myocardial infarction. J Fam Pract 2000 Jun;49(6):550-6.

40. Ebell MH, White LL, Weismantel D. A systematic review of troponin T and I values as a prognostic tool for patients with chest pain. J Fam Pract 2000 Aug;49(8):746-53.

41. Engels EA, Terrin N, Barza M, et al. Meta-analysis of diagnostic tests for acute sinusitis. J Clin Epidemiol 2000 Aug;53(8):852-62.

42. Engman ML. An update on EBCT (Ultrafast CT) scans for coronary artery disease. J Insur Med 1998;30(3):175-9.

43. Fahey MT, Irwig L, Macaskill P. Meta-analysis of Pap test accuracy. Am J Epidemiol 1995 Apr 1;141(7):680-9.

44. Faron G, Boulvain M, Irion O, et al. Prediction of preterm delivery by fetal fibronectin: a meta-analysis. Obstet Gynecol 1998 Jul;92(1):153-8.

45. Farquhar C, Ekeroma A, Furness S, et al. A systematic review of transvaginal ultrasonography, sonohysterography and hysteroscopy for the investigation of abnormal uterine bleeding in premenopausal women. Acta Obstet Gynecol Scand 2003 Jun;82(6):493-504.

46. Fleischmann KE, Hunink MG, Kuntz KM, et al. Exercise echocardiography or exercise SPECT imaging? A meta-analysis of diagnostic test performance. JAMA 1998 Sep 9;280(10):913-20.

47. Fraile M, Rull M, Julian FJ, et al. Sentinel node biopsy as a practical alternative to axillary lymph node dissection in breast cancer patients: an approach to its validity. Ann Oncol 2000 Jun;11(6):701-5.

48. Gianrossi R, Detrano R, Colombo A, et al. Cardiac fluoroscopy for the diagnosis of coronary artery disease: a meta analytic review. Am Heart J 1990 Nov;120(5):1179-88.

49. Gill CJ, Lau J, Gorbach SL, et al. Diagnostic accuracy of stool assays for inflammatory bacterial gastroenteritis in developed and resource-poor countries. Clin Infect Dis 2003 Aug 1;37(3):365-75.

50. Gisbert JP, Pajares JM. Diagnosis of Helicobacter pylori infection by stool antigen determination: a systematic review. Am J Gastroenterol 2001 Oct;96(10):2829-38.

51. Glas AS, Roos D, Deutekom M, et al. Tumor markers in the diagnosis of primary bladder cancer. A systematic review. J Urol 2003 Jun;169(6):1975-82.

52. Goodman CM, Cohen V, Thornby J, et al. The life span of silicone gel breast implants and a comparison of mammography, ultrasonography, and magnetic resonance imaging in detecting implant rupture: a meta-analysis. Ann Plast Surg 1998 Dec;41(6):577-85.

53. Gordon I, Barkovics M, Pindoria S, et al. Primary vesicoureteric reflux as a predictor of renal damage in children hospitalized with urinary tract infection: a systematic review and meta-analysis. J Am Soc Nephrol 2003 Mar;14(3):739-44.

54. Gould MK, Maclean CC, Kuschner WG, et al. Accuracy of positron emission tomography for diagnosis of pulmonary nodules and mass lesions: a meta-analysis. JAMA 2001 Feb 21;285(7):914-24.

55. Gould MK, Kuschner WG, Rydzak CE, et al. Test performance of positron emission tomography and computed tomography for mediastinal staging in patients with non-small-cell lung cancer: a meta-analysis. Ann Intern Med 2003 Dec 2;139(11):879-92.

56. Greco S, Girardi E, Masciangelo R, et al. Adenosine deaminase and interferon gamma measurements for the diagnosis of tuberculous pleurisy: a meta-analysis. Int J Tuberc Lung Dis 2003 Aug;7(8):777-86.

57. Hallan S, Asberg A. The accuracy of C-reactive protein in diagnosing acute appendicitis--a meta-analysis. Scand J Clin Lab Invest 1997 Aug;57(5):373-80.

58. Harvey RT, Gefter WB, Hrung JM, et al. Accuracy of CT angiography versus pulmonary angiography in the diagnosis of acute pulmonary embolism: evaluation of the literature with summary ROC curve analysis. Acad Radiol 2000 Oct;7(10):786-97.

59. Hobby JL, Tom BD, Bearcroft PW, et al. Magnetic resonance imaging of the wrist: diagnostic performance statistics. Clin Radiol 2001 Jan;56(1):50-7.

60. Hoffman RM, Clanon DL, Littenberg B, et al. Using the free-to-total prostate-specific antigen ratio to detect prostate cancer in men with nonspecific elevations of prostate-specific antigen levels. J Gen Intern Med 2000 Oct;15(10):739-48.

61. Honest H, Bachmann LM, Gupta JK, et al. Accuracy of cervicovaginal fetal fibronectin test in predicting risk of spontaneous preterm birth: systematic review. BMJ 2002 Aug 10;325(7359):301.

62. Hoogendam A, Buntinx F, de Vet HC. The diagnostic value of digital rectal examination in primary care screening for prostate cancer: a meta-analysis. Fam Pract 1999 Dec;16(6):621-6.

63. Hrung JM, Sonnad SS, Schwartz JS, et al. Accuracy of MR imaging in the work-up of suspicious breast lesions: a diagnostic meta-analysis. Acad Radiol 1999 Jul;6(7):387-97.

64. Huicho L, Campos M, Rivera J, et al. Fecal screening tests in the approach to acute infectious diarrhea: a scientific overview. Pediatr Infect Dis J 1996 Jun;15(6):486-94.

65. Huicho L, Campos-Sanchez M, Alamo C. Metaanalysis of urine screening tests for determining the risk of urinary tract infection in children. Pediatr Infect Dis J 2002 Jan;21(1):1-11, 88.

66. Hurley JC. Concordance of endotoxemia with gram-negative bacteremia. A meta-analysis using receiver operating characteristic curves. Arch Pathol Lab Med 2000 Aug;124(8):1157-64.

67. Imran MB, Khan MA, Aslam MN, et al. Diagnosis of coronary artery disease by stress echocardiography and perfusion scintigraphy. J Coll Physicians Surg Pak 2003 Aug;13(8):465-70.

68. Ioannidis JP, Salem D, Chew PW, et al. Accuracy and clinical effect of out-of-hospital electrocardiography in the diagnosis of acute cardiac ischemia: a meta-analysis. Ann Emerg Med 2001 May;37(5):461-70.

69. Jensen JE, Nielsen SH, Foged L, et al. The MICRAL test for diabetic microalbuminuria: predictive values as a function of prevalence. Scand J Clin Lab Invest 1996 Apr;56(2):117-22.

70. Jorgensen K, Andersen TJ, Dam H. The diagnostic efficiency of the Rorschach Depression Index and the Schizophrenia Index: a review. Assessment 2000 Sep;7(3):259-80.

71. Jorm AF. Methods of screening for dementia: a meta-analysis of studies comparing an informant questionnaire with a brief cognitive test. Alzheimer Dis Assoc Disord 1997 Sep;11(3):158-62.

72. Kardaun JW, Kardaun OJ. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. Methods Inf Med 1990 Jan;29(1):12-22.

73. Kearon C, Julian JA, Newman TE, et al. Noninvasive diagnosis of deep venous thrombosis. McMaster Diagnostic Imaging Practice Guidelines Initiative. Ann Intern Med 1998 Apr 15;128(8):663-77.

74. Kellen M, Aronson S, Roizen MF, et al. Predictive and diagnostic tests of renal failure: a review. Anesth Analg 1994 Jan;78(1):134-42.

75. Kertai MD, Boersma E, Bax JJ, et al. A meta-analysis comparing the prognostic accuracy of six diagnostic tests for predicting perioperative cardiac risk in patients undergoing major vascular surgery. Heart 2003 Nov;89(11):1327-34.

76. Kinkel K, Hricak H, Lu Y, et al. US characterization of ovarian masses: a meta-analysis. Radiology 2000 Dec;217(3):803-11.

77. Kittler H, Pehamberger H, Wolff K, et al. Diagnostic accuracy of dermoscopy. Lancet Oncol 2002 Mar;3(3):159-65.

78. Klassen TP, Rowe PC. Selecting diagnostic tests to identify febrile infants less than 3 months of age as being at low risk for serious bacterial infection: a scientific overview. J Pediatr 1992 Nov;121(5 Pt 1):671-6.

79. Koelemay MJ, Lijmer JG, Stoker J, et al. Magnetic resonance angiography for the evaluation of lower extremity arterial disease: a meta-analysis. JAMA 2001 Mar 14;285(10):1338-45.

80. Koumans EH, Johnson RE, Knapp JS, et al. Laboratory testing for Neisseria gonorrhoeae by recently introduced nonculture tests: a performance review with clinical and public health considerations 15. Clin Infect Dis 1998 Nov;27(5):1171-80.

81. Kowalski J, Tu XM, Jia G, et al. A comparative meta-analysis on the variability in test performance among FDA-licensed enzyme immunosorbent assays for HIV antibody testing. J Clin Epidemiol 2001 May;54(5):448-61.

82. Kwok Y, Kim C, Grady D, et al. Meta-analysis of exercise testing to detect coronary artery disease in women. Am J Cardiol 1999 Mar 1;83(5):660-6.

83. Lensing AW, Hirsh J. 125I-fibrinogen leg scanning: reassessment of its role for the diagnosis of venous thrombosis in post-operative patients. Thromb Haemost 1993 Jan 11;69(1):2-7.

84. LeSar CJ, Meier GH, DeMasi RJ, et al. The utility of color duplex ultrasonography in the diagnosis of temporal arteritis. J Vasc Surg 2002 Dec;36(6):1154-60.

85. Lewis JD, Ng K, Hung KE, et al. Detection of proximal adenomatous polyps with screening sigmoidoscopy: a systematic review and meta-analysis of screening colonoscopy. Arch Intern Med 2003 Feb 24;163(4):413-20.

86. Li J. Capnography alone is imperfect for endotracheal tube placement confirmation during emergency intubation
16. J Emerg Med 2001 Apr;20(3):223-9.

87. Liberman M, Sampalis F, Mulder DS, et al. Breast cancer diagnosis by scintimammography: a meta-analysis and review of the literature. Breast Cancer Res Treat 2003 Jul;80(1):115-26.

88. Littenberg B, Mushlin AI. Technetium bone scanning in the diagnosis of osteomyelitis: a meta-analysis of test performance. Diagnostic Technology Assessment Consortium. J Gen Intern Med 1992 Mar;7(2):158-64.

89. Maas JW, Evers JL, ter RG, et al. Pregnancy rate following normal versus abnormal hysterosalpingography findings: a meta-analysis. Gynecol Obstet Invest 1997;43(2):79-83.

90. Mackenzie R, Palmer CR, Lomas DJ, et al. Magnetic resonance imaging of the knee: diagnostic performance studies. Clin Radiol 1996 Apr;51(4):251-7.

91. Makrydimas G, Sotiriadis A, Ioannidis JP. Screening performance of first-trimester nuchal translucency for major cardiac defects: a meta-analysis. Am J Obstet Gynecol 2003 Nov;189(5):1330-5.

92. Mantha S, Roizen MF, Barnard J, et al. Relative effectiveness of four preoperative tests for predicting adverse cardiac outcomes after vascular surgery: a meta-analysis. Anesth Analg 1994 Sep;79(3):422-33.

93. Mijnhout GS, Hoekstra OS, van Tulder MW, et al. Systematic review of the diagnostic accuracy of (18)F-fluorodeoxyglucose positron emission tomography in melanoma patients. Cancer 2001 Apr 15;91(8):1530-42.

94. Mirvis SE, Shanmuganathan K, Miller BH, et al. Traumatic aortic injury: diagnosis with contrast-enhanced thoracic CT--five-year experience at a major trauma center. Radiology 1996 Aug;200(2):413-22.

95. Mitchell MF, Schottenfeld D, Tortolero-Luna G, et al. Colposcopy for the diagnosis of squamous intraepithelial lesions: a meta-analysis. Obstet Gynecol 1998 Apr;91(4):626-31.

96. Mitchell MF, Cantor SB, Ramanujam N, et al. Fluorescence spectroscopy for diagnosis of squamous intraepithelial lesions of the cervix. Obstet Gynecol 1999 Mar;93(3):462-70.

97. Mol BW, Dijkman B, Wertheim P, et al. The accuracy of serum chlamydial antibodies in the diagnosis of tubal pathology: a meta-analysis. Fertil Steril 1997 Jun;67(6):1031-7.

98. Mol BW, Bayram N, Lijmer JG, et al. The performance of CA-125 measurement in the detection of endometriosis: a meta-analysis. Fertil Steril 1998 Dec;70(6):1101-8.

99. Mol BW, Lijmer JG, Ankum WM, et al. The accuracy of single serum progesterone measurement in the diagnosis of ectopic pregnancy: a meta-analysis. Hum Reprod 1998 Nov;13(11):3220-7.

100. Mol BW, Meijer S, Yuppa S et al. Sperm penetration assay in predicting successful in vitro fertilization. A meta-analysis. J Reprod Med 1998;43:503-8.

101. Nallamothu BK, Saint S, Bielak LF, et al. Electron-beam computed tomography in the diagnosis of coronary artery disease: a meta-analysis. Arch Intern Med 2001 Mar 26;161(6):833-8.

102. Nelemans PJ, Leiner T, de Vet HC, et al. Peripheral arterial disease: meta-analysis of the diagnostic performance of MR angiography. Radiology 2000 Oct;217(1):105-14.

103. Ng PC, Dear PR. The predictive value of a normal ultrasound scan in the preterm baby--a meta-analysis. Acta Paediatr Scand 1990 Mar;79(3):286-91.

104. Oei EH, Nikken JJ, Verstijnen AC, et al. MR imaging of the menisci and cruciate ligaments: a systematic review. Radiology 2003 Mar;226(3):837-48.

105. Oei SG, Helmerhorst FM, Keirse MJ. When is the post-coital test normal? A critical appraisal. Hum Reprod 1995 Jul;10(7):1711-4.

106. Olaniyan OB. Validity of colposcopy in the diagnosis of early cervical neoplasia--a review. Afr J Reprod Health 2002 Dec;6(3):59-69.

107. Olatidoye AG, Wu AH, Feng YJ, et al. Prognostic role of troponin T versus troponin I in unstable angina pectoris for cardiac events with meta-analysis comparing published studies. Am J Cardiol 1998 Jun 15;81(12):1405-10.

108. Orr RK, Porter D, Hartman D. Ultrasonography to evaluate adults for appendicitis: decision making based on meta-analysis and probabilistic reasoning. Acad Emerg Med 1995 Jul;2(7):644-50.

109. Pai M, Flores LL, Pai N, et al. Diagnostic accuracy of nucleic acid amplification tests for tuberculous meningitis: a systematic review and meta-analysis. Lancet Infect Dis 2003 Oct;3(10):633-43.

110. Palomaki GE, Neveux LM, Haddow JE. Can reliable Down's syndrome detection rates be determined from prenatal screening intervention trials? J Med Screen 1996;3(1):12-7.

111. Patel SR, Wiese W, Patel SC, et al. Systematic review of diagnostic tests for vaginal trichomoniasis. Infect Dis Obstet Gynecol 2000;8(5-6):248-57.

112. Patrick DL, Cheadle A, Thompson DC, et al. The validity of self-reported smoking: a review and meta-analysis. Am J Public Health 1994 Jul;84(7):1086-93.

113. Paulson WD, Ram SJ, Birk CG, et al. Does blood flow accurately predict thrombosis or failure of hemodialysis synthetic grafts? A meta-analysis. Am J Kidney Dis 1999 Sep;34(3):478-85.

114. Petersen JR, Smith E, Okorodudu AO, et al. Comparison of four methods (L/S ratio, TDx FLM, lamellar bodies, PG) for fetal lung maturity using meta-analysis. Clin Lab Manage Rev 1996 Mar;10(2):169-75.

115. Picano E, Bedetti G, Varga A, et al. The comparable diagnostic accuracies of dobutamine-stress and dipyridamole-stress echocardiographies: a meta-analysis. Coron Artery Dis 2000 Mar;11(2):151-9.

116. Rao JK, Weinberger M, Oddone EZ, et al. The role of antineutrophil cytoplasmic antibody (c-ANCA) testing in the diagnosis of Wegener granulomatosis. A literature review and meta-analysis. Ann Intern Med 1995 Dec 15;123(12):925-32.

117. Reed JF, III. Meta-analysis of the reliability of noninvasive carotid studies. Biomed Instrum Technol 1991 Nov;25(6):465-71.

118. Revah A, Hannah ME, Sue AQA. Fetal fibronectin as a predictor of preterm birth: an overview. Am J Perinatol 1998;15(11):613-21.

119. Romagnuolo J, Bardou M, Rahme E, et al. Magnetic resonance cholangiopancreatography: a meta-analysis of test performance in suspected biliary disease. Ann Intern Med 2003 Oct 7;139(7):547-57.

120. Rosado B, Menzies S, Harbauer A, et al. Accuracy of computer diagnosis of melanoma: a quantitative meta-analysis. Arch Dermatol 2003 Mar;139(3):361-7.

121. Rosenberg D, Cretin S. Use of meta-analysis to evaluate tolonium chloride in oral cancer screening. Oral Surg Oral Med Oral Pathol 1989 May;67(5):621-7.

122. Samson DJ, Flamm CR, Pisano ED, et al. Should FDG PET be used to decide whether a patient with an abnormal mammogram or breast finding at physical examination should undergo biopsy? Acad Radiol 2002 Jul;9(7):773-83.

123. Sarmiento OL, Weigle KA, Alexander J, et al. Assessment by meta-analysis of PCR for diagnosis of smear-negative pulmonary tuberculosis. J Clin Microbiol 2003 Jul;41(7):3233-40.

124. Scheidler J, Hricak H, Yu KK, et al. Radiological evaluation of lymph node metastases in patients with cervical cancer. A meta-analysis. JAMA 1997 Oct 1;278(13):1096-101.

125. Schinkel AF, Bax JJ, Geleijnse ML, et al. Noninvasive evaluation of ischaemic heart disease: myocardial perfusion imaging or stress echocardiography? Eur Heart J 2003 May;24(9):789-800.

126. Scholten RJ, Deville WL, Opstelten W, et al. The accuracy of physical diagnostic tests for assessing meniscal lesions of the knee: a meta-analysis. J Fam Pract 2001 Nov;50(11):938-44.

127. Scholten RJ, Opstelten W, van der Plas CG, et al. Accuracy of physical diagnostic tests for assessing ruptures of the anterior cruciate ligament: a meta-analysis. J Fam Pract 2003 Sep;52(9):689-94.

128. Schreiber G, McCrory DC. Performance characteristics of different modalities for diagnosis of suspected lung cancer: summary of published evidence. Chest 2003 Jan;123(1 Suppl):115S-28S.

129. Schwimmer J, Essner R, Patel A, et al. A review of the literature for whole-body FDG PET in the management of patients with melanoma. Q J Nucl Med 2000 Jun;44(2):153-67.

130. Siegman-Igra Y, Anglim AM, Shapiro DE, et al. Diagnosis of vascular catheter-related bloodstream infection: a meta-analysis. J Clin Microbiol 1997 Apr;35(4):928-36.

131. Silvestri GA, Littenberg B, Colice GL. The clinical evaluation for detecting metastatic lung cancer. A meta-analysis. Am J Respir Crit Care Med 1995 Jul;152(1):225-30.

132. Skupski DW, Rosenberg CR, Eglinton GS. Intrapartum fetal stimulation tests: a meta-analysis. Obstet Gynecol 2002 Jan;99(1):129-34.

133. Smart SC, Sagar KB. Diagnostic and prognostic use of stress echocardiography in stable patients. Echocardiography 2000 Jul;17(5):465-77.

134. Smith-Bindman R, Hosmer W, Feldstein VA, et al. Second-trimester ultrasound to detect fetuses with Down syndrome: a meta-analysis. JAMA 2001 Feb 28;285(8):1044-55.

135. Sonnad SS, Langlotz CP, Schwartz JS. Accuracy of MR imaging for staging prostate cancer: a meta-analysis to examine the effect of technologic change. Acad Radiol 2001 Feb;8(2):149-57.

136. Stengel D, Bauwens K, Sehouli J, et al. Systematic review and meta-analysis of emergency ultrasonography for blunt abdominal trauma. Br J Surg 2001 Jul;88(7):901-12.

137. Swart P, Mol BW, van d, V, et al. The accuracy of hysterosalpingography in the diagnosis of tubal pathology: a meta-analysis. Fertil Steril 1995 Sep;64(3):486-91.

138. Tan KT, van Beek EJ, Brown PW, et al. Magnetic resonance angiography for the diagnosis of renal artery stenosis: a meta-analysis. Clin Radiol 2002 Jul;57(7):617-24.

139. Tebas P, Nease RF, Storch GA. Use of the polymerase chain reaction in the diagnosis of herpes simplex encephalitis: a decision analysis model. Am J Med 1998 Oct;105(4):287-95.

140. Tsao H, Nadiminti U, Sober AJ, et al. A meta-analysis of reverse transcriptase-polymerase chain reaction for tyrosinase mRNA as a marker for circulating tumor cells in cutaneous melanoma. Arch Dermatol 2001 Mar;137(3):325-30.

141. Urbach DR, Khajanchee YS, Jobe BA, et al. Cost-effective management of common bile duct stones: a decision analysis of the use of endoscopic retrograde cholangiopancreatography (ERCP), intraoperative cholangiography, and laparoscopic bile duct exploration. Surg Endosc 2001 Jan;15(1):4-13.

142. van Beek EJ, Brouwers EM, Song B, Bongaerts AH, Oudkerk M. Lung scintigraphy and helical computed tomography for the diagnosis of pulmonary embolism: a meta-analysis. Clin Appl Thromb Hemost 2001;7:87-92.

143. van Gelder JM. Computed tomographic angiography for detecting cerebral aneurysms: implications of aneurysm size distribution for the sensitivity, specificity, and likelihood ratios. Neurosurgery 2003 Sep;53(3):597-605.

144. Varonen H, Makela M, Savolainen S, et al. Comparison of ultrasound, radiography, and clinical examination in the diagnosis of acute maxillary sinusitis: a systematic review. J Clin Epidemiol 2000 Sep;53(9):940-8.

145. Vasbinder GB, Nelemans PJ, Kessels AG, et al. Diagnostic tests for renal artery stenosis in patients suspected of having renovascular hypertension: a meta-analysis. Ann Intern Med 2001 Sep 18;135(6):401-11.

146. Vasquez TE, Rimkus DS, Hass MG, et al. Efficacy of morphine sulfate-augmented hepatobiliary imaging in acute cholecystitis. J Nucl Med Technol 2000 Sep;28(3):153-5.

147. Visser K, Hunink MG. Peripheral arterial disease: gadolinium-enhanced MR angiography versus color-guided duplex US--a meta-analysis. Radiology 2000 Jul;216(1):67-77.

148. Watson EJ, Templeton A, Russell I, et al. The accuracy and efficacy of screening tests for Chlamydia trachomatis: a systematic review. J Med Microbiol 2002 Dec;51(12):1021-31.

149. Wellnitz U, Binder B, Fritz P, et al. Reliability of telepathology for frozen section service. Anal Cell Pathol 2000;21(3-4):213-22.

150. Wells PS, Lensing AW, Davidson BL, et al. Accuracy of ultrasound for the diagnosis of deep venous thrombosis in asymptomatic patients after orthopedic surgery. A meta-analysis. Ann Intern Med 1995 Jan 1;122(1):47-53.

151. White RH, McGahan JP, Daschbach MM, et al. Diagnosis of deep-vein thrombosis using duplex ultrasound. Ann Intern Med 1989 Aug 15;111(4):297-304.

152. Whitsel EA, Boyko EJ, Siscovick DS. Reassessing the role of QTc in the diagnosis of autonomic failure among patients with diabetes: a meta-analysis. Diabetes Care 2000 Feb;23(2):241-7.

153. Wiese W, Patel SR, Patel SC, et al. A meta-analysis of the Papanicolaou smear and wet mount for the diagnosis of vaginal trichomoniasis. Am J Med 2000 Mar;108(4):301-8.

154. Wijnberger LD, Huisjes AJ, Voorbij HA, et al. The accuracy of lamellar body count and lecithin/sphingomyelin ratio in the prediction of neonatal respiratory distress syndrome: a meta-analysis. BJOG 2001 Jun;108(6):583-8.

155. Williams JW, Jr., Noel PH, Cordes JA, Ramirez G, Pignone M. Is this patient clinically depressed? JAMA

156. Wu AH, Lane PL. Metaanalysis in clinical chemistry: validation of cardiac troponin T as a marker for ischemic heart diseases. Clin Chem 1995 Aug;41(8 Pt 2):1228-33.

157. Zandbergen EG, de Haan RJ, Stoutenbeek CP, et al. Systematic review of early prediction of poor outcome in anoxic-ischaemic coma. Lancet 1998 Dec 5;352(9143):1808-12.

Appendix B. Bayesian Model for Bivariate Meta-Analysis of Sensitivity and Specificity

For each study we used the 2×2 table of test results (positive or negative) and true disease status (Table B-1).

Table B-1. 2×2 table of test results

		Disease status	
		Disease	**Healthy**
Test result	Positive	*TP*	*FP*
	Negative	*FN*	*TN*

Here, *TP*, *FP*, *FN*, and *TN* are respectively the counts of true positive, false positive, false negative, and true negative findings from each study. The accuracy metrics of interest were the logit-transformed sensitivity and false positive rate (=1-specificity); at the study-level these were estimated as $\hat{\eta} = logit(\frac{TP}{TP+FN})$, and $\hat{\xi} = logit(\frac{FP}{FP+TN})$, respectively.

Throughout the report, we use the logit transformation of a proportion (p): $logit(p) = \ln\left(\frac{p}{1-p}\right)$. The inverse-logit of any number (x) is then defined as $logit^{-1}(x) = \frac{\exp(x)}{1+\exp)(x)}$.

Bivariate meta-analysis of sensitivity and specificity

The within studies model is specified as $TP_i \sim Bin(D_i, logit^{-1}(\eta_i))$ and $FP_i \sim Bin(H_i, logit^{-1}(\xi_i))$, where D_i and H_i are the total numbers of affected and unaffected individuals, respectively, for the i-th study. We then model the "true" logit-transformed true positive rate (η_i) and false positive rate (ξ_i) of the i-th study as drawn from a bivariate normal distribution: $\binom{\eta_i}{\xi_i} \sim N\left(\binom{\bar{\eta}}{\bar{\xi}}, \Omega\right)$.

Here, $\bar{\eta}$ and $\bar{\xi}$ are the "summary" logit-transformed summary true and false positive rate (across studies), with covariance matrix $\Omega = \begin{pmatrix} \sigma_\eta^2 & \rho\sigma_\eta\sigma_\xi \\ \rho\sigma_\eta\sigma_\xi & \sigma_\xi^2 \end{pmatrix}$; σ_η^2 and σ_ξ^2 are the variances of the logit-transformed sensitivity and false positive rate, respectively; ρ is their correlation (at the between-study level); and $\rho\sigma_\eta\sigma_\xi$ is their covariance.

This model was used both for the maximum likelihood and the Bayesian bivariate meta-analyses. The key difference between the two approaches is that in the latter, the model parameters (e.g., the average true and false positive rates and the covariance matrix) are treated as random variables and prior distributions need to be specified for them (hyper-priors). For the average logit-transformed true and false positive rates we used vague normal priors $N(0,100)$; for the between-study variance we assumed that its square root (i.e., the standard deviation) followed a uniform prior distribution $U(0.01,5)$. Finally, we assumed that the within study correlation had a uniform prior distribution

$U(-1,1)$. We used alternative prior distributions both for the variances and the correlation. One set of analyses used a normal prior $N^{0,25}$ for the Fisher-transformed correlation coefficient across studies (as suggested in Paul et al, Statistics in Medicine, 2010). Another set of analyses used a Wishart prior for the covariance matrix (as suggested in Verde et al., Statistics in Medicine 2010).

BUGS language Bayesian model for bivariate meta-analysis of sensitivity and false positive rate

```
#BUGS model: bivariate meta-analysis of test accuracy
#exact binomial likelihood for within-study variability

model {
for( i in 1 : n_studies ) {
tp[i] ~ dbin(tpr[i], n1[i])
fp[i] ~ dbin(fpr[i], n2[i])
logit(tpr[i]) <- m[i,1]
logit(fpr[i]) <- m[i,2]
m[i,1:2] ~ dmnorm(mu0[1:2], sigma.inv[1:2, 1:2])
}
# Priors for means
mu0[1] ~ dnorm(0, 0.01)
mu0[2] ~ dnorm(0, 0.01)

#priors for the elements of the covariance matrix
sigma.inv[1:2, 1:2] <- inverse(sigma[1:2,1:2])
sigma[1,1] <- sigma1*sigma1
sigma[1,2] <- sigma1*sigma2*rho
sigma[2,1] <- sigma1*sigma2*rho
sigma[2,2] <- sigma2*sigma2
sigma1 ~ dunif(0.01,5)
sigma2 ~ dunif(0.01,5)
rho ~ dunif(-1, 1)

# Pooled summaries
x.pool <- mu0[1]
y.pool <- mu0[2]
poolse <- exp(x.pool) / ( 1 + exp(x.pool) )
poolsp <- 1 - exp(y.pool) / ( 1 + exp(y.pool) )

# Variance-covariance matrix for random-effects
sigmaSens <- sigma[1,1]
sigmaSpec <- sigma[2,2]
}
```

Appendix C. Alternative Parameterizations of the Hierarchical Summary Receiver Operating Characteristic Curve

We followed Arends 2008 [25] in constructing alternative hierarchical summary receiver operating (HSROC) curves based on the bivariate random effects meta-analysis model. Here, $\bar{\eta}$ is the summary logit-transformed true positive rate, $\bar{\xi}$ is the summary logit-transformed false positive rate (1 – specificity), η_i and ξ_i are respectively the logit-transformed true positive rate and the false positive rate in the i-th study; σ_η^2 and σ_ξ^2 are the variances of the logit-transformed true and false positive rates, respectively; $\sigma_{\xi\eta}$ is their covariance.

After fitting the bivariate model, we obtained the HSROC lines, as shown in Table C-1.

Table C-1. Alternative HSROC lines

Model	Regression line	Intercept	Slope
η ~ ξ regression	$\eta_i = \bar{\eta} + \frac{\sigma_{\xi\eta}}{\sigma_\xi^2}(\xi_i - \bar{\xi})$	$\bar{\eta} - \frac{\sigma_{\xi\eta}}{\sigma_\xi^2}\bar{\xi}$	$\frac{\sigma_{\xi\eta}}{\sigma_\xi^2}$
ξ ~ η regression	$\eta_i = \bar{\eta} + \frac{\sigma_\eta^2}{\sigma_{\xi\eta}}(\xi_i - \bar{\xi})$	$\bar{\eta} - \frac{\sigma_\eta^2}{\sigma_{\xi\eta}}\bar{\xi}$	$\frac{\sigma_\eta^2}{\sigma_{\xi\eta}}$
D ~ S regression	$\eta_i = \bar{\eta} + \frac{\sigma_\eta^2 + \sigma_{\xi\eta}}{\sigma_\xi^2 + \sigma_{\xi\eta}}(\xi_i - \bar{\xi})$	$\bar{\eta} - \frac{\sigma_\eta^2 + \sigma_{\xi\eta}}{\sigma_\xi^2 + \sigma_{\xi\eta}}\bar{\xi}$	$\frac{\sigma_\eta^2 + \sigma_{\xi\eta}}{\sigma_\xi^2 + \sigma_{\xi\eta}}$
Rutter-Gatsonis model	$\eta_i = \bar{\eta} + \frac{\sigma_\eta}{\sigma_\xi}(\xi_i - \bar{\xi})$	$\bar{\eta} - \frac{\sigma_\eta}{\sigma_\xi}\bar{\xi}$	$\frac{\sigma_\eta}{\sigma_\xi}$
MAR	$\eta_i = \bar{\eta} + \frac{\sigma_\eta^2 - \sigma_\xi^2 + \sqrt{(\sigma_\eta^2 - \sigma_\xi^2)^2 + 4\sigma_{\xi\eta}^2}}{2\sigma_{\xi\eta}}(\xi_i - \bar{\xi})$	$\bar{\eta} - \frac{\sigma_\eta^2 - \sigma_\xi^2 + \sqrt{(\sigma_\eta^2 - \sigma_\xi^2)^2 + 4\sigma_{\xi\eta}^2}}{2\sigma_{\xi\eta}}\bar{\xi}$	$\frac{\sigma_\eta^2 - \sigma_\xi^2 + \sqrt{(\sigma_\eta^2 - \sigma_\xi^2)^2 + 4\sigma_{\xi\eta}^2}}{2\sigma_{\xi\eta}}$

D = logit(sensitivity) – logit(1-specificity)
S = logit(sensitivity) + logit(1-specificity)
MAR = major axis regression
η = logit(sensitivity)
ξ = logit(1-specificity).
"Intercept" and "slope" refer to the estimated intercept and slope of the hsROC in the logit space. They describe a line that can be transformed to the HSROC curve. For each model, the corresponding hsROC curve is represented by the following function: sensitivity = invlogit(intercept + slope * logit(specificity)). Estimates of the intercept and slope can be derived from the output of the bivariate regression model in all major statistical packages.

Appendix D. Worked Meta-Analysis Example

Here we present a worked meta-analysis example using several of the methods we employed in the report. In addition to the results presented in the main text of the report, we present additional model diagnostics for maximum likelihood and Bayesian methods. Our data were derived from Table 1 of Arends et al. (Med Decis Making, 2008) and pertain to the test performance of aspiration cytologic examination of the breast for the detection of cancer. The original source of the data is Giard and Hermans (Cancer, 1992). We reproduce the data in Table D-1.

Table D-1. Example meta-analysis data

TP	FP	FN	TN
979	70	89	939
51	3	22	163
1569	55	152	894
35	25	15	259
59	4	12	121
56	18	4	216
329	602	39	3117
125	10	17	213
211	88	63	499
49	0	1	31
336	26	178	643
210	147	42	746
16	5	3	25
258	16	53	356
56	9	18	107
162	16	28	112
116	6	13	112
65	99	12	145
94	5	10	78
26	0	4	70
1318	28	249	136
569	55	120	539
46	1	16	287
64	13	6	76
39	1	4	104
132	16	20	426
470	17	22	161
28	25	4	200
42	43	3	22

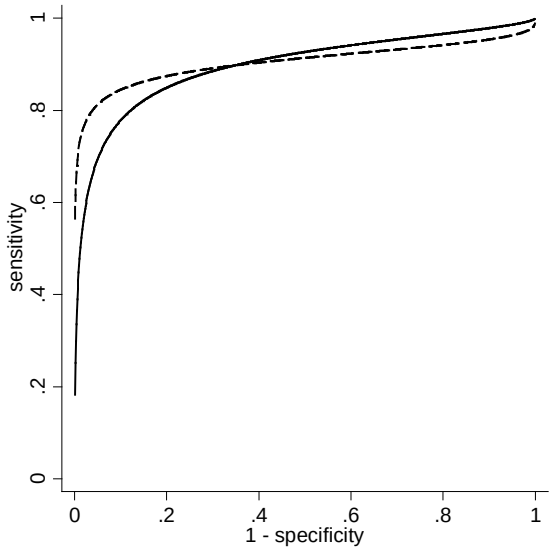
We used Stata version IC/12 (Stata Corp., College Station, TX) to implement all non-Bayesian analyses presented in the report; nonetheless, code should be easy to translate into other statistical packages that provide maximum likelihood implementations for general and generalized linear models. To run our code users will need a version of Stata that includes packages for mixed effects generalized linear models (version 10 or later), as well as the user-contributed packages metan, metandi, metareg, and mvmeta. Our data is saved in a file named example.dta.

```
/* enter the data */
use example.dta , clear

/* some data transformations */
/* logit sensitivity, with continuity correction */
generate b1 = logit(TP/(TP + FN))
replace b1 = logit((TP+0.5)/(TP+FN+2*0.5)) if TN == 0 | FP == 0 | TP == 0 | FN == 0
generate V11 = 1/TP + 1/FN if TP != 0 & FN != 0
replace V11 = 1/(TP+0.5) + 1/(FN+0.5) if TN == 0 | FP == 0 | TP == 0 | FN == 0
generate se_b1 = sqrt(V11)

/* logit specificity, with continuity correction */
generate b2 = logit(TN/(TN+FP))
replace b2 = logit((TN+0.5)/(TN+FP+2*0.5)) if TN == 0 | FP == 0 | TP == 0 | FN == 0
generate V22 = 1/TN + 1/FP
replace V22 = 1/(TN + 0.5) + 1/(FP + 0.5) if TN == 0 | FP == 0 | TP == 0 | FN == 0
generate se_b2 = sqrt(V22)

/*********************************************************/
/* univariate meta-analyses of sensitivity and specificity */
/*********************************************************/
/* normal within-study likelihood */
/* FE inverse variance */
metan b1 se_b1 , fixedi nograph z /* summary logit sensitivity */
metan b2 se_b2 , fixedi nograph z /* summary logit specificity */

/* normal within-study likelihood */
/* DerSimonian-Laird method for RE*/
metan b1 se_b1 , randomi nograph z /* summary logit sensitivity */
metan b2 se_b2 , randomi nograph z /* summary logit specificity */

/* normal within-study likelihood */
/* REML estimation for RE */
metareg b1 , wsse(se_b1) reml z /* summary logit sensitivity */
metareg b2 , wsse(se_b2) reml z /* summary logit specificity */

/* exact within study likelihood */
/* ML for random effects */
generate id = _n /* study id */
generate disease = TP + FN /* total individuals with disease */
xtmelogit TP || id: , binomial(disease) intp(5) /* intercept only model */

/* exact within study likelihood */
/* ML for random effects */
generate healthy = FP + TN /* total individuals without disease */
xtmelogit TN || id: , binomial(healthy)  intp(5) /* intercept only model */
```

```
/*********************************************************/
/* bivariate meta-analyses of sensitivity and specificity */
/*********************************************************/
/* normal within-study likelihood */
/* generalized DerSimonian-Laird method for RE*/
mvmeta b V, corr(0) mm /* the within-study correlation is set to zero */

/* normal within-study likelihood */
/* REML estimation for RE */
mvmeta b V, corr(0) reml /* the within-study correlation is set to zero */

/* exact within study likelihood */
/* ML for random effects */
metandi TP FP FN TN



/**************/
/* ROC curves */
/**************/

/* sROC unweighted */
generate D = b1 - (1 - b2) /* logit sensitivity minus FPR */
generate S = b1 + (1 - b2) /* logit sensitivity plus FPR */
regress D S /*unweighted Moses-Littenberg model*/

/*save estimates to obtain the graph*/
matrix estimates = e(b)
 local beta_sroc_unweighted = estimates[1,1]
 local alpha_sroc_unweighted = estimates[1,2]
 local a_un = `alpha_sroc_unweighted'/(1 - `beta_sroc_unweighted')
 local b_un = (1 + `beta_sroc_unweighted') / (1 - `beta_sroc_unweighted')

/* sROC weighted */
/* obtain the weights, with continuity correction if needed */
generate se = sqrt(1/TP + 1/FP + 1/FN + 1/TN)
replace se = sqrt(1/(TP+0.5) + 1/(FP+0.5) + 1/(FN+0.5) + 1/(TN+0.5)) if se == .
vwls D S, sd(se) /*weighted Moses-Littenberg model*/

matrix estimates = e(b)
 local beta_sroc_weighted = estimates[1,1]
 local alpha_sroc_weighted = estimates[1,2]
 local a_w = `alpha_sroc_weighted'/(1 - `beta_sroc_weighted')
 local b_w = (1 + `beta_sroc_weighted') / (1 - `beta_sroc_weighted')

/* joint graph for comparison */
graph two ( function y = invlogit(`a_un' + `b_un' * logit(x)) , ///
           lcol(black) lpat(dash) n(1000) range(0 1) ) ///
          ( function y = invlogit(`a_w' + `b_w' * logit(x)) , ///
           lcol(black) n(1000) range(0 1) ) ///
          ||, ylabel(0 0.2 0.4 0.6 0.8 1.0) ///
           aspectratio(1) scheme(s1mono) ///
           plotregion(style(none)) ///
           xtitle("1 - specificity") ///
           ytitle("sensitivity") ///
           legend(off)
```

After the last command, we obtain Figure D-1.

Figure D-1. SROC curves for the example in Table D-1



The dashed curve is derived from the unweighted analysis. The solid line is derived from the weighted analysis.

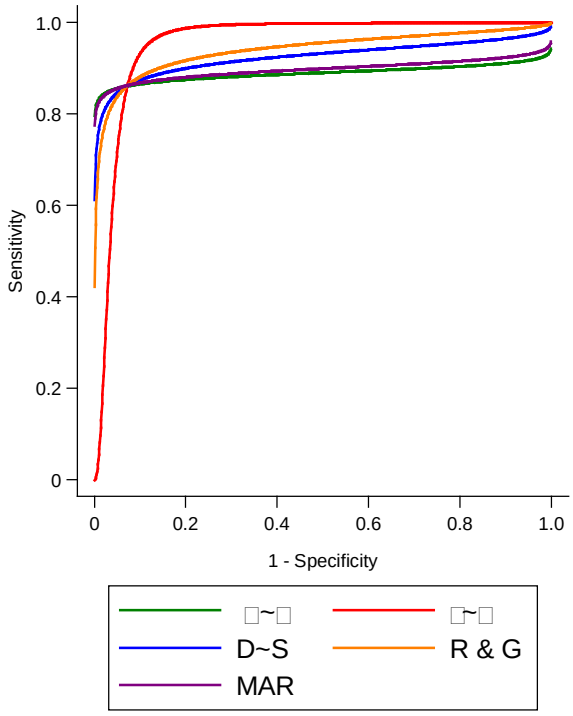
```
/* use mixed effects logistic regression to fit the bivariate model */
/* then use the estimates to fit different curves as in Arends et al. 2008*/

use example.dta , clear

generate persons = _n
generate n1 = TP + FN
generate n0 = TN + FP
generate detect1 = TP
generate detect0 = TN
reshape long n detect, i(persons) j(d1)
generate d0 = -(1 - d1 ) /* data transformation to replicate analyses in Arends et al. 2008 */

/* fit the bivariate model */
xtmelogit detect d1 d0 , nocons  || persons: d1 d0, ///
 nocons covariance(un) ///
 binomial(n) diff intp(10) refineopts(iterate(3))
 matrix estimates = e(b)
 matrix variances = e(V)

 local mean_se = estimates[1,1]
 local mean_sp = estimates[1,2]

nlcom exp(2 * [lns1_1_1]_b[_cons])
 matrix var_mean_se = r(b)
 local var_mean_se = var_mean_se[1,1]
 nlcom exp(2 * [lns1_1_2]_b[_cons])
 matrix var_mean_sp = r(b)
 local var_mean_sp = var_mean_sp[1,1]

nlcom exp([lns1_1_1]_b[_cons]) * exp([lns1_1_2]_b[_cons]) * tanh([atr1_1_1_2]_b[_cons])
 matrix estimates = r(b)
 local cov_se_sp = estimates[1,1]


/* now obtain and store estimates for the 5 ROC curves */
```

```
/* eta on ksi */
nlcom  (exp([lns1_1_1]_b[_cons]) * exp([lns1_1_2]_b[_cons]) * tanh([atr1_1_1_2]_b[_cons])) /
exp(2 * [lns1_1_2]_b[_cons])
    matrix estimate = r(b)
    local beta_eta_on_ksi = estimate[1,1]
nlcom   _b[d1] - exp([lns1_1_1]_b[_cons]) * exp([lns1_1_2]_b[_cons])  *
tanh([atr1_1_1_2]_b[_cons]) / exp(2 * [lns1_1_2]_b[_cons])  * _b[d0]
    matrix estimate = r(b)
    local alpha_eta_on_ksi = estimate[1,1]

/*ksi on eta*/
nlcom  exp(2 * [lns1_1_1]_b[_cons]) / ((exp([lns1_1_1]_b[_cons]) * exp([lns1_1_2]_b[_cons])  *
tanh([atr1_1_1_2]_b[_cons]))  )
    matrix estimate = r(b)
    local beta_ksi_on_eta = estimate[1,1]
nlcom   _b[d1] - exp(2 * [lns1_1_1]_b[_cons])  / (exp([lns1_1_1]_b[_cons]) *
exp([lns1_1_2]_b[_cons])  * tanh([atr1_1_1_2]_b[_cons])) * _b[d0]
    matrix estimate = r(b)
    local alpha_ksi_on_eta = estimate[1,1]

/* D on S */
nlcom (exp(2 * [lns1_1_1]_b[_cons]) + (exp([lns1_1_1]_b[_cons]) *
exp([lns1_1_2]_b[_cons])*tanh([atr1_1_1_2]_b[_cons])) ) / (exp(2 * [lns1_1_2]_b[_cons])  +
(exp([lns1_1_1]_b[_cons]) * exp([lns1_1_2]_b[_cons]) * tanh([atr1_1_1_2]_b[_cons])))
    matrix estimate = r(b)
    local beta_d_on_s = estimate[1,1]
nlcom  _b[d1]  - (exp(2 * [lns1_1_1]_b[_cons]) + (exp([lns1_1_1]_b[_cons]) *
exp([lns1_1_2]_b[_cons])  * tanh([atr1_1_1_2]_b[_cons])) ) / ( exp(2 * [lns1_1_2]_b[_cons])  +
(exp([lns1_1_1]_b[_cons]) * exp([lns1_1_2]_b[_cons])  * tanh([atr1_1_1_2]_b[_cons])))* _b[d0]
    matrix estimate = r(b)
    local alpha_d_on_s = estimate[1,1]

/* R & G */
nlcom sqrt(exp(2 * [lns1_1_1]_b[_cons]))/sqrt(exp(2 * [lns1_1_2]_b[_cons]) )
    matrix estimate = r(b)
    local beta_r_g = estimate[1,1]
nlcom _b[d1] - sqrt(exp(2 * [lns1_1_1]_b[_cons]))/sqrt(exp(2 * [lns1_1_2]_b[_cons]))* _b[d0]
    matrix estimate = r(b)
    local alpha_r_g = estimate[1,1]

/* MAR */
nlcom  ( exp(2 * [lns1_1_1]_b[_cons])  - exp(2 * [lns1_1_2]_b[_cons])  + sqrt( (exp(2 *
[lns1_1_1]_b[_cons]) - exp(2 * [lns1_1_2]_b[_cons]) )^2 + 4*(exp([lns1_1_1]_b[_cons]) *
exp([lns1_1_2]_b[_cons]) * tanh([atr1_1_1_2]_b[_cons]))^2 ))/( 2*exp([lns1_1_1]_b[_cons]) *
exp([lns1_1_2]_b[_cons]) * tanh([atr1_1_1_2]_b[_cons]) )
    matrix estimate = r(b)
    local beta_mar = estimate[1,1]

nlcom _b[d1] - _b[d0]*((exp(2*[lns1_1_1]_b[_cons])-
exp(2*[lns1_1_2]_b[_cons])+sqrt((exp(2*[lns1_1_1]_b[_cons])-exp(2*[lns1_1_2]_b[_cons]))^2 +
4*(exp([lns1_1_1]_b[_cons])*exp([lns1_1_2]_b[_cons])*tanh([atr1_1_1_2]_b[_cons]))^2)) /
(2*exp([lns1_1_1]_b[_cons]) * exp([lns1_1_2]_b[_cons])  * tanh([atr1_1_1_2]_b[_cons])))
    matrix estimate = r(b)
    local alpha_mar = estimate[1,1]
```

```
/* plot the curves */
/* ROC space */
graph two (function y = invlogit(`alpha_eta_on_ksi' + `beta_eta_on_ksi'*logit(x)) ///
, lcol(green) range(0 1) n(1000)  )  ///
  (function y = invlogit(`alpha_ksi_on_eta' + `beta_ksi_on_eta'*logit(x)) , lcol(red) range(0 1)
n(1000))  ///
  (function y = invlogit(`alpha_d_on_s' + `beta_d_on_s'*logit(x)), lcol(blue) range(0 1) n(1000)
)  ///
  (function y = invlogit(`alpha_r_g' + `beta_r_g'*logit(x)), lcol(orange) range(0 1) n(1000)  )
///
  (function y = invlogit(`alpha_mar' + `beta_mar'*logit(x)), lcol(purple) range(0 1) n(1000)  )
///
  ||, scheme(s1mono) plotregion(style(none))  ///
    aspectratio(1) ///
    xtitle(" " "1 - Specificity" , size(*0.7)) ytitle(" " "Sensitivity" , size(*0.7)) ///
    xlabel(0 "0" 0.2 "0.2" 0.4 "0.4" 0.6 "0.6" 0.8 "0.8" 1 "1.0" , labsize(*0.7)) ///
    ylabel(0 "0" 0.2 "0.2" 0.4 "0.4" 0.6 "0.6" 0.8 "0.8" 1 "1.0" , angle(0) labsize(*0.7)) ///
    legend( lab(1 " {&eta}~{&xi}") lab(2 "{&xi}~{&eta}") lab(3 "D~S") lab(4 "R & G") lab(5
"MAR"))
```

After the last command, we obtain Figure D-2.

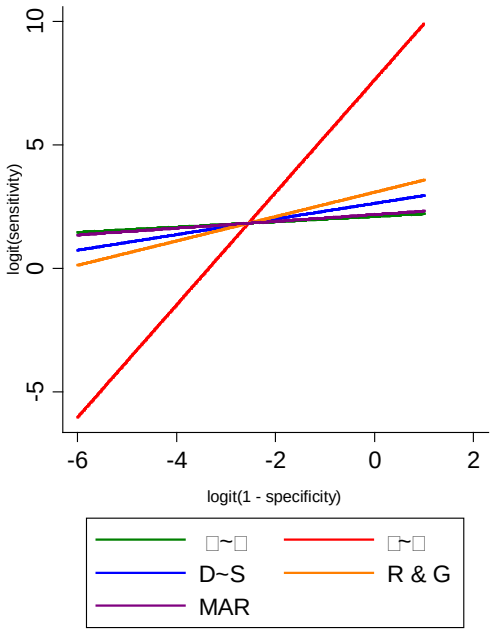
Figure D-2. Alternative HSROC curves for the example in Table D-1



```
/* logit space */
graph two (function y = (`alpha_eta_on_ksi' + `beta_eta_on_ksi'*(x)) , lcol(green) range(-6 1)
n(1000)  )  ///
  (function y = (`alpha_ksi_on_eta' + `beta_ksi_on_eta'*(x)) , lcol(red) range(-6 1) n(1000)  )  ///
  (function y = (`alpha_d_on_s' + `beta_d_on_s'*(x)) , lcol(blue) range(-6 1) n(1000)  )  ///
  (function y = (`alpha_r_g' + `beta_r_g'*(x)) , lcol(orange) range(-6 1) n(1000)  )  ///
  (function y = (`alpha_mar' + `beta_mar'*(x)) , lcol(purple) range(-6 1) n(1000)  )  ///
  ||, scheme(s1mono) plotregion(style(none))  ///
     xtitle(" " "logit(1 - specificity)" , size(*0.7)) ytitle(" " "logit(sensitivity)" , size(*0.7)) ///
     legend( lab(1 " {&eta}~{&xi}") lab(2 "{&xi}~{&eta}") lab(3 "D~S") lab(4 "R & G") lab(5
"MAR") ) ///
     aspectratio(1)
```

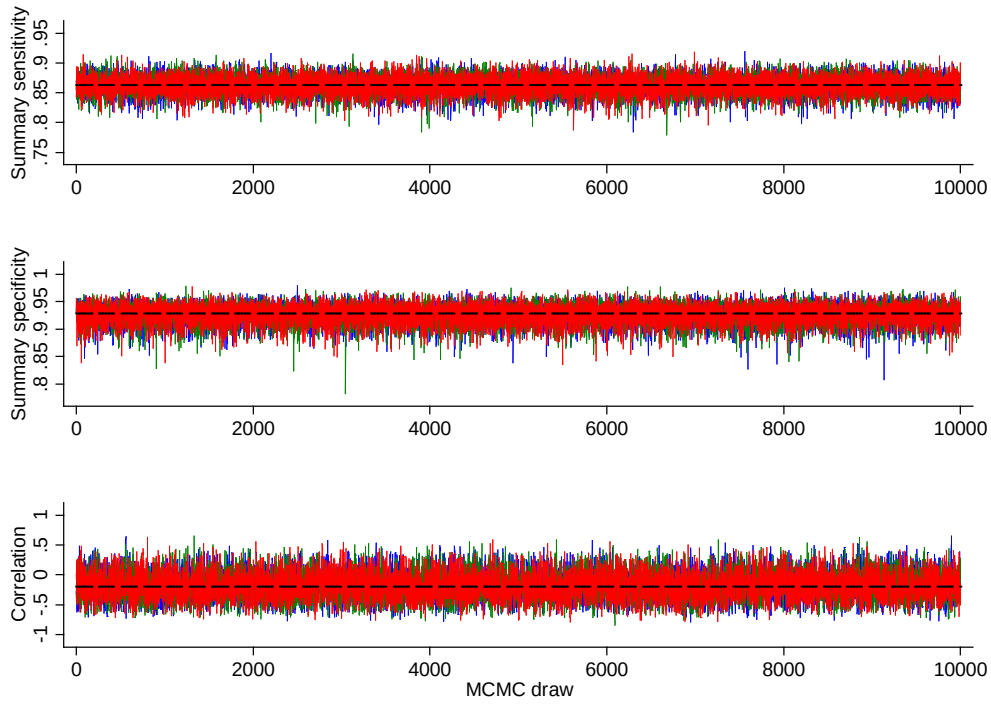
After the last command, we obtain Figure D-3.

Figure D-3. Alternative HSROC curves for the example in Table D-1 (logit space)



For Bayesian analyses, users should use the model presented in Appendix B. In this specific example, for the last 10,000 iterations (of the 20,000 run as burn-in) for three chains initialized using different starting values, we obtained the following trace plots for logit-sensitivity, logit-specificity, and the between-study correlation (Figure D-4; of course other parameters need to be monitored as well).

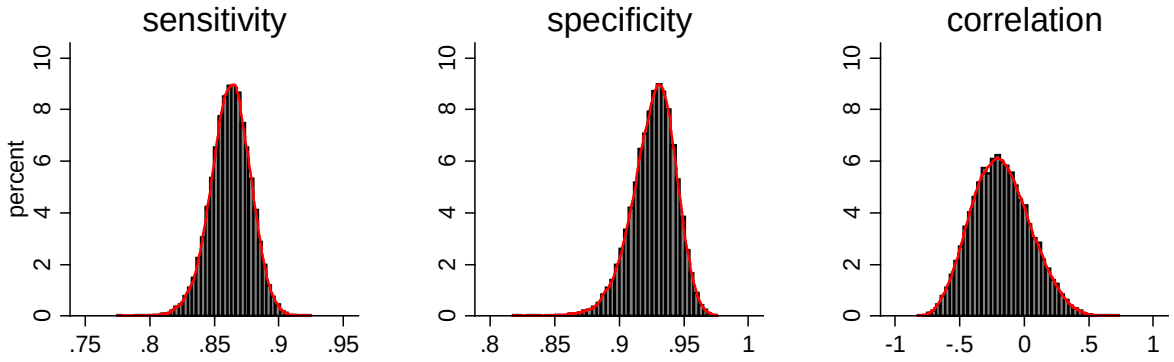
Figure D-4. Trace plots for summary sensitivity, specificity, and correlation



Dashed lines indicate the medians of the posterior distributions.

Convergence was also assessed with the Gelman-Rubin diagnostic. In this example, after 20,000 iterations, the final median values for the statistic were 0.99< R <1.01, for logit-sensitivity, logit-specificity, the between-study correlation, and the between-study variances of sensitivity and specificity, indicating that the model had converged. After convergence, we run the model for an additional 10,000 iterations and used the results to obtain density plots and summary statistics for the parameters of interest. For example, we obtained the following density plots for the posterior distributions of the summary sensitivity, specificity, and their correlation (Figure D-5).

Figure D-5. Posterior densities for sensitivity, specificity, and correlation



Red lines are kernel densities.

For comparison, summary results from all meta-analysis methods for this example are summarized in Table D-2.

Table D-2. Summary results from all meta-analysis methods (for the example in Table D1)

Model	Estimation (within-study likelihood)	Logit-sensitivity (95 CI or CrI*)	Logit-specificity (95% CI or CrI*)
Univariate, FE	IV (normal)	1.708 (1.646, 1.770)	1.855 (1.797, 1.912)
Univariate, RE	DL (normal)	1.803 (1.561, 2.044)	2.354 (2.049, 2.659)
Univariate, RE	REML (normal)	1.799 (1.569, 2.029)	2.422 (2.017, 2.828)
Univariate, RE	ML (binomial)	1.840 (1.607, 2.074)	2.556 (2.110, 3.002)
Bivariate, RE	multivariate DL (normal)	1.805 (1.564, 2.047)	2.352 (2.048, 2.657)
Bivariate, RE	REML (normal)	1.801 (1.569, 2.033)	2.416 (2.005, 2.827)
Bivariate, RE	ML (binomial)	1.839 (1.605, 2.072)	2.547 (2.104, 2.990)
Bivariate, RE	Fully Bayesian	1.840 (1.589, 2.105)	2.558 (2.080, 3.076)

CI = confidence interval; CrI = credibility interval; DL = DerSimonian-Laird; FE = fixed effect; IV = inverse variance; ML = maximum likelihood; RE = random effects; REML = restricted maximum likelihood;
*Credibility intervals are presented for Bayesian analyses.