BMJ Open

Empirical comparisons of meta-analysis methods for diagnostic studies: a meta-epidemiological study

Kristine J Rosenberger,[1] Haitao Chu,[2,3] Lifeng Lin [1]

To cite: Rosenberger KJ, Chu H, Lin L. Empirical comparisons of meta-analysis methods for diagnostic studies: a meta-epidemiological study. *BMJ Open* 2022;**12**:e055336. doi:10.1136/bmjopen-2021-055336

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-055336).

Received 11 July 2021
Accepted 15 April 2022

[1]Department of Statistics, Florida State University, Tallahassee, Florida, USA
[2]Statistical Research and Innovation, Global Biometrics and Data Management, Pfizer Inc, New York, New York, USA
[3]Division of Biostatistics, University of Minnesota School of Public Health, Minneapolis, Minnesota, USA

Correspondence to
Dr Lifeng Lin; linl@stat.fsu.edu

ABSTRACT

Objectives Several methods are commonly used for meta-analyses of diagnostic studies, such as the bivariate linear mixed model (LMM). It estimates the overall sensitivity, specificity, their correlation, diagnostic OR (DOR) and the area under the curve (AUC) of the summary receiver operating characteristic (ROC) estimates. Nevertheless, the bivariate LMM makes potentially unrealistic assumptions (ie, normality of within-study estimates), which could be avoided by the bivariate generalised linear mixed model (GLMM). This article aims at investigating the real-world performance of the bivariate LMM and GLMM using meta-analyses of diagnostic studies from the Cochrane Library.

Methods We compared the bivariate LMM and GLMM using the relative differences in the overall sensitivity and specificity, their 95% CI widths, between-study variances, and the correlation between the (logit) sensitivity and specificity. We also explored their relationships with the number of studies, number of subjects, overall sensitivity and overall specificity.

Results Among the extracted 1379 meta-analyses, point estimates of overall sensitivities and specificities by the bivariate LMM and GLMM were generally similar, but their CI widths could be noticeably different. The bivariate GLMM generally produced narrower CIs than the bivariate LMM when meta-analyses contained 2–5 studies. For meta-analyses with <100 subjects or the overall sensitivities or specificities close to 0% or 100%, the bivariate LMM could produce substantially different AUCs, DORs and DOR CI widths from the bivariate GLMM.

Conclusions The variation of estimates calls into question the appropriateness of the normality assumption within individual studies required by the bivariate LMM. In cases of notable differences presented in these methods' results, the bivariate GLMM may be preferred.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ Two commonly used methods for meta-analyses of diagnostic studies, that is, the bivariate linear mixed model and bivariate generalised linear mixed model, were empirically compared.
- ⇒ This empirical study is based on a large-scale database containing 1379 meta-analyses of diagnostic studies from the Cochrane Library.
- ⇒ We assessed overall sensitivity and specificity, their 95% CI widths, between-study variances, and the correlation between the (logit) sensitivity and specificity produced by the two meta-analysis methods.
- ⇒ We investigated the impact of the number of studies, number of subjects, overall sensitivity, and overall specificity on the difference between the two meta-analysis methods.
- ⇒ This study has been restricted to two commonly used methods for meta-analyses of diagnostic studies, while alternative methods (eg, Bayesian methods) exist in the literature of evidence synthesis.

INTRODUCTION

Diagnostic tests are commonly implemented in clinical settings to confirm or rule out conditions and diseases. Systematic reviews and meta-analyses are widely used to combine results from multiple diagnostic test accuracy studies. The reporting guidelines of systematic reviews and meta-analyses of diagnostic accuracy studies have been well established.[1–4] The results of diagnostic test accuracy studies are frequently reported as a pair of sensitivity and specificity.[5] Sensitivity (ie, the true positive fraction) is the conditional probability of a positive test in diseased subjects, and specificity (ie, the true negative fraction) is the conditional probability of a negative test in non-diseased subjects. In general, increasing sensitivity decreases specificity and vice versa.

Several tools beyond sensitivity and specificity are also used to summarise results from diagnostic studies. The diagnostic OR (DOR) measures how the odds of a positive test result differ among patients with the disease condition compared with patients without the condition.[6 7] It combines sensitivity and specificity into a single measurement. However, it cannot distinguish between the overall sensitivity and specificity, which are essential measurements in clinical settings. Thus, it is considered more difficult to interpret than sensitivity or specificity alone. The receiver operating characteristic (ROC) curve is another popular tool, which plots sensitivity versus specificity (or 1–specificity). The area under the curve (AUC) can be subsequently calculated to measure the performance of a diagnostic test for identifying diseased and non-diseased populations.[8]

Traditionally, the summary ROC (SROC) approach was the standard method for meta-analyses of diagnostic studies when both sensitivity and specificity are available.[9–15] The SROC approach converts each pair of sensitivity and specificity into the log DOR and regresses it on the logit difference between sensitivity and specificity.[6] This approach has several drawbacks and is no longer commonly used. It requires an ad hoc continuity correction for zero cell counts, and the independent variable in the SROC model is subject to measurement error; not taking this into account can lead to bias in threshold and accuracy parameters.[16] Additionally, the conventional SROC is a fixed-effect model and assumes model parameters do not vary across studies; this can underestimate the SE and produce biased estimates when between-study heterogeneity is present. Substantial heterogeneity may exist in diagnostic test accuracy studies;[17] it is thus of interest to employ random-effects models for combining these studies. The SROC model may be either weighted or unweighted. Weighting is typically implemented via the inverse of variance; this is not optimal when the between-study variance is large.

The bivariate linear mixed model (LMM) approach, an improvement and extension of the SROC approach, has been proposed to preserve the two-dimensional nature of diagnostic data testing.[18 19] The bivariate approach is more intuitive than the SROC for the comparison of multiple diagnostic tests; it directly models sensitivity and specificity and can produce multiple summary statistics such as the AUC, DOR and SROC curves. It is advantageous to the SROC approach as it can estimate both sensitivity and specificity, their correlation, their 95% CIs, and the amount of between-study variation. Additionally, the bivariate LMM can adjust for study-level covariates.[18] However, it still requires an ad hoc continuity correction for zero cell counts as in the SROC approach, and the logit sensitivity and specificity within each study are approximated to normal distribution.[20] These assumptions might not be valid in certain cases, such as small sample sizes or rare events.

The bivariate LMM was further improved by Chu and Cole[20] to avoid the unnecessary normality assumption within studies. Specifically, the diagnostic data can feasibly be analysed via a bivariate generalised linear mixed model (GLMM), which uses binomial distributions to model the counts of true positives and true negatives. In instances with sparse data or small sample sizes, the bivariate GLMM provides an unbiased estimate that is more efficient than the bivariate LMM.[21] This approach does not require the logit transformation or the assumption that logit sensitivity and specificity approximately follow normal distributions.[22] Additionally, the binomial distribution is better equipped for modelling within-study heterogeneity.[23] However, the bivariate GLMM can lead to convergence problems during the parameter estimation, particularly in cases of a large number of parameters or a limited number of studies.

As an improvement to the SROC model, Rutter and Gatsonis[24] derived a hierarchal SROC (HSROC) model that allows for both between-study and within-study variations to partly address the SROC model's shortcomings. The HSROC model and the bivariate random-effects meta-analysis model are closely related; therefore, this article will not consider the HSROC model. Specifically, Harbord *et al*[21] showed that the HSROC model without covariates affecting accuracy and threshold is equivalent to a bivariate model without covariates for sensitivity and specificity. However, there is no unique definition of an SROC across multiple studies with different accuracies.[25 26] Because the ROC curve measures accuracy along with varying thresholds, SROC curves may not be interpreted as the ROC across all studies without further assumptions.[16]

Another commonly used method in practice is the univariate meta-analysis. This method separately combines sensitivity and specificity, and it ignores the correlation between the two measures. It may be performed under either the fixed-effects or random-effects setting. The univariate model is simpler than the bivariate models, and it may avoid computation problems in meta-analyses of few studies or sparse data. An equally powerful alternative is to use an HSROC model that assumes a symmetrical SROC curve.[27] Furthermore, it has been shown that bivariate and univariate measures produce similar estimates of likelihood ratios with differences that are not sufficiently large enough to alter clinical decisions, and that the bivariate measure with no continuity correction for zero-event studies is more robust than the univariate model.[28] A simulation study further concludes that bivariate random-effects meta-analyses return superior point estimates with smaller standard errors than univariate analyses.[29]

It is generally recognised that hierarchical models outperform simple combining methods. Furthermore, the exact method may produce less biased estimates than approximate methods, especially in instances with small samples sizes, large between-study variance and large overall sensitivity.[30] Nevertheless, the bivariate LMM remains popular in current meta-analyses of diagnostic studies. Reitsma and Zwinderman[31] recommend making use of both the bivariate GLMM and LMM. The clinical differences between estimates produced by the two models may be insignificant, and it is important to identify when the within-study normal approximation by the bivariate LMM performs poorly.

To address these research gaps, this article empirically compares the differences in estimates produced by the bivariate LMM and GLMM methods among a large database of meta-analyses of diagnostic studies in the Cochrane Library. We aim to identify how the inclusion of exact event count affects diagnostic results and explore how the differences in results are related to the size of a meta-analysis and the overall sensitivity and specificity.

METHODS

Data sources

We extracted meta-analyses of diagnostic studies from systematic reviews published from 2003 Issue 1 to 2020 Issue 1 in the Cochrane Library. All reviews with statistical data were included, and the withdrawn reviews were excluded. A similar search strategy has been detailed in our earlier work on Cochrane meta-analyses of comparative intervention studies.[32] The extracted data were the counts of true positives, true negatives, false positives and false negatives for each study within each meta-analysis. Studies reporting no subjects (ie, studies that were included in the Cochrane Library but for which meta-analyses were not run) were removed prior to running the meta-analysis models.

Statistical analyses

We applied both the bivariate LMM and GLMM methods to each extracted Cochrane meta-analysis using the R (V.4.0.3) package 'altmeta' (V.3.3). The Supplementary Material presents the details of these methods. For both methods, we calculated the AUC of the SROC, estimated the DOR with its 95% CI, overall sensitivity and specificity with their 95% CIs, their between-study variances, and the correlation coefficient ρ between logit sensitivity and specificity. The implementation of some meta-analysis methods could fail in some meta-analyses (eg, due to the parameter estimation algorithm's non-convergence); such meta-analyses were excluded from our final analyses.

We compared the point estimates of the AUC, DOR, overall sensitivity and overall specificity using the relative differences. The relative difference was calculated by dividing the estimate produced by the bivariate LMM approach over the corresponding estimate produced by the bivariate GLMM. The bivariate GLMM is the recommended method for Cochrane reviews and was thus treated as the reference method in our analyses.[33] Of note, as we did not aim to perform a simulation study and did not know the true parameters of the bivariate GLMM and LMM, the relative difference between estimates is not synonymous with bias. To compare interval estimates of the DOR, overall sensitivity and overall specificity, we calculated the relative difference in interval length. A relative difference greater than 1.0 indicates that the bivariate GLMM produces a smaller estimate. To compare the point estimate of ρ, we calculated the absolute difference between the bivariate LMM and GLMM.

Table 1 Methods used in the original analyses as reported in the Cochrane systematic reviews

Method	Count (%)
Bivariate LMM	44 (39.29%)
Bivariate GLMM	25 (22.32%)
HSROC	32 (28.57%)
Univariate	12 (10.71%)
Meta-analysis not performed	5 (4.46%)
Not listed	9 (8.04%)

The HSROC and bivariate LMM methods have been shown to be equivalent in cases of no covariates. A Cochrane review might use more than one method.
GLMM, generalised linear mixed model; HSROC, hierarchal summary receiver operating characteristics; LMM, linear mixed model.

Patient and public involvement

This study did not have patient and public involvement because it focused on statistical methods for meta-analyses of diagnostic studies. All analyses were performed based on published data in the literature.

RESULTS

Basic characteristics

We identified 1379 meta-analyses with full data sets. The bivariate LMM and GLMM failed to achieve convergence when implementing the algorithm for the parameter estimation in 38 and 7 meta-analyses, respectively; these meta-analyses were excluded from our final analysis. Thus, the final analysis included 1334 meta-analyses with a total of 11 007 studies and 12 924 404 subjects. The number of studies in the meta-analyses ranged from 2 to 116 (median=4, IQR=2–9). The number of subjects in the meta-analyses ranged from 34 to 1173 853 (median=1132, IQR=451–3481).

Table 1 summarises the methodologies included in the original Cochrane reviews. The included meta-analyses came from a total of 112 reviews. Of the included reviews, the bivariate LMM method was implemented in 44 (39.29%), the HSROC in 32 (28.57%), the bivariate GLMM in 25 (22.32%), the univariate approach in 12 (10.71%); 9 (8.04%) did not list which methodology was used and 5 (4.46%) did not perform a meta-analysis. Notably, of the 12 meta-analyses that implemented the univariate method, 11 did so in conjunction with other methodologies and only for studies with sparse data or

Table 2 Comparisons of the results produced by the bivariate LMM and GLMM among the Cochrane meta-analyses of diagnostic studies, with the bivariate GLMM as the reference

Estimate*	Median	IQR
Sensitivity	0.99	0.95–1.01
Sensitivity CI width	1.05	0.86–1.28
Sensitivity variance	1.11	0.73–1.63
Specificity	1.00	0.98–1.00
Specificity CI width	1.08	0.92–1.30
Specificity variance	1.16	0.84–1.69
ρ	0.00	–0.05–0.05

*The absolute difference was calculated for the correlation coefficient estimates, ρ, while the relative difference was calculated for other estimates.
GLMM, generalised linear mixed model; LMM, linear mixed model.

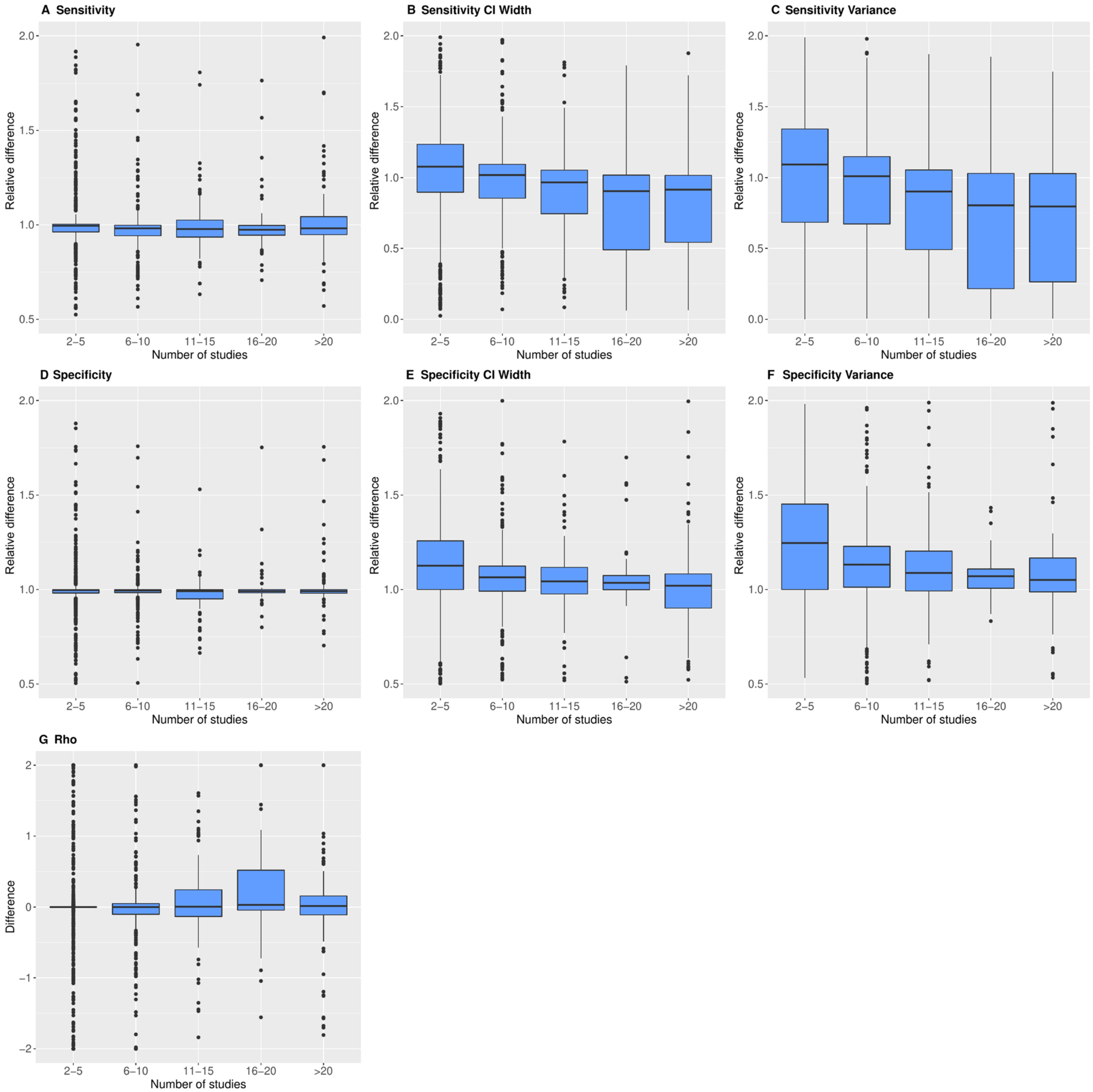

Figure 1 Comparison of the bivariate linear mixed model (LMM) versus bivariate generalised linear mixed model (GLMM), sorted by the number of studies in each meta-analysis.

small sample size (ie, <5). Only one meta-analysis was performed using the univariate method solely.

Overall comparisons of point and interval estimates

Table 2 presents a summary of comparisons for point and interval estimates. The relative differences tended to be positively skewed owing to some extreme estimates produced by the bivariate LMM method relative to the bivariate GLMM. The relative differences for the point estimates of sensitivity had a median of 0.99 (IQR=0.95–1.01), those for the CI widths had a median of 1.05 (IQR=0.86–1.28) and those for the variances had a median of 1.11 (IQR=0.73–1.63). The relative differences for the point estimate of specificity appeared to be approximately symmetrically distributed, with a median of 1.00 (IQR=0.98–1.00); those for the CI widths had a median of 1.08 (IQR=0.92–1.30), and those for the variances had a median of 1.16 (IQR=0.84–1.69). The point estimate for ρ had a median of 0.00 (IQR= −0.05–0.05). Of note, ρ was incalculable in 175 meta-analyses for the bivariate GLMM.

Comparisons categorised by meta-analysis characteristics

Figure 1 summarises the comparison between the bivariate LMM and GLMM based on the number of studies in each meta-analysis. In general, more outliers appeared for meta-analyses with fewer studies. For the relative differences in both sensitivity and specificity, the medians were close to 1, and the IQRs were roughly the same regardless of the number of studies. For the sensitivity CI width and sensitivity variance, the median moved closer to 1 as the number of studies increased. For the specificity CI width

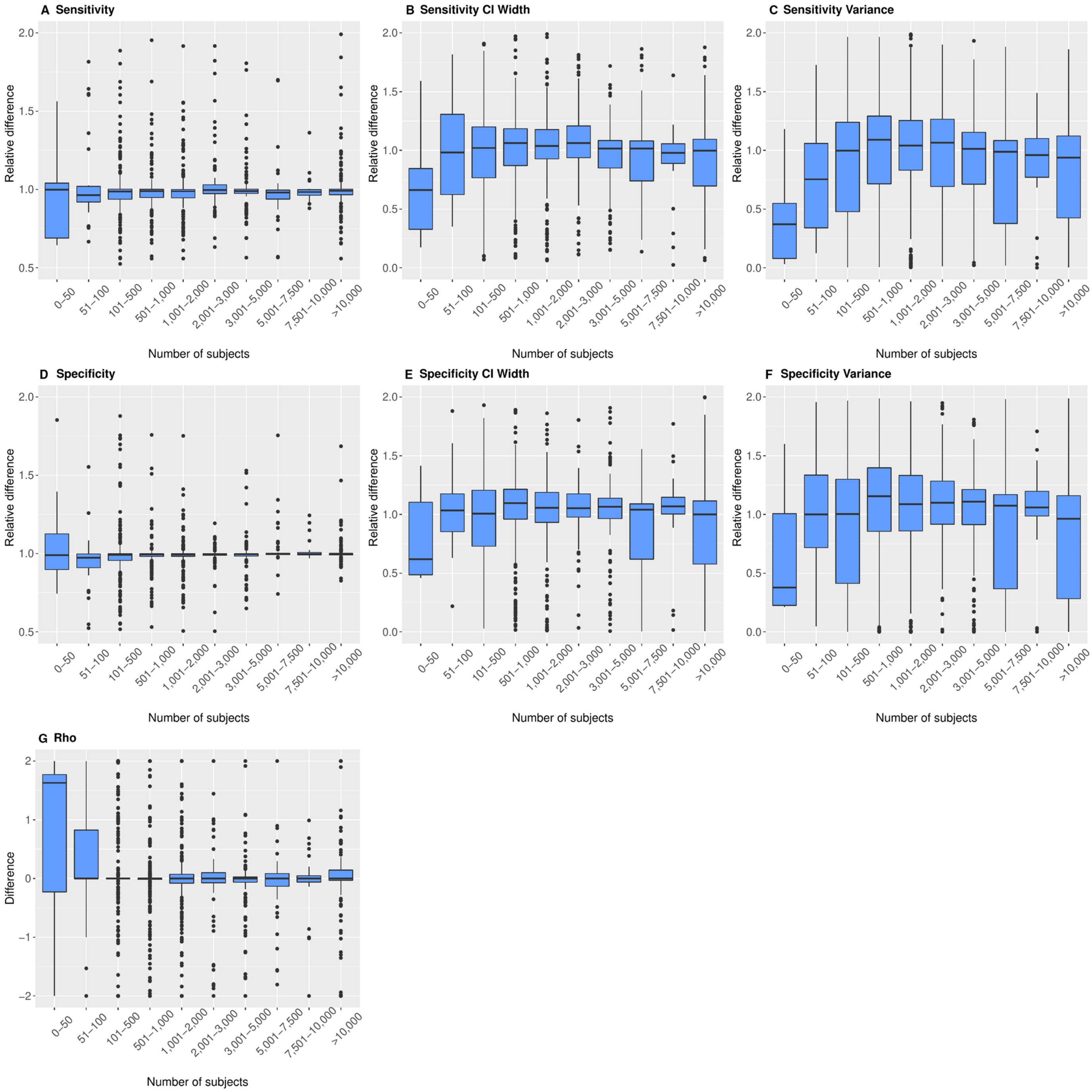

Figure 2 Comparison of the bivariate linear mixed model (LMM) versus bivariate generalised linear mixed model (GLMM), sorted by the number of subjects in each meta-analysis.

and specificity variance, the median was also closer to 1 as the number of studies increased, with the widest IQR occurring for meta-analyses with 2–5 studies. The absolute differences of ρ had a median of approximately 0 regardless of the number of included studies.

Figure 2 summarises the comparisons of the bivariate LMM versus bivariate GLMM based on the number of subjects in each meta-analysis. For meta-analyses with more than 100 subjects, the bivariate LMM and GLMM produced approximately the same sensitivities and specificities. For meta-analyses with more than 100 subjects, the relative differences of sensitivities were mostly greater than 1, while those of specificities were slightly smaller than 1. The sensitivity CI widths by the bivariate LMM and GLMM were approximately equal regardless of the number of subjects. The sensitivity variances produced by the two methods were approximately equal when the number of subjects was greater than 100. The specificity CI widths by the bivariate LMM were lower than those by the bivariate GLMM for meta-analyses with less than 50 subjects, as were the specificity variances. The differences of ρ were slightly smaller than 0 for meta-analyses with less than 50 subjects.

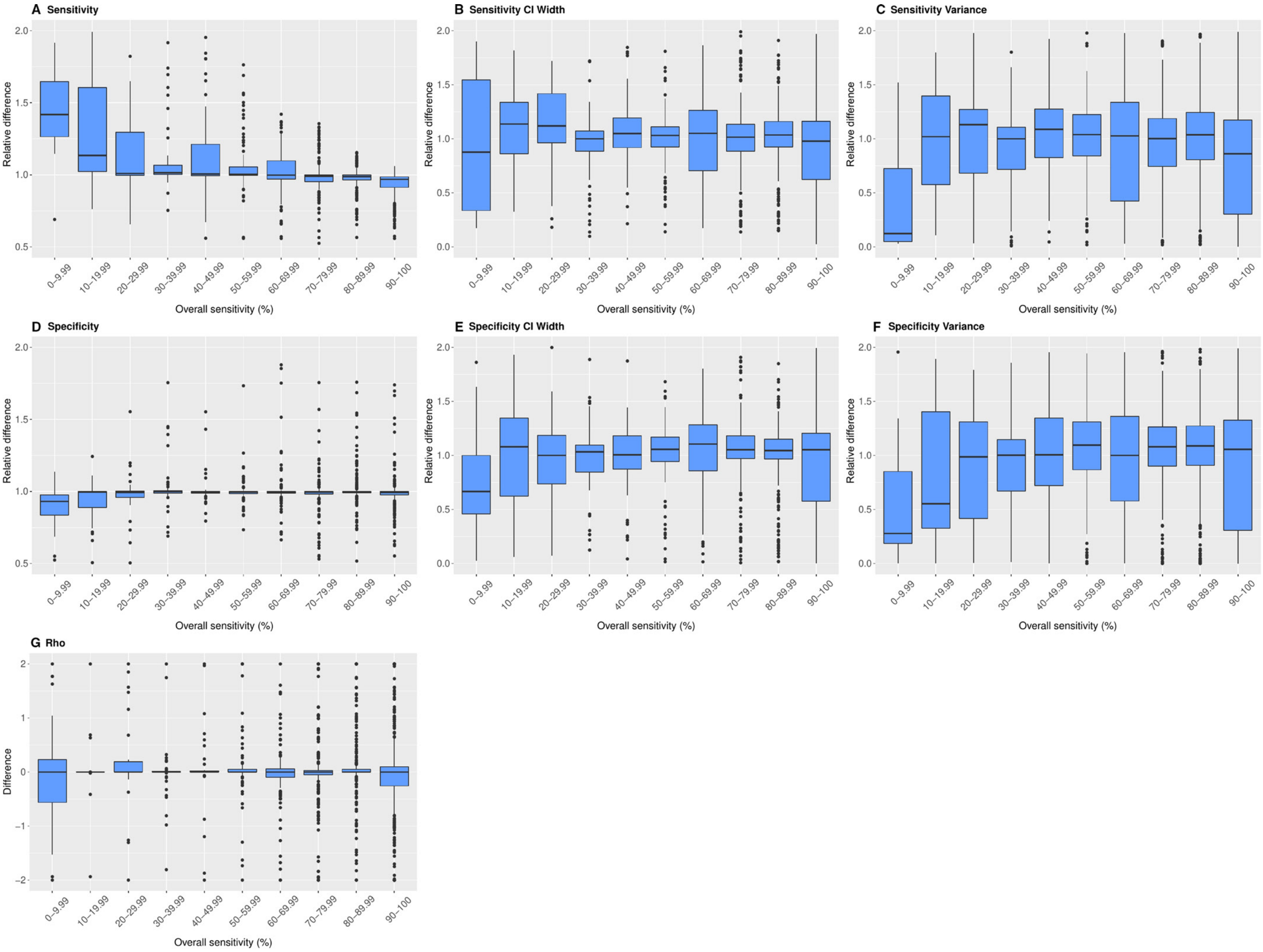

Figure 3 Comparison of the bivariate linear mixed model (LMM) versus bivariate generalised linear mixed model (GLMM), sorted by the sensitivity from the bivariate GLMM in each meta-analysis.

Figure 3 summarises the comparisons of the bivariate LMM versus bivariate GLMM method based on the overall sensitivity produced by the bivariate GLMM in each meta-analysis. Many outliers were present when the overall sensitivity was 90% or greater. The sensitivities produced by the bivariate LMM and GLMM were approximately equal when the overall sensitivity was at least 20%. The sensitivity CI widths by the bivariate LMM and GLMM were generally similar regardless of the overall sensitivity. The sensitivity variances produced by the two estimates were approximately equal when the overall sensitivity was at least 10%. The specificities and specificity CI widths produced by the bivariate LMM and GLMM were approximately equal when the overall sensitivity was at least 10%. The specificity variances were approximately equal when the overall sensitivity was at least 20%. The differences of ρ did not substantially change as the overall sensitivity varied.

Figure 4 summarises the comparisons of the bivariate LMM versus bivariate GLMM method based on the overall specificity produced by the bivariate GLMM in each meta-analysis. The sensitivities produced by the bivariate LMM and GLMM methods were nearly the same across different overall specificities. When the overall specificity was at least 20%, the two methods produced approximately the same specificities. When the overall specificity was less than 10%, the CI widths of both sensitivities and specificities produced by the bivariate LMM method were noticeably smaller than those by the bivariate GLMM. When the overall specificity was at least 40%, the variances of both sensitivity and specificity produced by the two estimates were generally similar; otherwise, the bivariate GLMM produced noticeably larger values. The differences of ρ did not substantially change as the overall specificity varied.

We explored the causes of outlying results and found that they usually appeared in meta-analyses with a few studies, small sample sizes, or overall sensitivity or specificity close to boundary values. For example, one meta-analysis that produced extreme estimates had four studies, 78 subjects, an overall sensitivity of 0 and an overall specificity of 1. A second such meta-analysis had an overall sensitivity of 0.96, an overall specificity of 0.44, and contained 2 studies and 34 subjects.

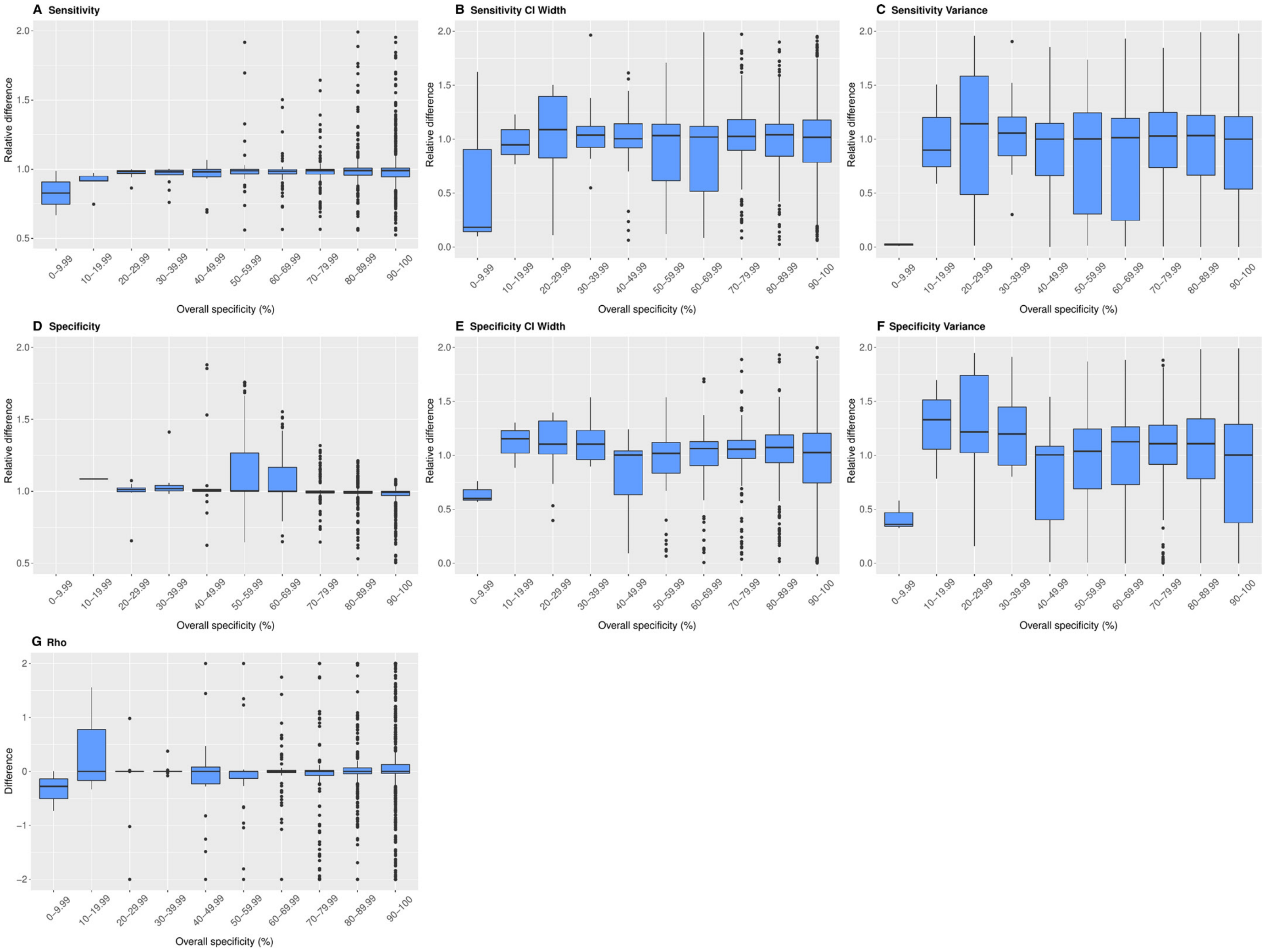

Figure 4 Comparison of the bivariate linear mixed model (LMM) versus bivariate generalised linear mixed model (GLMM), sorted by the specificity from the bivariate GLMM in each meta-analysis.

Supplemental analyses

Online supplemental table S1 in the Supplementary Material summarises additional results of DOR and AUC. The median relative difference of the DORs was 0.85 (IQR=0.45–1.00). For the DOR CI widths, the relative differences of the bivariate LMM had a median of 0.85 (IQR=0.27–1.18). The median relative differences for the AUCs were nearly identical for both methods; those by the bivariate LMM had a median of 0.99 (IQR=0.93–1.03).

Online supplemental figures S1–S4 further present the comparisons between the bivariate LMM and GLMM. As is illustrated in online supplemental figure S1, the two methods produced similar median AUCs for two to five studies, though the bivariate LMM method generally produced narrower IQRs with medians closer to the bivariate GLMM as the number of studies increased. For meta-analyses with at least six studies, the bivariate LMM produced smaller DOR estimates than the bivariate GLMM. For meta-analyses with two to five studies, the DOR CI widths produced by the bivariate LMM were closest to those produced by the bivariate GLMM.

Online supplemental figure S2 shows that both methods produced approximately the same AUC for meta-analyses with more than 100 subjects, while the relative differences of AUCs were mostly smaller than 0 for meta-analyses with less than 100 subjects. As the number of subjects increased, the relative differences of DORs and their CI widths increased towards 1.

The relative differences of AUCs did not substantially change as the overall sensitivity or specificity varied (online supplemental figures S3 and S4). The relative differences of DORs produced by the two methods were similar for overall sensitivities or specificities less than 90%, while they were noticeably smaller than 1 for overall sensitivities or specificities greater than 90%.

DISCUSSION

In this empirical study of 1379 meta-analyses of diagnostic studies, we found that while the bivariate GLMM is the recommended method for Cochrane reviews, it is not as commonly used as the bivariate LMM or HSROC method. Point estimates of the overall sensitivities and specificities produced by the bivariate LMM and GLMM were generally similar. However, their CI widths could be noticeably different, and the bivariate GLMM generally produced narrower CIs than the bivariate LMM when a meta-analysis contained two to five studies. Additionally, when the number of subjects was less than 100 or the overall sensitivities or specificities were close to 0% or 100%, the bivariate LMM could produce substantially different AUCs, DORs and DOR CI widths from the bivariate GLMM. In general, when the number of studies and subjects in a meta-analysis is small and either the sensitivity or specificity is close to 0% or 100%, the bivariate LMM and GLMM may produce substantially different results. It has been shown that the approximate method produces biased estimates in cases of large between-study heterogeneity;[23] thus, differences in heterogeneity between studies included in different meta-analyses are likely to result in more discrepancies between methods.

This study had several limitations. Both methods exhibited convergence issues in several meta-analyses. It may be of interest to employ other transformations such as the complementary log-log transformation and examine if the model fitting could be improved.[22] The convergence issues may also be addressed by employing Bayesian analyses with informative priors; multiple methods for random-effects meta-analysis approaches exist under the Bayesian framework for simultaneously combining sensitivity and specificity.[16 24 34] In lieu of the SROC curve, some researchers have recommended supplying confidence regions for the mean sensitivity and specificity of the bivariate model;[16 25] this could be an expansion for further research, for example, by comparing the area of the confidence regions. It may also be of interest to compare the area of prediction regions; in a random-effects study setting, the prediction region gives a range for the estimate in a new study.[35] The inclusion of prediction intervals can facilitate the application of meta-analysis for diagnostic testing by making it easier to apply the results to the clinical setting.[36]

Several studies have been conducted in the literature to compare alternative methods for diagnostic test accuracy. Using data from two available studies, Ma *et al*[4] empirically compared the results of the SROC, bivariate LMM, HSROC and bivariate GLMM methods. Zapf *et al*[37] proposed a non-parametrical meta-analysis that addresses the issue of convergence; through simulation studies, they showed that the resulting bias, empirical coverage and mean squared error are generally superior to those produced by the bivariate GLMM. Both the bivariate GLMM and LMM rely on a single pair of sensitivity and specificity point estimates from each available study in the meta-analysis and cannot combine results from studies that report on diagnostic contingency tables with multiple thresholds. To combat this shortcoming, Steinhauser *et al*[38] introduced an approach for modelling multiple thresholds that can account for between-study heterogeneity and dependence of sensitivity and specificity. This approach uses all available data and can estimate diagnostic test accuracy and the threshold of optimal performance. Hoyer and Kuss[39] developed a quadrivariate GLMM to compare results from two diagnostic tests to a gold standard. This method can calculate the differences between sensitivities and specificities, and it accounts for the correlation between tests and heterogeneity between studies.

The variation of estimates produced by the bivariate GLMM and LMM methods calls into question the appropriateness of the normality assumption within individual studies required by the bivariate LMM method. In such cases, extra caution is needed when interpreting the results from this method, and the bivariate GLMM may be recommended.

Contributors Conceptualisation: HC, LL; Methodology: KJR, LL; Formal analysis: KJR; Data curation: LL; Writing of the original draft: KJR; Writing of the review and editing: HC, LL; Visualisation: KJR; Supervision: LL; Funding acquisition: HC, LL; Guarantor: LL.

Funding This research was supported in part by the US National Institutes of Health/National Institute of Mental Health grant R03 MH128727, the National Institutes of Health/National Library of Medicine grant R01 LM012982, and the National Institutes of Health/National Center for Advancing Translational Sciences grant UL1 TR001427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iD
Lifeng Lin http://orcid.org/0000-0002-3562-9816

REFERENCES

1 Devillé WL, Buntinx F, Bouter LM, *et al*. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Med Res Methodol* 2002;2:9.
2 Bossuyt PM, Reitsma JB, Bruns DE, *et al*. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Ann Intern Med* 2003;138:40–4.
3 Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001;323:157–62.
4 Ma X, Nie L, Cole SR, *et al*. Statistical methods for multivariate meta-analysis of diagnostic tests: an overview and tutorial. *Stat Methods Med Res* 2016;25:1596–619.

5 Honest H, Khan KS. Reporting of measures of accuracy in systematic reviews of diagnostic literature. *BMC Health Serv Res* 2002;2:4.
6 Glas AS, Lijmer JG, Prins MH, *et al*. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol* 2003;56:1129–35.
7 Reitsma JB, Glas AS, Rutjes AWS, *et al*. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005;58:982–90.
8 Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978;8:283–98.
9 Irwig L, Macaskill P, Glasziou P, *et al*. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol* 1995;48:119–30.
10 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.
11 Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull* 1995;117:167–78.
12 Irwig L, Tosteson AN, Gatsonis C, *et al*. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994;120:667–76.
13 Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making* 1993;13:313–21.
14 Midgette AS, Stukel TA, Littenberg B. A meta-analytic method for summarizing diagnostic test performances: receiver-operating-characteristic-summary point estimates. *Med Decis Making* 1993;13:253–7.
15 Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med* 2002;21:1237–56.
16 Arends LR, Hamza TH, van Houwelingen JC, *et al*. Bivariate random effects meta-analysis of ROC curves. *Med Decis Making* 2008;28:621–38.
17 Lijmer JG, Bossuyt PMM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Stat Med* 2002;21:1525–37.
18 van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med* 2002;21:589–624.
19 Van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. *Stat Med* 1993;12:2273–84.
20 Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *J Clin Epidemiol* 2006;59:1331–2.
21 Harbord RM, Deeks JJ, Egger M, *et al*. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007;8:239–51.
22 Chu H, Guo H, Zhou Y. Bivariate random effects meta-analysis of diagnostic studies using generalized linear mixed models. *Med Decis Making* 2010;30:499–508.
23 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.
24 Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001;20:2865–84.
25 Rücker G, Schumacher M. Letter to the editor. *Biostatistics* 2009;10:806–7.
26 Chu H, Guo H. Letter to the editor. *Biostatistics* 2009;10:201–3.
27 Takwoingi Y, Guo B, Riley RD, *et al*. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. *Stat Methods Med Res* 2017;26:1896–911.
28 Simel DL, Bossuyt PMM. Differences between univariate and bivariate models for summarizing diagnostic accuracy may not be large. *J Clin Epidemiol* 2009;62:1292–300.
29 Riley RD, Abrams KR, Sutton AJ, *et al*. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Med Res Methodol* 2007;7:3.
30 Harbord RM, Whiting P, Sterne JAC, *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.
31 Reitsma JB, Zwinderman AH. Response to Chu and Cole: bivariate meta-analysis of sensitivity and specificity with sparse data. *J Clin Epidemiol* 2006;59:1332–3.
32 Lin L, Shi L, Chu H, *et al*. The magnitude of small-study effects in the *Cochrane Database of Systematic Reviews*: an empirical study of nearly 30 000 meta-analyses. *BMJ Evid Based Med* 2020;25:27–32.
33 Higgins JPT, Thomas J, Chandler J. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, UK: John Wiley & Sons, 2019.
34 Ma X, Lian Q, Chu H, *et al*. A Bayesian hierarchical model for network meta-analysis of multiple diagnostic tests. *Biostatistics* 2018;19:87–102.
35 Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc* 2009;172:137–59.
36 Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011;342:d549.
37 Zapf A, Hoyer A, Kramer K, *et al*. Nonparametric meta-analysis for diagnostic accuracy studies. *Stat Med* 2015;34:3831–41.
38 Steinhauser S, Schumacher M, Rücker G. Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. *BMC Med Res Methodol* 2016;16:97.
39 Hoyer A, Kuss O. Meta-analysis for the comparison of two diagnostic tests to a common gold standard: a generalized linear mixed model approach. *Stat Methods Med Res* 2018;27:1410–21.