Received: 20 October 2021 | Revised: 17 March 2022 | Accepted: 18 March 2022

DOI: 10.1111/jebm.12467

PERSPECTIVE

WILEY

Overconfident results with the bivariate random effects model for meta-analysis of diagnostic accuracy studies

Luis Furuya-Kanamori[1] | Eletherios Meletis[2] | Chang Xu[3] | Polychronis Kostoulas[2] | Suhail AR Doi[3]

[1]UQ Centre for Clinical Research, Faculty of Medicine, The University of Queensland, Herston, Australia

[2]Laboratory of Epidemiology and Artificial Intelligence, Faculty of Public and One Health, School of Health Sciences, University of Thessaly, Karditsa, Greece

[3]Department of Population Medicine, College of Medicine, QU Health, Qatar University, Doha, Qatar

Correspondence
Dr. Luis Furuya-Kanamori, UQ Centre for Clinical Research, The University of Queensland, Herston, Australia.
Email: l.furuya@uq.edu.au

Received on 20-Oct-2021; Accepted on 18-Mar-2022

KEYWORDS
Bayesian, diagnosis, latent class, split component synthesis

Meta-analyses of diagnostic accuracy studies are a fundamental component of evidence-based medicine, and they are extensively used in medical imaging and the clinical laboratory. Techniques specifically developed to combine independent studies of diagnostic accuracy and provide pooled estimates for sensitivity (Se), specificity (Sp), positive (pLR) and negative (nLR) likelihood ratios are relatively new. In 2001, Rutter and Gatsonis proposed the hierarchical summary receiver operating characteristic (HSROC) model,[1] and in 2004 Macaskill described an empirical Bayes approach.[2] Soon after, in 2005, Reitsma et al. proposed the bivariate random effects model,[3] which has been widely adopted and is the most commonly used method for diagnostic meta-analysis.[4]

However, as pointed out by Diaz,[5] the statistical performance of the bivariate model has not been scrutinized. Diaz found that the performance of the bivariate model deteriorates when between-study heterogeneity increases and the number of studies decrease.[5] Our simulation studies found similar results—with moderate levels of heterogeneity ($tau^2 = 1$), the coverage probabilities of Se, Sp, and the diagnostic odds ratio (DOR) with the bivariate model dropped below the nominal level.[6] Diagnostic accuracy studies usually favor sensitivity over specificity, or vice versa leading to diagnostic 2 × 2 tables with one or more of the cells with low frequency or zero counts. Thus, extreme DORs are more commonly observed in diagnostic than in intervention meta-analyses, which leads to high levels of heterogeneity (despite the wide confidence intervals of the studies).[7]

1 | CASE STUDY: ELISA FOR DETECTING RABIES ANTIBODIES

We report the results of a meta-analysis with five studies estimating the operating characteristics of enzyme-linked immunosorbent assay (ELISA) when compared against the reference standard, fluorescent focus inhibition test (RFFIT), for detection of immune response (i.e., seropositive or seronegative status) after a rabies vaccine.[8] The sample size of the studies ranged from 28 to 990. Despite the studies not having 'extreme' values (range of Se: 84.2 to 100; and range of Sp: 87.1 to 100) (Table 1), large between-study heterogeneity ($tau^2 = 8.3$) was observed, due to low frequency of false positives and false negatives, including cells with zero counts. The pooled estimates were calculated using the:

- Bivariate random effects model[3]
- An extension of the bivariate model proposed by Chu and Cole[9] which is a generalized linear mixed model with a modeling approach for sparse data recommended when there are low cell counts

TABLE 1 Data from the five studies included in the meta-analysis

Author, year	Sample size	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Feyssaguet 2007	655	191	3	2	459	99.0 (96.3–99.9)	99.4 (98.1–99.9)
Muhamuda 2007	990	740	0	0	250	100 (99.5–100)	100 (98.5–100)
Pandit 1991	28	16	1	3	8	84.2 (60.4–96.6)	88.9 (51.8–99.7)
Welch 2009	82	32	2	2	46	94.1 (80.3–99.3)	95.8 (85.8–99.5)
Zhao 2019	428	374	4	23	27	94.2 (91.4–96.3)	87.1 (70.2–96.4)

TP, true positive; FP, false positive; FN, false negative; TN, true negative; CI, confidence interval.

TABLE 2 Comparison of pooled estimates using the bivariate models, split component method, and Bayesian HSROC

	Bivariate model	Bivariate model (Chu and Cole extension)	Split component method	Bayesian HSROC[a]
Sensitivity	98.4 (90.2–99.8)	98.5 (88.6–99.8)	95.6 (62.6–99.6)	92.7 (67.4–99.8)
Specificity	98.3 (85.4–99.8)	98.5 (82.5–99.9)	95.4 (61.9–99.6)	92.8 (75.5–99.5)
pLR	58.86 (5.93–584.63)	65.60 (4.78–899.62)	20.91 (2.04–214.02)	12.88 (2.75–199.60)
nLR	0.02 (0.00–0.11)	0.02 (0.00–0.13)	0.05 (0.01–0.48)	0.08 (0.01–0.43)
DOR	3608.5 (62.3–208848.8)	4413.1 (42.2–461117.2)	451.1 (16.7–12205.6)	163.7 (6.4–99301.0)
AUC	0.98 (0.89–0.99)[b]	0.99 (0.87–0.99)[b]	0.96 (0.80–0.99)	–

HSROC, hierarchical summary receiver operating characteristic; pLR, positive likelihood ratio; nLR, negative likelihood ratio; DOR, diagnostic odds ratio; AUC, area under the curve.
[a]Reference standard for sensitivity 99.9 (99.5–100) and specificity 99.7 (99.2–100).
[b]AUC estimated from the DOR.

- Split component synthesis (SCS) method[6] which summarizes the study-specific ln(DOR) using the inverse variance heterogeneity model,[10] and then splits the summary ln(DOR) into its component parts (i.e., logit(Se) and logit(Sp))
- Hierarchical summary receiver operating characteristic model in a Bayesian latent class meta-analysis framework (Bayes-HSROC),[11] which assumes the absence of a perfect reference standard

The analyses were conducted in Stata MP version 14.1 using the *metandi* module[12] for the bivariate models and the *diagma* module[13] for the SCS method. The Bayes-HSROC model was implemented in the R programming language using the *rjags*[14] and *runjags*[15] packages. The Bayes-HSROC model applies Bayesian inference, where the posterior distribution of the parameters of interest depends on the likelihood function and the prior information provided. The likelihood function was computed as a statistical model for the observed data (Supplementary Material S1). Noninformative priors were used, meaning that no external information was provided to the model. Parameter estimates were based on analytical summaries of 500,000 iterations of two chains after a burn-in phase of 10,000 iterations. Time series plots were used to assessed convergence.[16] The two chains converged to the same solution and autocorrelation plots dropped-off fast (Supplementary Material S2).

The results of the four models are presented in Table 2, and point estimates and confidence intervals (credible intervals for the Bayes-HSROC model) were more conservative with the SCS methods and the Bayes-HSROC model than with both bivariate models. The Se was 98.4 (95% CI 90.2–99.8) with the bivariate model, while it was 95.6 (95% CI 62.6–99.6) and 92.7 (95% CI 67.4–99.8) with the SCS method and Bayesian HSROC respectively, with similar results for the Sp.

2 | SIMULATION STUDY

The five studies included in the case study were simulated fixing the sample size to original study and fixing the true value of Se and Sp = 0.96 (based on the pooled estimates in Table 2). The number of diseased (*dis*) and nondiseased (*ndis*) individuals were drawn from a binomial distribution using the sample size and the actual prevalence of seropositivity and seronegativity, respectively in each study. The four cell counts (*tp*, *fp*, *fn*, *tn*) were then derived from *dis* and *ndis*, and the Se and Sp. Next, the four counts were divided by a scale parameter (minimum value = 1) that was derived from a transformation of a

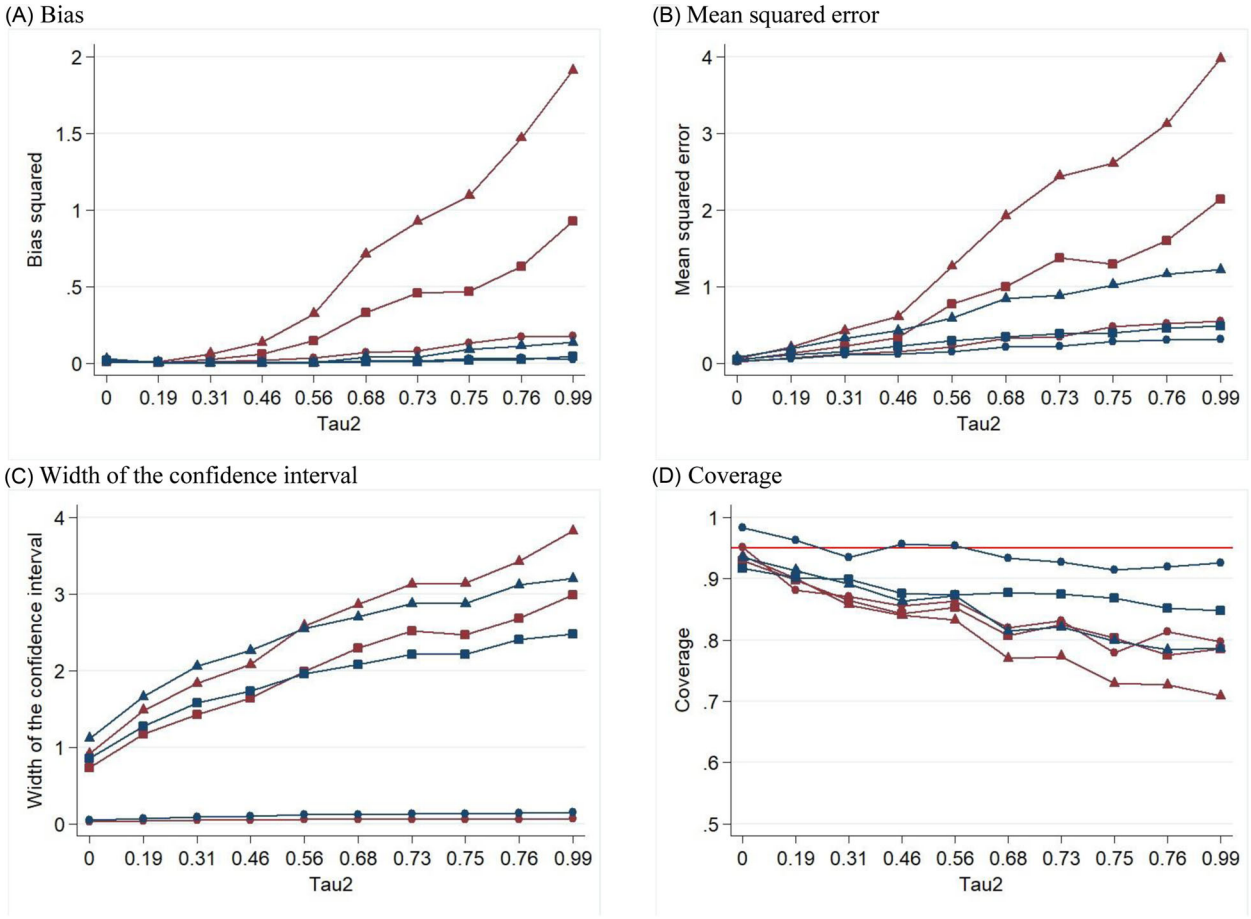

FIGURE 1 Performance comparison of the diagnostic odds ratio (triangle), sensitivity (circle), and specificity (square) between the split component synthesis method (blue) and the bivariate model (red) at different levels of heterogeneity

hypothetically imputed bias variance to introduce systematic error.[17] Both random and systematic error were introduced by regenerating a simulated Se and Sp from a beta distribution with parameters *tp/f* and *fn/f*; and *tn/f* and *fp/f*, respectively. Next, the studies were generated and meta-analyzed, and 1000 meta-analyses were simulated in each of 10 runs, with run 1 representing random error alone (scale parameter = 1) and runs 2–10 having increasing level of between-study heterogeneity. The Stata codes for the data generation are provided in the Supplementary Material S3.

For each level of heterogeneity, summary DOR, Se, and Sp estimated by the extension of the bivariate model (proposed by Chu and Cole) and SCS method were compared based on mean absolute estimation error squared (bias squared), mean squared error (MSE), width of the confidence interval, and coverage probability.[18] The distribution of Se, Sp, and tau^2 generated for each of the 10 runs are reported in the Supplementary Material S4. The bivariate model did not converge in 19% of the simulated meta-analyses and these were excluded from the performance analyses for both the bivariate model and SCS method.

The simulation study revealed that the SCS method's DOR, Se, and Sp estimates were less biased (Figure 1A) and had smaller MSE than the bivariate model estimates (Figure 1B). As heterogeneity increased, the width of the 95% confidence interval became wider with the bivariate model (Figure 1C), yet it had lower coverage probability of the confidence interval compared to the SCS method (Figure 1D). It was not possible to compare the performance of the models when moderate or extensive heterogeneity was introduced as the bivariate model did not converge in > 50% of the meta-analyses.

3 | DISCUSSION

In our case study with small number of studies and large heterogeneity, discrepancies were observed in the confidence/credible intervals - very narrow confidence intervals with the bivariate models, while the confidence/credible intervals were wide with the SCS method and the Bayes-HSROC. The simulation study revealed that when heterogeneity was introduced, there was a considerable decline in the performance of the bivariate model. Therefore, it is very likely that the results of the case study and other studies using the bivariate model would generate spuriously overconfident results due to overdispersion of the data relative to the model.

Between-study heterogeneity is the norm in meta-analyses of diagnostic accuracy studies. In a methodological review, Dinnes et al.[19] found that there was statistical heterogeneity in 79% of diagnostic meta-analyses; thus pooling methods have to be able to properly

maintain performance when heterogeneity is present. This study therefore suggests that newer SCS method can resolve the issue of overdispersion with the bivariate model and needs to be prioritized in research. Alternatively, a Bayesian approach can be used, especially when the reference method is imperfect.

In conclusion, the bivariate model suffers from the same issue of overdispersion as the random effects model in standard meta-analysis[20] and the SCS method seems to be a viable alternative. The latter also avoids the issue of nonconvergence and is not unduly affected by varying implicit thresholds given that it starts with synthesis of the DOR. Further evaluation is therefore recommended to independently verify these findings, so that the necessary recommendations can be made for the research community

ACKNOWLEDGMENTS

Open Access Funding provided by The University of Queensland.

FUNDING

LFK was supported by Australian National Health and Medical Research Council Early Career Fellowships (APP1158469).

DATA SHARING

The data that supports the findings of this study are available in the Supplementary Material of this article.

ORCID

Suhail AR Doi https://orcid.org/0000-0002-2630-2125

REFERENCES

1. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med*. 2001;20(19), 2865–2884.
2. 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(9), 925–932.
3. 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(10), 982–990.
4. Ochodo EA, Reitsma JB, Bossuyt PM, Leeflang MMG. Survey revealed a lack of clarity about recommended methods for meta-analysis of diagnostic accuracy data. *J Clin Epidemiol*. 2013;66(11), 1281–1288.
5. Diaz M. Performance measures of the bivariate random effects model for meta-analyses of diagnostic accuracy. *Comput Stat Data Anal*. 2015;83:82–90.
6. Furuya-Kanamori L, Kostoulas P, Doi SAR. A new method for synthesizing test accuracy data outperformed the bivariate method. *J Clin Epidemiol*. 2021;132:51–58.
7. Begg CB. Systematic reviews of diagnostic accuracy studies require study by study examination: first for heterogeneity, and then for sources of heterogeneity. *J Clin Epidemiol*. 2005;58(9), 865–866.
8. Ling W, Doi SAR, Lau CL, Mills DJ, Kostoulas P, Furuya-Kanamori L. Diagnostic accuracy of ELISA kits for measurement of rabies antibodies. *J Travel Med*. 2021;28(5).
9. 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(12), 1331–1332.
10. Doi SA, Barendregt JJ, Khan S, Thalib L, Williams GM. Advances in the meta-analysis of heterogeneous clinical trials I: the inverse variance heterogeneity model. *Contemp Clin Trials*. 2015;45:130–138. Pt A.
11. Dendukuri N, Schiller I, Joseph L, Pai M. Bayesian meta-analysis of the accuracy of a test for tuberculous pleuritis in the absence of a gold standard reference. *Biometrics*. 2012;68(4), 1285–1293.
12. Harbord RM, Whiting P. metandi: meta-analysis of diagnostic accuracy using hierarchical logistic regression; 2009.
13. Furuya-Kanamori L, Doi SAR. DIAGMA: Stata module for the split component synthesis method of diagnostic meta-analysis; 2020.
14. Plummer M, Stukalov A, Denwood M, Plummer MM. Package 'rjags'. 2018.
15. Denwood MJ. Runjags: an R Package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *J Stat Soft*. 2016;71(9), 1–25.
16. Toft N, Innocent GT, Gettinby G, Reid SW. Assessing the convergence of Markov Chain Monte Carlo methods: an example from evaluation of diagnostic tests in absence of a gold standard. *Prev Vet Med*. 2007;79(2–4), 244–256.
17. Doi SAR, Furuya-Kanamori L. Selecting the best meta-analytic estimator for evidence-based practice: a simulation study. *Int J Evid Based Healthc*. 2020;18(1), 86–94.
18. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med*. 2006;25(24), 4279–4292.
19. 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(12), 1–113.
20. Doi SAR, Furuya-Kanamori L, Thalib L, Barendregt JJ. Meta-analysis in evidence-based healthcare: a paradigm shift away from random effects is overdue. *Int J Evid Based Healthc*. 2017;15(4), 152–160.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Furuya-Kanamori L, Meletis E, Xu C, Kostoulas P, Doi SAR Overconfident results with the bivariate random effects model for meta-analysis of diagnostic accuracy studies. *J Evid Based Med*. 2022;15:6–9. https://doi.org/10.1111/jebm.12467.