Research Synthesis Methods (2026), **17**: 693–713
doi:10.1017/rsm.2025.10070

Research Synthesis Methods

RESEARCH ARTICLE

Diagnostic test accuracy meta-analysis based on an exact within-study variance calculation method

Olana Angesa Dabi and Zelalem Firisa Negeri

Department of Statistics and Actuarial Science, University of Waterloo, Canada

Corresponding author: Zelalem Firisa Negeri; Email: znegeri@uwaterloo.ca

Received: 29 June 2025; **Revised:** 30 November 2025; **Accepted:** 10 December 2025; published online 4 February 2026

Keywords: diagnostic test accuracy; exact variance calculation; meta-analysis; sensitivity; specificity

Abstract

A meta-analysis of diagnostic test accuracy (DTA) studies typically synthesizes study-specific test sensitivity (Se) and specificity (Sp) to quantify the accuracy of an index test of interest. The bivariate linear mixed effects model with logit transformation of Se and Sp (BLMM-Logit) is commonly used to make statistical inferences, but may lead to misleading results due to the need for Haldane–Anscombe correction and an approximate estimation of variance within the study. Alternative models based on the arcsine square root and Freeman–Tukey double arcsine transformation have been proposed to address these issues; however, they still rely on approximate variance estimation, which is suitable only for large sample sizes. The bivariate generalized linear mixed effects model (BGLMM) is another option, but it faces convergence issues with small meta-analyses or sparse primary studies. To address these limitations, we proposed an exact within-study variance calculation method that does not require Haldane–Anscombe correction and is applicable regardless of the transformation used or the number of studies and participants. We evaluated this method against existing approaches using real-life and simulated DTA meta-analyses. The methods were comparable for large meta-analyses. However, BLMM-Logit demonstrated substantial negative bias in estimating variances between studies and consistently underestimated summary Se and Sp in all simulation scenarios. In contrast, the proposed exact methods (Exact-Logit, Exact-ASR, and Exact-FTDA) and BGLMM had minimal bias and better performance metrics, particularly for meta-analyses with sparse primary studies. Thus, the proposed exact methods should be preferred for DTA meta-analyses with small or sparse studies.

Highlights

What is already known?

- The bivariate linear mixed effects model (BLMM) using logit-transformed sensitivity and specificity is the standard approach in meta-analysis of diagnostic test accuracy (DTA) studies.
- However, BLMM has limitations, such as reliance on Haldane–Anscombe correction and approximate within-study variance estimates, especially problematic for meta-analyses with small or sparse studies.
- The bivariate generalized linear mixed model (BGLMM) offers a different framework but often faces convergence issues when meta-analyses are small or sparse.

What is new?

- The article proposes a novel exact method for within-study variance calculation that:
  - does not require Haldane–Anscombe correction,

This article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2026. Published by Cambridge University Press on behalf of The Society for Research Synthesis Methodology. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NoDerivatives licence (https://creativecommons.org/licenses/by-nd/4.0), which permits re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited.

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

  - is transformation-invariant, and
  - works well regardless of sample size or number of studies.
- The method is evaluated against existing models using both real and simulated datasets.

Potential impact for RSM readers

- Readers involved in meta-analysis of DTA studies will gain a more accurate and flexible toolset, particularly for small or sparse datasets, where existing models often break down.
- The proposed exact method may reduce bias and improve the reliability of inference to support better clinical and methodological decisions based on DTA meta-analyses.
- The work helps to advance the methodological frontier in synthesis science, thereby addressing a persistent practical challenge in diagnostic research.

1. Introduction

Diagnostic tests or screening tests are essential in clinical practice for identifying individuals who have diseases and those who do not. The accuracy of these tests and the appropriateness of the thresholds employed to define positive results are assessed by diagnostic test accuracy (DTA) studies.[1] These studies present a range of diagnostic accuracy metrics that are obtained by comparing the relevant diagnostic test (index test) result to the gold standard, which is the test that can accurately differentiate between people who are positive for a disease and those who are not. A diagnostic test is considered inaccurate if it yields a positive result for a patient who does not have the condition or a negative result for a patient who does have a condition. Therefore, it is essential to compare a diagnostic test's accuracy to that of the reference or gold standard test to mitigate the risk of misdiagnosis. The results of DTA studies are composed of the probability of a positive test result in diseased individuals (sensitivity) and the probability of a negative test result in healthy individuals (specificity). The sensitivity (Se) and specificity (Sp) are the two main DTA metrics that indicate the diagnostic performance of the index test.[2]

Diagnostic tests in research often encounter errors and bias, which can complicate clinical decision-making. Meta-analyses address this by synthesizing data from multiple studies to provide reliable assessments of diagnostic performance, mitigating errors, and accounting for study heterogeneity. Meta-analysis is a statistical method for combining evidence from multiple studies that aim to address the same research question.[3] Diagnostic test meta-analyses are crucial for evaluating the performance of specific diagnostic tests by combining data from multiple DTA studies.[4] This method considered the highest level of study design in medical research, enabling clinicians to make informed decisions based on robust evidence.[5] In an aggregate data meta-analysis (ADMA) of DTA studies, information on the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) outcomes for a particular diagnostic test across multiple studies is collected.[6] From this data, one can derive study-specific observed values for Se, Sp, and other diagnostic accuracy measures. Researchers then combine data from multiple studies to generate comprehensive estimates of these test accuracy measures, taking into account the variability and potential biases of each study.

Various statistical models for meta-analysis of DTA have been developed over the past two decades, each with its strengths and weaknesses. Bivariate random-effects and hierarchical models have also been developed during the last two decades to synthesize DTA measures, particularly logit-transformed Se and Sp. Reitsma et al.'s[7] approach is easier to use and less complicated than Rutter and Gatsonis's[8] hierarchical model. Both models are parameterization of one another without covariates, and when applied to the same data, they produce similar results (Negeri et al.).[9] However, their development and parameter estimation methods are different: Rutter and Gatsonis' model uses a Bayesian approach, whereas Reitsma et al.'s model uses a classical Frequentist approach. The standard bivariate linear mixed model (BLMM) of Reitsma et al.[7] meta-analyzes DTA using maximum likelihood estimates (MLEs) through numerical optimization methods, as the model lacks a closed-form solution for

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

the between-study heterogeneity parameter. A bivariate random effects model that utilizes the logit transformation of Se and Sp, and accounts for both within- and between-study heterogeneity is commonly used to make statistical inferences about the unknown test characteristics. However, it has been well reported that this model may lead to misleading inference,[6,10–12] particularly with small sample sizes or zero cell counts for TPs, TNs, FNs and FPs, requiring an arbitrary Haldane–Anscombe correction[13,14] when the data are sparse. Also, the logit transformed Se and Sp within each study are approximated to a normal distribution,[15] and these assumptions might not be valid in certain cases, such as small sample sizes or rare events. Alternative transformations which do not require Haldane–Anscombe correction, such as the arcsine square root (ASR) and Freeman–Tukey double arcsine (FTDA), were recently proposed by Negeri et al.[9] in the context of bivariate ADMA of DTA studies. However, these methods also suffer from using approximate or asymptotic within-study variance estimates, which can only be justified when within-study sample sizes are large.

Chu and Cole[15] developed a bivariate generalized linear mixed model (BGLMM) to avoid the within-studies normality assumption required by the BLMM of Reitsma et al.[7] The BGLMM uses the exact binomial distribution and does not require Haldane–Anscombe correction when any of the four cell frequencies in a DTA study contain zero counts.[6] It provides unbiased and efficient estimates,[16] especially with sparse data or small sample sizes, without requiring logit transformation or normality assumptions to describe within-study variation in test characteristics.[17] However, the BGLMM for sparse data meta-analyses may fail to converge or provide unreliable parameters, leading to overestimated results due to publication bias.[4] Additionally, if only a few studies exist, the BGLMM may underestimate variance and is not recommended due to variations in thresholds, spectrum effect, and systematic error.[4] Moreover, the BGLMM may cause convergence issues during parameter estimation, especially when dealing with a large number of parameters or a limited number of studies.[18]

To address the limitations of existing methods, we propose an exact or analytical within-study variance calculation approach that does not require a Haldane–Anscombe correction and remains invariant to transformations. To the best of our knowledge, there is no literature evaluating the performance of this analytical method in the context of an ADMA of DTA studies. The proposed analytical method avoids convergence issues and addresses the shortcomings of the standard BLMM, which often necessitates a Haldane–Anscombe correction when small or zero cell counts are present in the DTA data. To fill this research gap, this study aimed to evaluate the performance of the proposed exact within-study variance calculation method compared to the existing standard approaches for ADMA of DTA studies. A comprehensive simulation study evaluated the new exact method and the existing asymptotic approaches in terms of absolute bias, root mean squared error (RMSE), coverage probability, and 95% confidence interval (CI) width.

The remainder of this article is structured as follows: Section 2 presents motivating examples using two real-life datasets. Section 3 introduces the statistical methods and our simulation study design. In Section 4, we discuss the outcomes of our simulation study and illustrate the proposed within-study variance calculation method using two real-life datasets. Finally, we give a summary and concluding thoughts in Section 5.

2. Motivating examples

This section provides an overview of two real-life datasets from Doria et al.[19] and Arevalo-Rodriguez et al.[20] to motivate the statistical methods presented in Section 3. Section 4 will analyze these datasets further to demonstrate the applicability of the techniques covered in this manuscript. These datasets were selected to compare the proposed within-study variance calculation method with standard approaches when data with zero cell counts are present and absent. We display a forest plot for each dataset showing the studies under consideration, their 95% CIs, and test characteristics (i.e., observed Se or Sp).

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

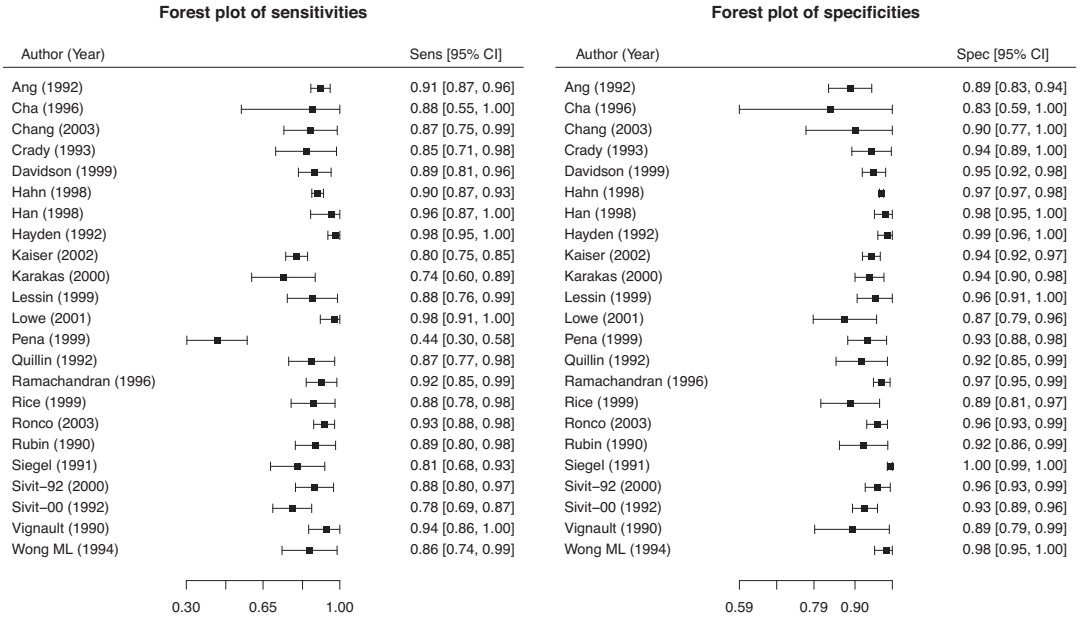


Figure 1. *Forest plot for sensitivity (left) and specificity (right) of the US-Children data obtained after a Haldane–Anscombe correction is applied.*

2.1. The US-Children data

The first dataset is about the ultrasonography test. Doria et al.[19] examined the effectiveness of this test in diagnosing appendicitis in both children and adults. The children-specific version of the ultrasonography data (US-Children) includes 23 studies, with an average of 77 children with appendicitis and 254 children without appendicitis across the studies (see Table A.1 in the Supplementary Material). In this dataset, there is an indication of sparsity in the counts of both FNs and FPs. Specifically, three of the primary studies (Cha et al.[21]; Lowe et al.[22]; Han[23]) reported an observed Se of 100% (i.e., had FNs=0) and one study (Siegel et al.[24]) reported an observed Sp of 100% (i.e., had FP=0). The forest plot of the US-Children data is given in Figure 1.

2.2. The mini-mental state examination data

The second meta-analysis is from the review by Arevalo-Rodriguez et al.,[20] which examined the mini-mental state examination (MMSE) test for detecting Alzheimer's disease and dementia in people with mild cognitive impairment. The MMSE data includes eight studies, with an average of 47 participants with and 95 without the condition (see Table A.2 in the Supplementary Material). This meta-analysis showed no indication of sparsity in either counts of FNs or FPs; that is, none of the included primary studies had observed Se or Sp values close to 0 or 1. Figure 2 displays the forest plot of the MMSE data.

3. Methods

Given the observed data for any DTA study as in Table 1, we can easily estimate the diagnostic test's accuracy parameters of interest as follows: $\widehat{Se} = TP/n_1$ and $\widehat{Sp} = TN/n_2$, where n_1 and n_2 are the study-specific sample sizes that denote the total number of diseased and non-diseased individuals, respectively. In the following sections, we describe the standard methods for ADMA of DTA studies assuming the data structure presented in Table 1.

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

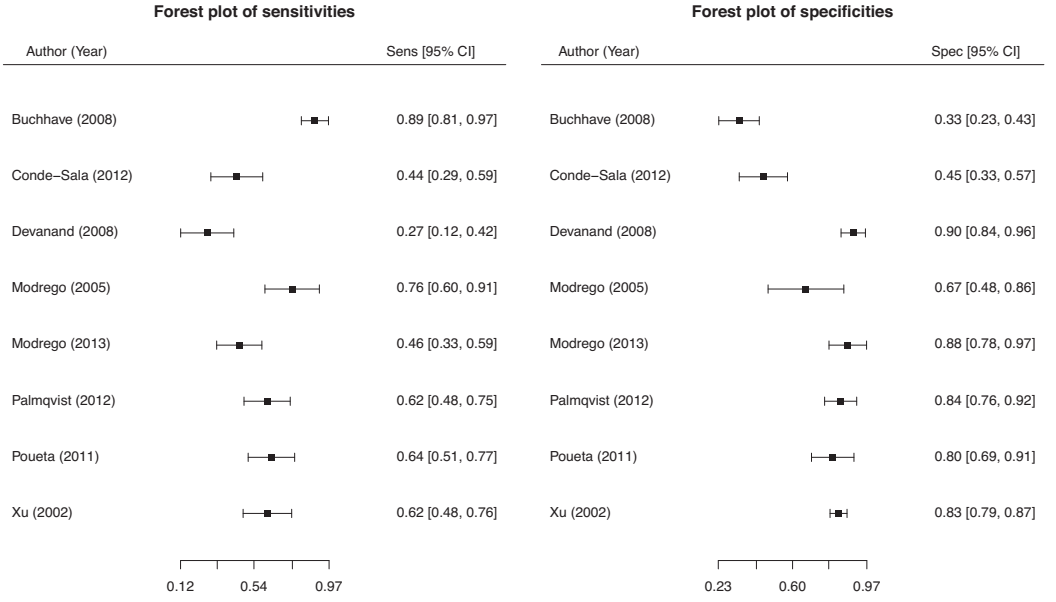


Figure 2. *Forest plot for sensitivity (left) and specificity (middle) of the Mini-Mental State Examination (MMSE) data obtained without applying a Haldane–Anscombe correction.*

Table 1. *Data structure of a diagnostic test result for a single study.*

		True disease status	
		Diseased	Non-diseased
Test result	Positive	TP	FP
	Negative	FN	TN
	Total (n)	$n_1 = TP + FN$	$n_2 = FP + TN$

3.1. The bivariate linear mixed-effects model

The BLMM of Reitsma et al.[7] models both the within- and between-study variability in test characteristics using the bivariate normal distribution.

According to the BLMM, the observed vector of response for study i, $\mathbf{y}_i = (y_{1i}, y_{2i})^T = \left[\mathrm{g}(\widehat{Se_i}), \mathrm{g}(\widehat{Sp_i})\right]^T$, is modeled within-study in the following way:

$$\mathbf{y}_i|\boldsymbol{\mu}_i \sim N_2\left[\boldsymbol{\mu}_i = \begin{pmatrix}\phi_{1i}\\ \phi_{2i}\end{pmatrix}, \boldsymbol{S_i} = \begin{pmatrix}s_{1i}^2 & 0\\ 0 & s_{2i}^2\end{pmatrix}\right], \quad i = 1, \ldots, k, \tag{3.1}$$

where g is any transformation, such as the logit, $\boldsymbol{\phi}_{1i}$ denotes the true study-specific $\mathrm{g}(Se_i)$, $\boldsymbol{\phi}_{2i}$ denotes the true study-specific $\mathrm{g}(Sp_i)$, $\mathbf{S}_i = \begin{pmatrix}s_{1i}^2 & 0\\ 0 & s_{2i}^2\end{pmatrix}$represents the assumed to be known within-study variance–covariance matrix of the response vector $\mathbf{y}_i$, and k denotes the number of studies. In practice, however, $\boldsymbol{S_i}$ is estimated from the data using the first-order delta method (i.e., for logit, ASR, and FTDA).

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

At the second stage, the BLMM assumes heterogeneity between the true study-specific $g(Se_i)$ and $g(Sp_i)$. Thus,

$$\boldsymbol{\mu}_i = \begin{pmatrix} \phi_{1i} \\ \phi_{2i} \end{pmatrix} \sim \mathrm{N}_2 \left[\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right], \quad i = 1, 2, \ldots, k, \tag{3.2}$$

where $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ denotes the true mean of $g(Se_i)$ and $g(Sp_i)$, and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$is the between-study covariance matrix, which denotes the heterogeneity parameter such that σ_1^2 and σ_2^2 are the variances of ϕ_{1i} and ϕ_{2i}, respectively, and σ_{12} is the covariance between ϕ_{1i} and ϕ_{2i}.

By combining (3.1) and (3.2) and assuming that the studies are independent, we arrive at the marginal model

$$\mathbf{y}_i \sim N_2 \left(\boldsymbol{\mu}, \boldsymbol{\Sigma}_i \right), i = 1, \ldots, k, \tag{3.3}$$

where $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma} + \boldsymbol{S}_i = \begin{pmatrix} s_{1i}^2 + \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & s_{2i}^2 + \sigma_2^2 \end{pmatrix}$.

3.1.1. The approximate or asymptotic variance estimation methods

The standard BLMM [7] employs the logit transformation to model the observed Se and Sp using the bivariate normal distribution. Thus, for the within-study variance of a logit-transformed proportion estimate, $g(\hat{p}_{ji}) = \mathrm{logit}(\hat{p}_{ji}), i = 1, \ldots, k$, usually the delta method based on the first-order Taylor series approximation is used[9,25]:

$$\begin{aligned} s_{ji}^2 &= \mathrm{Var}[\mathrm{logit}(\hat{p}_{ji})] \\ &\approx \mathrm{Var}(\hat{p}_{ji}) \left\{ \frac{\partial}{\partial \hat{p}_{ji}} [\mathrm{logit}(\hat{p}_{ji})] \right\}^2 \\ &= \mathrm{Var}(\hat{p}_{ji}) \left\{ \frac{\partial}{\partial \hat{p}_{ji}} \left[\log \left(\frac{\hat{p}_{ji}}{1 - \hat{p}_{ji}} \right) \right] \right\}^2 \\ &= \frac{\hat{p}_{ji}(1 - \hat{p}_{ji})}{n_{ji}} \left[\frac{1}{\hat{p}_{ji}(1 - \hat{p}_{ji})} \right]^2 \\ &= \frac{1}{n_{ji} \hat{p}_{ji}(1 - \hat{p}_{ji})}, \text{ for } j = 1, 2 \text{ and } i = 1, \ldots, k. \end{aligned}$$

The above expression implies that the variance estimate of $\mathrm{logit}(\hat{p}_{ji})$ is defined only if the $\hat{p}_{ji}$ is different from 0 and 1; that is, only if there are no zero cell counts in Table 1. Otherwise, usually, 0.5 is added to all cells of the 2×2 contingency table. Negeri et al.[9] proposed two additional transformations, the ASR and FTDA transformations,[26] to overcome the limitations of the default logit transformation used by the BLMM. The respective approximate within-study variances for the ASR and FTDA transformations were given by $\frac{1}{4n_{ji}}$ and $\frac{1}{4n_{ji}+2}$, for $j = 1, 2$ and $i = 1, \ldots, k$.

3.1.2. The proposed exact or analytical variance calculation method

All three of the above within-study variance estimation methods are based on asymptotic results that can only be valid for a large number of studies in a meta-analysis. Moreover, the standard BLMM with the logit transformation requires a Haldane–Anscombe correction for sparse studies in a meta-analysis. Therefore, we propose a within-study variance calculation approach based on exact or analytical results[27] for ADMA, which works for any transformation and meta-analysis size. The exact variance $\mathrm{Var}[g(X)]$ for a given DTA frequency count $TP = X$ or $TN = X$ of a function g can be calculated using

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

the theoretical variance formula as

$$\mathrm{Var}[\mathrm{g}(X)] = E[\mathrm{g}(X)^2] - [E(\mathrm{g}(X))]^2, \tag{3.4}$$

where the function g can be any transformation, including the logit, ASR, or FTDA. Assuming the Binomial distribution for the counts in Table 1, given the study-specific sensitivity ($p_{1i} = \frac{X}{n_{1i}}$) or specificity ($p_{2i} = \frac{X}{n_{2i}}$) and the total number of participants with disease (n_{1i}) and without disease (n_{2i}), we define our within-study variance formula using the following finite Binomial sums:

$$s_{ji}^2 = \mathrm{Var}\left[g\left(p_{ji} = \frac{X}{n_{ji}}\right)\right] = \sum_{x=0}^{n_{ji}} \left[g\left(\frac{x}{n_{ji}}\right)\right]^2 \binom{n_{ji}}{x} p_{ji}^x (1-p_{ji})^{n_{ji}-x} - \left\{\sum_{x=0}^{n_{ji}} g\left(\frac{x}{n_{ji}}\right)\binom{n_{ji}}{x} p_{ji}^x (1-p_{ji})^{n_{ji}-x}\right\}^2, \tag{3.5}$$

for $j = 1, 2$ and $i = 1, \ldots, k$, where the function $g(p_{ji})$ is defined as

$$g(p_{ji}) = \begin{cases} \log\left(\dfrac{p_{ji}}{1-p_{ji}}\right), & \text{logit transformation,} \\ \sin^{-1}\left(\sqrt{p_{ji}}\right), & \text{ASR transformation,} \\ \dfrac{1}{2}\left\{\sin^{-1}\left(\sqrt{\dfrac{n_{ji}p_{ji}}{n_{ji}+1}}\right) + \sin^{-1}\left(\sqrt{\dfrac{n_{ji}p_{ji}+1}{n_{ji}+1}}\right)\right\}, & \text{FTDA transformation.} \end{cases}$$

In addition to the benefits listed above, unlike the BLMM of Reitsma et al.,[7] the proposed analytical variance formula does not fail to converge or require Haldane–Anscombe correction for sparse studies in a meta-analysis. However, the proposed approach with logit transformation requires a Haldane–Anscombe correction for the smallest and largest value in the binomial range, that is, only when $x = 0$ or $x = n_{ji}$.

3.2. The bivariate generalized linear mixed model

For completeness, we also describe the BGLMM of Chu and Cole[15] as a comparator model to the BLMM and the proposed within-study variance calculation approach. In a BGLMM for ADMA, the variability within the study for $i = 1, 2, \ldots, k$ is modeled as

$$\begin{aligned} TP_i \mid Se_i &\sim Binomial\,(n_{1i}, Se_i) \\ TN_i \mid Sp_i &\sim Binomial\,(n_{2i}, Sp_i)\,, \end{aligned} \tag{3.6}$$

where the linear component and random effects have the form:

$$\mathrm{g}\,(Se_i) = \mu_1 + \phi_{1i} \text{ and } \mathrm{g}\,(Sp_i) = \mu_2 + \phi_{2i}$$

and

$$\boldsymbol{\mu}_i = \begin{pmatrix} \phi_{1i} \\ \phi_{2i} \end{pmatrix} \sim \mathrm{N}_2\left[\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right]. \tag{3.7}$$

The BGLMM is recommended over the BLMM, as it performs better, especially for small meta-analyses and for values of Se or Sp closer to 1 or 0.[11] Hence, it is widely accepted as the preferred model

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

for ADMA of DTAs, although it also suffers from convergence issues unless appropriate numerical optimization methods are used.[6]

3.3. Parameter estimation

For the BLMM examined in this study, five parameters must be estimated to make statistical inferences: two fixed effects (μ_1 and μ_2) and three random effects (σ_1^2, σ_2^2, and σ_{12}). For k independent studies, observed sensitivities ($\widehat{Se_i}$) and specificities ($\widehat{Sp_i}$); the adjusted likelihood function from which the maximum likelihood estimators are derived is given as follows.

The likelihood function $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given by

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu})\right] \\ &= (2\pi)^{-k/2} \prod_{i=1}^{k} |\boldsymbol{\Sigma}_i|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu})\right]. \end{aligned}$$

The log-likelihood function is

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{k}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{k}(\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}). \tag{3.8}$$

To estimate the parameters of the bivariate random effects models, we employed maximum likelihood estimation via numerical optimization. Initial parameter values were set using empirical means and variance–covariance estimates from transformed data, providing data-driven starting parameter values for each model. Once the estimate of the between-study covariance matrix, $\hat{\boldsymbol{\Sigma}}$, is obtained through numerical methods, the parameter of interest, $\boldsymbol{\mu}$, is estimated using the weighted average method, which is equivalent to the MLE:

$$\hat{\boldsymbol{\mu}} = \left(\sum_{i=1}^{k}(\hat{\boldsymbol{\Sigma}} + \boldsymbol{S}_i)^{-1}\right)^{-1} \sum_{i=1}^{k}(\hat{\boldsymbol{\Sigma}} + \boldsymbol{S}_i)^{-1}\mathbf{y}_i.$$

Asymptotically, the estimated average effect size, $\widehat{\boldsymbol{\mu}}$, is approximately distributed as bivariate normal with mean $\boldsymbol{\mu}$ and variance–covariance, VARCOV $= \left(\sum_{i=1}^{k}(\hat{\boldsymbol{\Sigma}} + \boldsymbol{S}_i)^{-1}\right)^{-1}$. Therefore, the approximate $100(1-\alpha)\%$ CI for μ_j can be obtained as

$$\left(\hat{\mu}_j - Z_{\frac{\alpha}{2}}\sqrt{\text{VARCOV}(j,j)}, \quad \hat{\mu}_j + Z_{\frac{\alpha}{2}}\sqrt{\text{VARCOV}(j,j)}\right),$$

where $Z_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$th percentile of the standard normal distribution and VARCOV(j,j) is the jth diagonal element of VARCOV.

For the BGLMM, its marginal likelihood function can be written as

$$\begin{aligned} L\left(TP_i, TN_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \prod_{i=1}^{k} \int_0^1 \int_0^1 &\text{Binomial}\left(n_{1i}, Se_i\right) \text{Binomial}\left(n_{2i}, Sp_i\right) \\ &\times \phi_2\left(g(Se_i), g(Sp_i); \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \partial Se_i \, \partial Sp_i. \end{aligned} \tag{3.9}$$

However, the marginal likelihood function in (3.9) does not have a closed-form solution. Therefore, numerical algorithms like the Laplace approximation (LA),[28,29] iteratively reweighted least square

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

Table 2. *True parameter settings for the simulation study.*

Parameter	Setting 1	Setting 2	Setting 3	Setting 4
(Se, Sp)	$(0.70, 0.80)$	$(0.80, 0.90)$	$(0.95, 0.99)$	–
(σ_1^2, σ_2^2)	$(0.90, 0.55)$	$(1.51, 1.00)$	$(1.59, 1.83)$	–
σ_{12}	-0.03	-0.34	-0.70	–
(n_1, n_2)	$(50, 100)$	$(100, 200)$	$(200, 300)$	–
k	5	15	25	50

(ILRS),[30] and adaptive Gauss–Hermite quadrature (AGHQ),[28] which numerically integrate the likelihood, were used.[6] Since a recent study by Zhao et al.[6] recommended against using the LA in situations where data sparsity is a concern and suggested using either the IRLS or AGHQ regardless of data characteristics, we decided to use the IRLS as the primary computational method for this manuscript since it had a higher convergence rate than the AGHQ.

Furthermore, we constructed the summary receiver operating characteristics (SROC) curve and its area under the SROC curve (AUC) for the two motivating examples, following the approach of Chu et al.[17] Under the assumption of bivariate normality of $[g(Se_i), g(Sp_i)]$, the expected sensitivity for a chosen specificity on the transformed scale is given by

$$g(Se_i) = \mu_1 - \frac{\sigma_{12}}{\sigma_2^2}\mu_2 + \frac{\sigma_{12}}{\sigma_2^2}g(Sp_i). \tag{3.10}$$

Accordingly, the expected AUC is approximated by

$$AUC = \int_0^1 g^{-1}\left\{\mu_1 - \frac{\sigma_{12}}{\sigma_2^2}\mu_2 + \frac{\sigma_{12}}{\sigma_2^2}g(Sp_i)\right\} dSp_i. \tag{3.11}$$

3.4. Simulation study design

In this study, we developed a four-step simulation based on a strategy proposed by Zhao et al.[6] We considered three (small, medium, and large) pairs of true mean sensitivity and specificity (Se, Sp), the true between-study variances (σ_1^2, σ_2^2) and covariance (σ_{12}), the true sample sizes (n_1, n_2) for each study, and four different numbers of studies in the meta-analysis (k) as the true parameters. The specific settings are as given in Table 2. Accordingly, we explored a total of $3^4 \times 4 = 324$ scenarios in our simulation investigation and for each combination of parameters, DTA datasets were generated using three true models: bivariate normal distribution using the logit, ASR, and FTDA transformation, following the methodology outlined by Negeri et al.[9] The simulation was run 1000 times using different seeds for each replication to ensure reproducibility. We evaluated the models in terms of accuracy (bias), precision (RMSE), coverage (95% empirical coverage probability), and 95% CI width.

3.5. Software

Three `R` packages were utilized for data simulation and statistical analyses. The `faraway` package[31] was used to both transform and back-transform the simulated Se and Sp within the models. The `mada` package[32] was used to create forest plots for the real-life data examples. Bivariate data from the bivariate normal distribution were simulated using the `mvrnorm` function from the `MASS` package.[33,34] We applied the IRLS with nAGQ equal to 0 using the `glmer()` function from the `lme4` R package[35] to fit the BGLMM. We wrote an `R` code to fit the BLMM and proposed models.

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

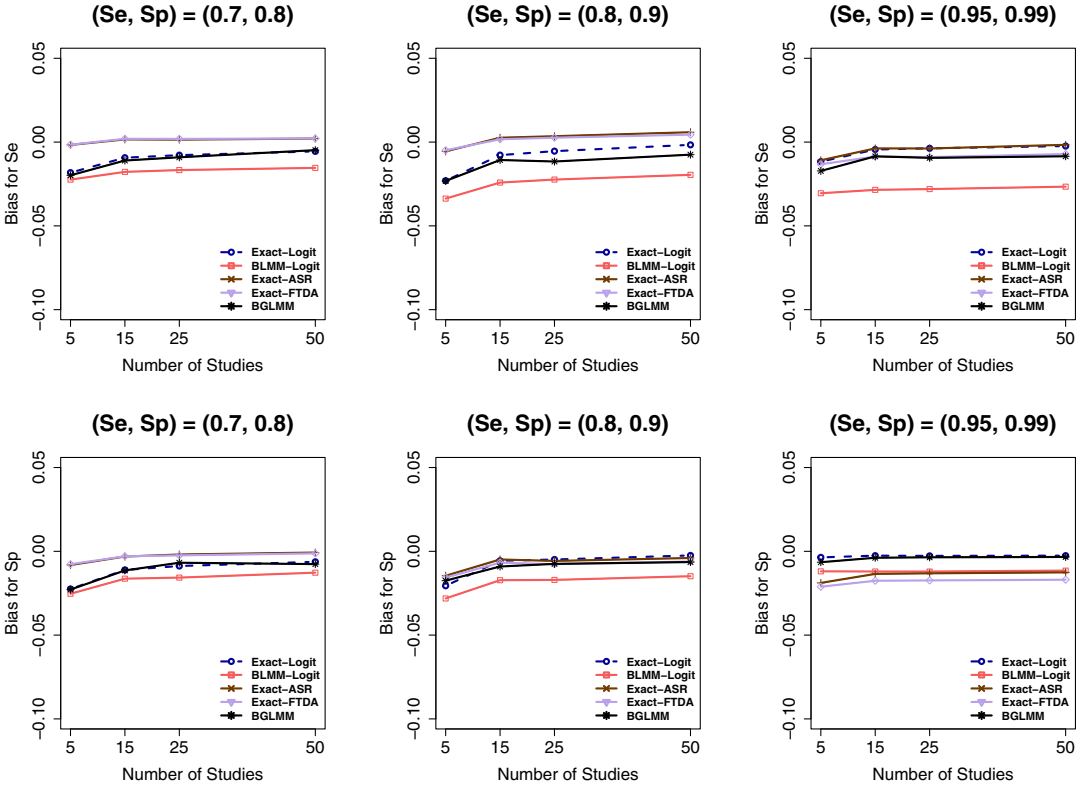


Figure 3. *Bias for sensitivity (Se) and specificity (Sp) when* $\sigma_1^2 = 1.59$, $\sigma_{12} = -0.03$, $\sigma_2^2 = 1.83$, $n_1 = 50$, *and* $n_2 = 100$.

4. Results

4.1. Simulation study results

This Section presents simulation results for scenarios characterized by small sample sizes per group, large between-study variability in logit-sensitivity and logit-specificity, and a small negative correlation between them, reflecting characteristics of our real-life data examples. Simulation results for the remaining scenarios are provided in the Supplementary Material.

Figure 3 displays the bias of the overall Se and Sp estimators for the proposed exact within-study variance calculation method and the existing approaches, given the true parameters $\sigma_1^2 = 1.59$, $\sigma_2^2 = 1.83$, and $\sigma_{12} = -0.7$. The BLMM based on the proposed exact within-study variance calculation method and arcsine transformations (Exact-ASR and Exact-FTDA) yielded unbiased estimates of overall Se and Sp when the meta-analysis did not contain sparse primary studies (Figure 3, first two panels). When the meta-analysis had sparse primary studies (Figure 3, third panel), while the Exact-Logit had the smallest bias for Se, the Exact-Logit and the BGLMM had the smallest bias for Sp. Conversely, the standard approximate BLMM-Logit had the largest negative bias for sensitivity and specificity, irrespective of the presence of sparse primary studies in a meta-analysis. Although similar results were observed when the number of participants was moderate ($n_1 = 100, n_2 = 200$; Figure B.20 in the Supplementary Material) or large ($n_1 = 200, n_2 = 300$; Figure B.21 in the Supplementary Material), the BLMM-Logit model's bias diminished with increasing sample sizes, as expected.

Figure 4 depicts the RMSE for each method's overall Se and Sp estimator. Relatively all methods had comparable RMSEs when there were no sparse primary studies in a meta-analysis (Figure 4, first two panels). However, the proposed Exact-Logit method and the standard BGLMM had the smallest

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

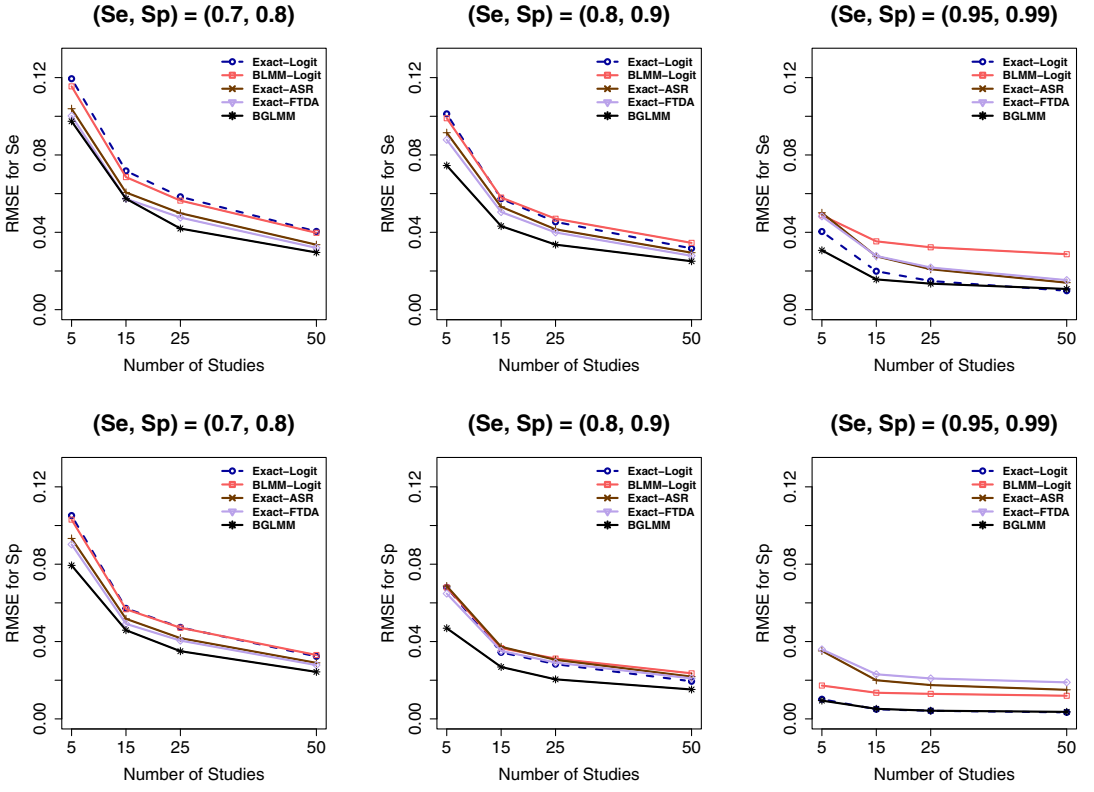


Figure 4. *RMSE for sensitivity (Se) and specificity (Sp) when* $\sigma_1^2 = 1.59$, $\sigma_{12} = -0.03$, $\sigma_2^2 = 1.83$, $n_1 = 50$, *and* $n_2 = 100$.

RMSE for *Se* and *SP* when there were sparse primary studies in a meta-analysis (Figure 4, third panel). Conversely, the standard approximate BLMM-Logit model had the largest RMSE for *Se* and *SP* for the latter scenario. Consistent results were observed for moderate (Figure C.20 in the Supplementary Material) and large (Figure C.21 in the Supplementary Material) numbers of participants in a meta-analysis study, although, as expected, the RMSEs improved with an increasing number of studies in the meta-analysis.

All methods had comparable 95% empirical coverage probability (Figure 5) when there were no sparse primary studies in the meta-analysis (first two panels), regardless of the other parameters varied in our simulation. However, when there were sparse primary studies in a meta-analysis (third panel), the proposed exact methods (Exact-Logit, Exact-ASR, and Exact-FTDA) and the BGLMM had superior coverage probabilities, achieving the nominal 95%, especially in *Se*, compared to the traditional approximate BLMM-Logit method, which had a poor coverage probability for this scenario. Similar trends were observed when the number of participants in a meta-analytic study was moderate (Figure D.20 in the Supplementary Material) or large (Figure D.21 in the Supplementary Material), except that the coverage probabilities of the asymptotic BLMM-Logit improved with increasing sample sizes, as expected.

In terms of 95% CI widths (Figure 6), we found that all methods had comparable average widths of the 95% CIs when the number of studies in a meta-analysis was large and the meta-analysis contained no sparse primary studies (Figure 6, first two panels). However, the standard asymptotic BLMM-Logit, the proposed Exact-Logit, and the BGLMM had narrower 95% CI widths when there were sparse primary studies in a meta-analysis (third panels). The Exact-ASR and Exact-FTDA had the narrowest 95% CI widths when there were no sparse primary studies and the number of studies was small (first

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

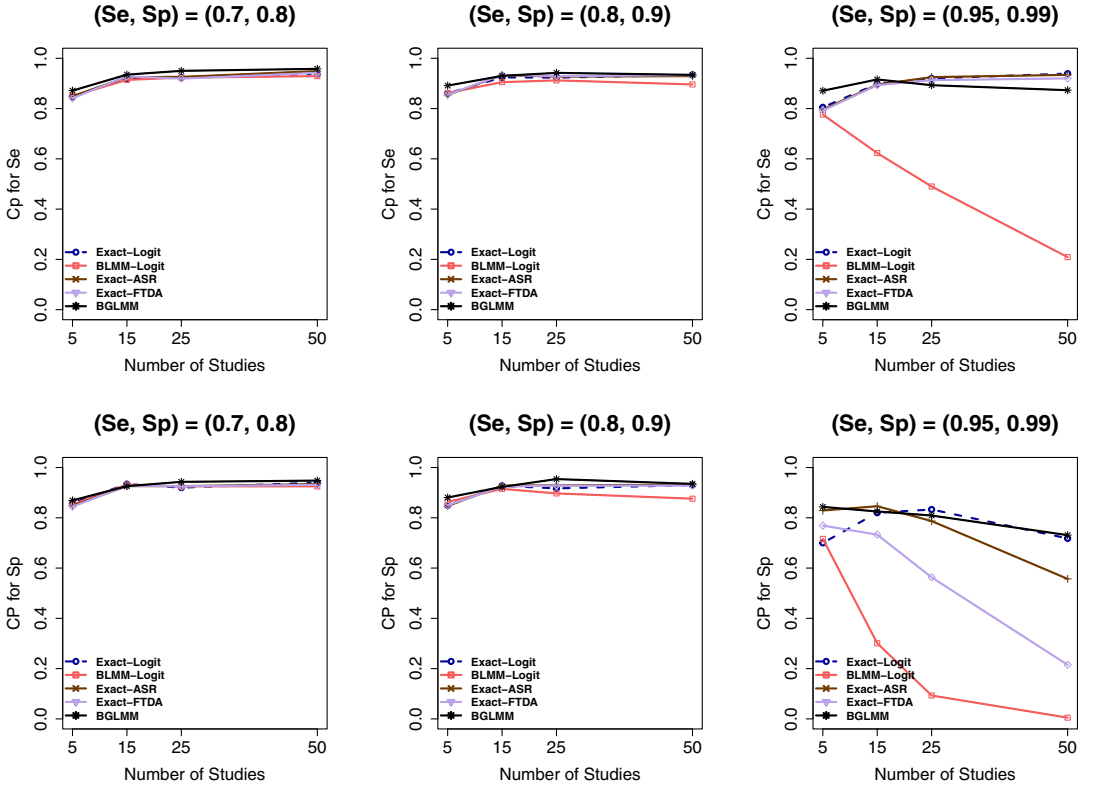


Figure 5. *Coverage probability for sensitivity (Se) and specificity (Sp) when* $\sigma_1^2 = 1.59$, $\sigma_{12} = -0.03$, $\sigma_2^2 = 1.83$, $n_1 = 50$, *and* $n_2 = 100$.

two panels), regardless of the varied parameters in our simulations. Moreover, we observed that models based on the arcsine transformations (Exact-ASR and Exact-FTDA) had the widest 95% CI widths for meta-analyses with sparse primary studies, irrespective of the number of participants or studies in a meta-analysis. This trend was also observed for moderate (Figure E.20 in the Supplementary Material) and large (Figure E.21 in the Supplementary Material) sample sizes in a meta-analysis study.

In terms of the bias of the between-study variances σ_1^2 and σ_2^2, the proposed exact variance calculation methods (Exact-Logit, Exact-ASR, and Exact-FTDA) and the BGLMM outperformed the standard BLMM-Logit across all scenarios (Figure 7). The standard BLMM-Logit method significantly underestimated σ_1^2 and σ_2^2 across all scenarios, whereas the other methods provided more accurate estimates. Although the Exact-Logit method also underestimated σ_2^2 when there were sparse primary studies in a meta-analysis, its bias decreased for moderate or large number of participants per study, as expected (Figures F.20 and F.21 in the Supplementary Material).

Considering the RMSE of σ_1^2 and σ_2^2 (Figure 8), the proposed exact method using the arcsine transformations (Exact-ASR and Exact-FTDA) had the least RMSE for σ_1^2 and σ_2^2, irrespective of the presence or absence of sparse primary studies in a meta-analysis. However, while the standard asymptotic BLMM-Logit method had the worst RMSE when a meta-analysis contained sparse primary studies, the proposed Exact-Logit method and the BGLMM had the second least RMSEs. Similar trends were observed when the number of participants was moderate or large (Figures G.20 and G.21 in the Supplementary Material), which also revealed diminishing RMSEs as the sample sizes increased.

Regarding bias and RMSE of the between-study covariance (σ_{12}), all methods had comparable bias when there were no sparse primary studies (Figure 9, top row, first two panels). For meta-analysis with

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

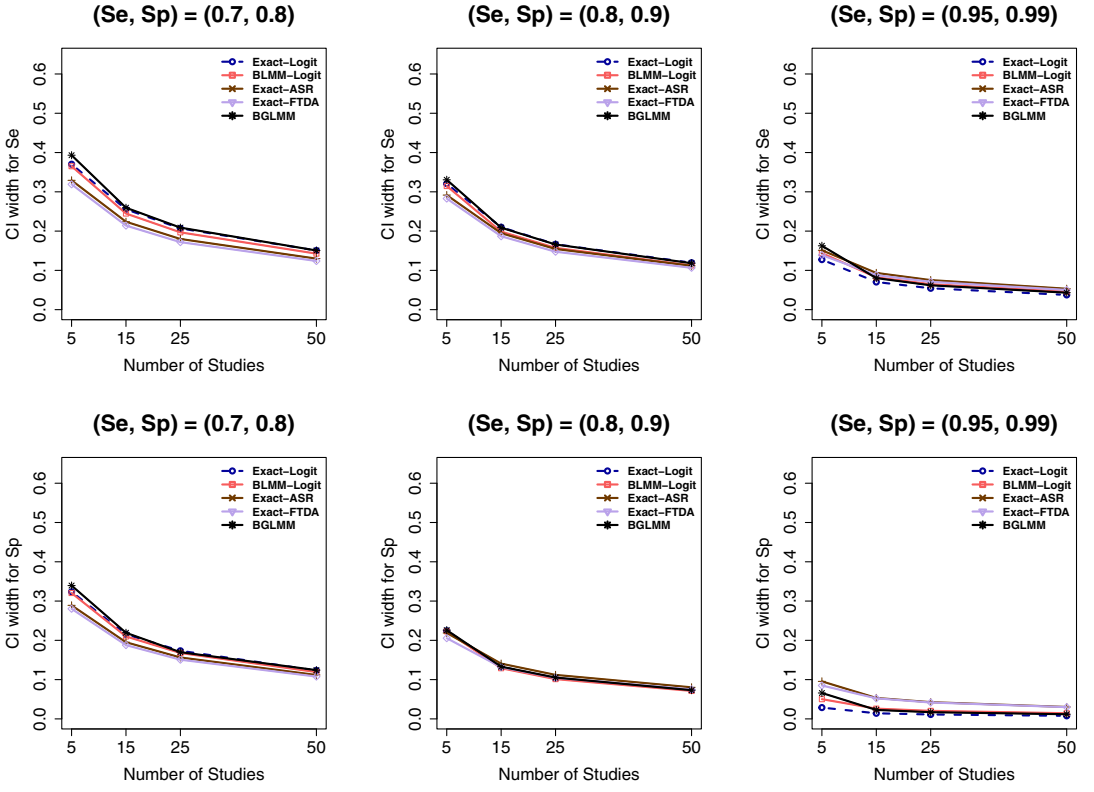


Figure 6. *CI widths for sensitivity (Se) and specificity (Sp) when* $\sigma_1^2 = 1.59$, $\sigma_{12} = -0.03$, $\sigma_2^2 = 1.83$, $n_1 = 50$, *and* $n_2 = 100$.

sparse primary studies, the proposed Exact-ASR and Exact-FTDA had the least bias, and the traditional approximate BLMM-Logit and the Exact-Logit had the largest bias. On the other hand, whereas the Exact-ASR and Exact-FTDA had the smallest RMSE, the Exact-Logit, BLMM-Logit, and BGLMM had the largest RMSE for σ_{12}. Similar patterns were observed when a meta-analysis study contained a moderate to large number of participants (Figures H.20 and H.21 in the Supplementary Material), which also resulted in improved bias and RMSE.

4.2. Illustrative examples

This section demonstrates the proposed exact within-study variance calculation method using the two datasets presented in Section 2. The datasets were chosen to highlight the performances of the proposed and existing approaches when a meta-analysis contains and does not contain sparse primary studies. It is essential to note that, to estimate the pooled sensitivity and specificity for both datasets using the standard approximate BLMM-Logit method, a Haldane–Anscombe correction was consistently applied whenever a primary study had a cell with zero counts, ensuring that the estimates and their variances exist. Without this correction, the model would not converge. To compare the within-study variance estimation methods, we calculated and presented the estimates of mean Se and mean Sp, along with their corresponding 95% CIs, and the between-study variance–covariance terms.

4.2.1. The US-Children data

This meta-analysis included three studies with zero cell counts for FN and one study with a single zero cell count for FP. Table 3 demonstrates that the traditional approximate BLMM-Logit approach's

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

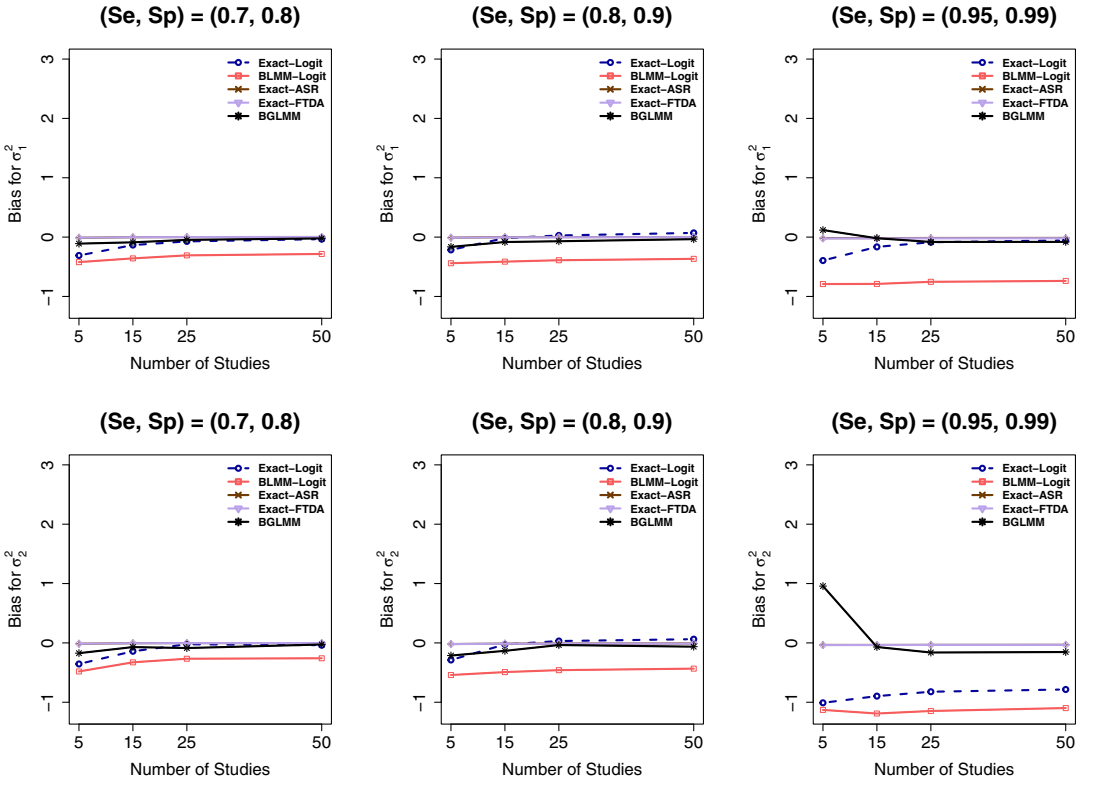


Figure 7. *Bias for* σ_1^2 *and* σ_2^2 *when the true* $\sigma_1^2 = 1.59$, $\sigma_{12} = -0.03$, $\sigma_2^2 = 1.83$, $n_1 = 50$, *and* $n_2 = 100$.

estimated pooled *Se* and *Sp* were lower by at least 2% and 1%, respectively, than the other methods, including the proposed exact approaches and the BGLMM. This result aligns with our simulation results, which suggested that the standard asymptotic BLMM-Logit method underestimates both *Se* and *Sp* in meta-analysis with sparse primary studies. Similarly, the approximate BLMM-Logit method yielded lower estimates of the between-study variances compared to the proposed Exact-Logit method and BGLMM (the two approaches that use the same transformation as the BLMM-Logit), reinforcing our simulation results that the BLMM-Logit method underestimates these parameters for meta-analysis with sparse primary studies. Finally, the three approaches based on the logit transformation (BLMM-Logit, Exact-Logit, and BGLMM) had narrower 95% CI widths than those based on the arcsine-based transformations. This also agrees with why the arcsine-based approaches had ideal coverage probabilities and the BLMM-Logit had less than ideal coverage for sparse meta-analytic datasets in our simulations.

Figure 10 shows the SROC curves for all methods on the US-Children dataset. The standard BGLMM and BLMM-Logit methods had higher AUC and more precise confidence and prediction regions than the proposed Exact-FTDA and Exact-ASR methods for this dataset. The Exact-Logit had the lowest AUC due to the unusually small number of participants in studies with 0 FNs (see Table A.1 in the Supplementary Material), leading to larger between-study variance estimates, since this model does not fully avoid the Haldane–Anscombe correction, as explained in Section 3.1.2.

To compare the methods on studies with 0 FPs or FNs but moderate to large number of participants, we applied all methods and constructed an SROC curve on the AuditC dataset[36] (see Table I.1 in the Supplementary Material). As expected, all methods performed comparably in *Sp* but Exact-Logit and BGLMM methods had higher pooled *Se* and larger σ_1^2 estimates (see Table I.2 in the Supplementary

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

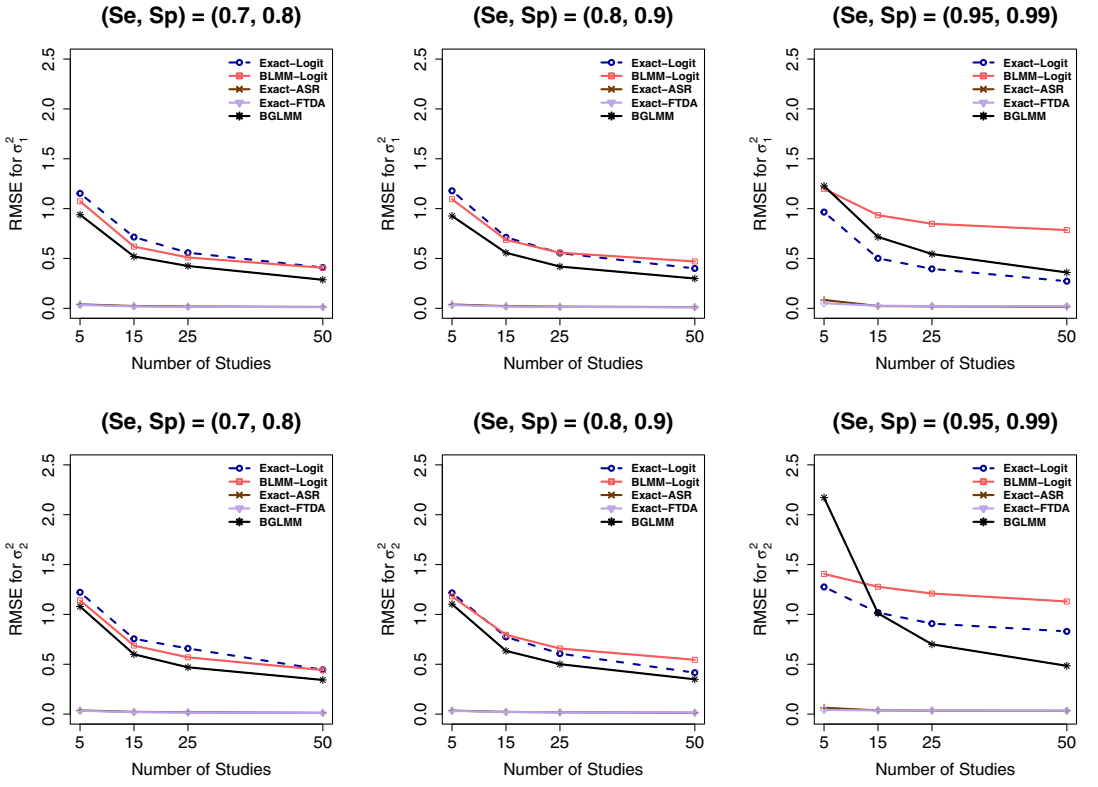


Figure 8. *RMSE for* σ_1^2 *and* σ_2^2 *when the true* $\sigma_1^2 = 1.59$, $\sigma_{12} = -0.03$, $\sigma_2^2 = 1.83$, $n_1 = 50$, *and* $n_2 = 100$.

Table 3. *Estimates of mean Se and mean Sp with their corresponding 95% CI and estimates of the between-study heterogeneity parameters for the US-Children dataset.*

Methods	Se (95% CI)	Sp (95% CI)	σ_1^2	σ_2^2	σ_{12}
BLMM-Logit	0.864 (0.821, 0.897)	0.943 (0.928, 0.955)	0.364	0.161	0.057
BGLMM	0.875 (0.832, 0.908)	0.948 (0.933, 0.960)	0.465	0.234	0.068
Exact-Logit	0.887 (0.843, 0.920)	0.952 (0.930, 0.968)	0.679	0.738	0.018
Exact-ASR	0.899 (0.847, 0.941)	0.952 (0.933, 0.967)	0.032	0.006	0.0002
Exact-FTDA	0.893 (0.849, 0.932)	0.954 (0.937, 0.970)	0.019	0.005	0.0005

Material), mainly due to the pulling effect of the two studies with 100% observed sensitivity (studies 7 and 8). However, the proposed arcsine-based models (Exact-ASR and Exact-FTDA) yielded better SROC curves with more precise confidence and prediction regions (see Figure I.1 in the Supplementary Material).

4.2.2. The mini-mental state examination data

The *MMSE* meta-analysis does not contain primary studies with zero cell counts for FNs or FPs. As a result, all methods had similar estimates of average *Se* and *Sp* along with their 95% CIs (Table 4). Among the methods based on the logit transformation, the BLMM-Logit approach yielded slightly lower between-study variance estimates than the BGLMM, and the Exact-Logit approach had the smallest covariance estimate marginally. Conversely, methods based on the arcsine transformations

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

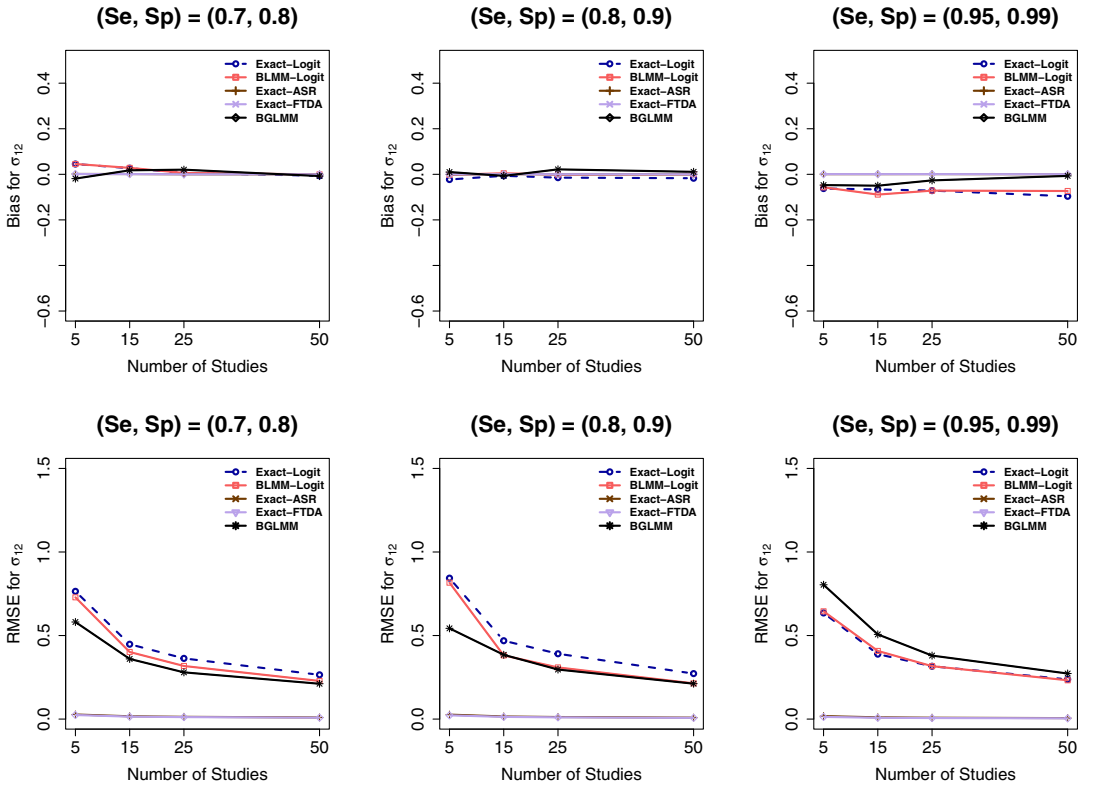


Figure 9. *Bias and RMSE for* σ_{12} *when the true* $\sigma_1^2 = 1.59$, $\sigma_{12} = -0.03$, $\sigma_2^2 = 1.83$, $n_1 = 50$, *and* $n_2 = 100$.

Table 4. *Estimates of mean Se and mean Sp with their corresponding 95% CI and estimates of the between-study heterogeneity parameters for the MMSE dataset.*

Methods	Se (95% CI)	Sp (95% CI)	σ_1^2	σ_2^2	σ_{12}
BLMM-Logit	0.601 (0.457, 0.729)	0.742 (0.589, 0.852)	0.590	0.907	−0.505
BGLMM	0.601 (0.455, 0.732)	0.742 (0.587, 0.853)	0.624	0.935	−0.532
Exact-Logit	0.599 (0.457, 0.725)	0.741 (0.586, 0.852)	0.556	0.907	−0.484
Exact-ASR	0.596 (0.462, 0.723)	0.728 (0.584, 0.852)	0.032	0.045	−0.026
Exact-FTDA	0.596 (0.463, 0.723)	0.729 (0.584, 0.853)	0.030	0.043	−0.025

(Exact-ASR and Exact-FTDA) had similar between-study variances and covariance for the transformed Se and Sp. These results also reinforced our simulation study findings that the methods performed comparably when a meta-analysis does not contain sparse primary studies.

Figure 11 shows the SROC curves for all methods based on the MMSE dataset. Consistent with our findings in Table 4, all methods produced almost identical SROC curves and AUCs. This further demonstrates that the methods are comparable in the absence of sparsity in the meta-analytic data.

5. Discussion and conclusions

This study proposed an exact or analytical within-study variance calculation method for ADMA of DTA studies and investigated its performance compared to existing approaches. Our proposed approach

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

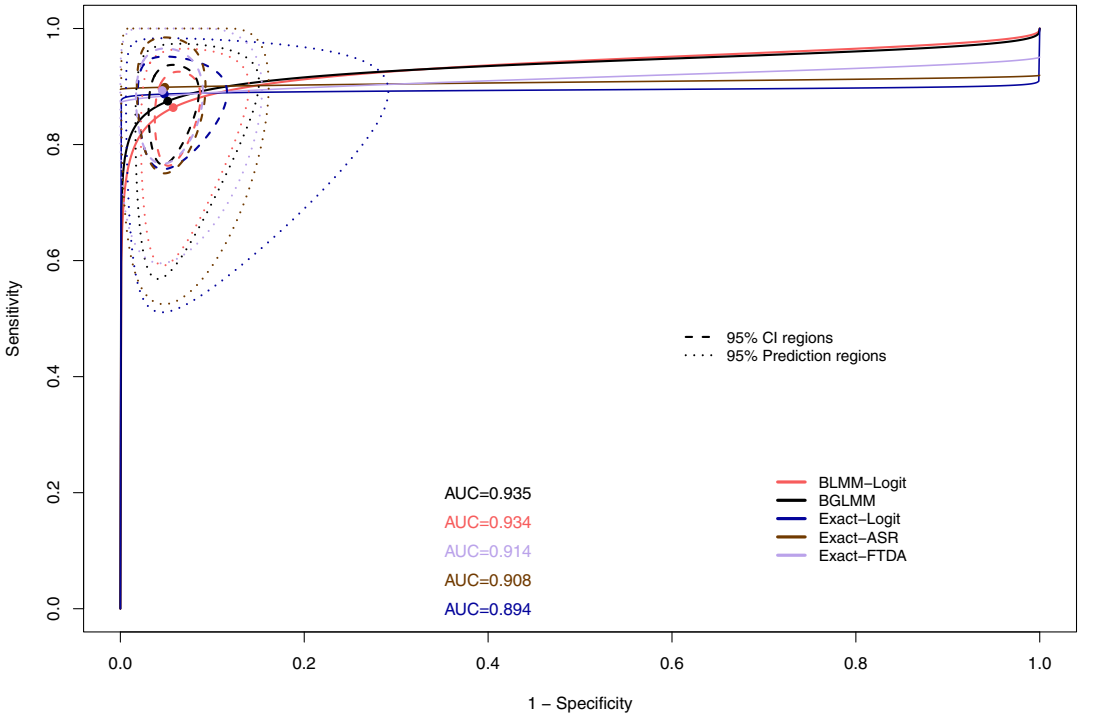


Figure 10. *SROC curves along with their AUCs and 95% confidence and prediction regions based on the BLMM, BGLMM, Exact-Logit, Exact-ASR, and Exact-FTDA methods for the US-Children dataset.*

eliminates the need for traditional, ad hoc Haldane–Anscombe correction to model parameter estimates, thereby avoiding artificial bias. We conducted an extensive simulation analysis using varying data characteristics and model parameters to evaluate the respective performances of the proposed and existing approaches. Our simulation and real-life data evaluations provided insights into the robustness and limitations of each method, particularly in the presence or absence of sparse primary studies in a meta-analysis. All methods generally showed comparable performances when the number of participants per study was large and there were no zero cell counts in primary studies of a meta-analysis. However, the BGLMM and the proposed exact methods (Exact-Logit, Exact-ASR, and Exact-FTDA) outperformed the standard BLMM-Logit approach in scenarios where meta-analyses contained sparse primary studies.

The standard approximate BLMM-Logit method had the highest bias as it underestimated *Se* and *Sp* when the number of participants per study was small or when there were primary studies in a meta-analysis with zero cell counts for TPs, TNs, FN and FP (i.e., when *Se* and *Sp* were closer to 0 or 1), meaning small sample sizes or zero cell counts could significantly impact the performance of the approximate BLMM-Logit method. This finding is consistent with prior studies by Doebler et al.,[37] Negeri et al.,[9] and Rosenberger et al.[18] Contrarily, methods based on the proposed analytical methods (Exact-Logit, Exact-ASR, and Exact-FTDA) tended to be more robust across various sample sizes and scenarios.

The methods based on arcsine transformations yielded relatively lower RMSEs for *Se* and *Sp* when there were no sparse primary studies in the meta-analysis, but larger RMSEs when such studies were included. Conversely, the proposed Exact-Logit, BGLMM, and BLMM-Logit had the smallest RMSE for *Se* and *Sp* in the presence of sparse primary studies in a meta-analysis. As sample sizes and number of studies increased, RMSE generally decreased for all methods, showing improved estimation accuracy with large sample sizes and number of studies, consistent with the results of Negeri et al.[9]

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

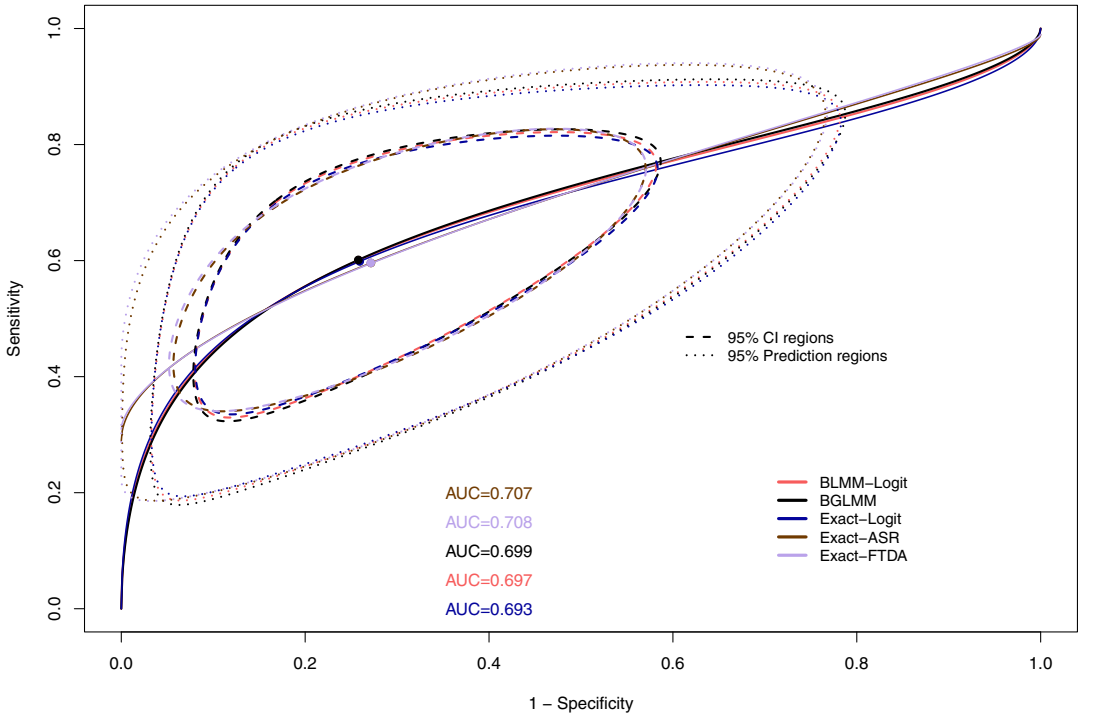


Figure 11. *SROC curves along with their AUCs and 95% confidence and prediction regions for the BLMM, BGLMM, Exact-Logit, Exact-ASR, and Exact-FTDA methods for the MMSE dataset.*

All methods exhibited similar performance in terms of coverage probabilities when there were no primary studies with zero cell counts, achieving the nominal 95% coverage values for some scenarios. However, the approximate BLMM-Logit method performed worse for meta-analyses with sparse primary studies, indicating less reliable coverage in such scenarios. Overall, the proposed analytical variance calculation approaches and the BGLMM demonstrated more consistent and robust coverage probabilities across varying scenarios, aligning with their general performance advantages observed in the bias and RMSE results. These findings also align with those of Negeri et al.[9] Relatively, however, the proposed Exact-ASR and Exact-FTDA methods sometimes had higher 95% coverage probabilities than the BGLMM for meta-analyses with small studies because the estimation of between-study heterogeneity and correlation in BGLMM could become unreliable due to convergence issues or estimated parameters being on the boundary of parameter space.[38,39] This instability can result in overly small standard errors and narrow CIs that fail to capture the true parameters as often as expected.

Furthermore, the standard BLMM-Logit method had the poorest bias when estimating the between-study variances and covariance compared to the other methods evaluated in this study. While the approximate BLMM-Logit method considerably underestimated σ_1^2 regardless of the presence of sparse primary studies and the sample sizes, the BGLMM, the Exact-Logit method, and the arcsine transformations-based exact methods (Exact-ASR and Exact-FTDA) generally offered unbiased σ_1^2 and σ_2^2 estimates. These observations further demonstrate the approximate BLMM-Logit method's limitations in variance–covariance estimation under challenging data conditions.

We further reinforced our simulation results by applying all methods to the *US-Children* and *MMSE* datasets. For the *US-Children* data, the approximate BLMM-Logit method yielded smaller estimates of summary sensitivity and specificity, consistent with the simulation trends. In contrast, all the other methods provided more robust estimates, although methods based on the arcsine transformations had

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

wider 95% CIs. As expected, all methods performed similarly in estimating the overall sensitivity, overall specificity, and the between-study variances and covariance for the *MMSE* data, which does not involve sparse primary studies with zero cell counts for TPs, TNs, FNs and FPs. This result also agreed with our simulation findings. Overall, these examples further demonstrate the flexibility of the proposed exact methods for calculating the within-study variances for ADMA of DTA studies.

Of the transformations considered in this manuscript, the logit transformation has been commonly used as a standard transformation for proportions, although it suffers from instability, especially when proportions are close to 0 or 1, which often require Haldane–Anscombe correction that may introduce bias. In contrast, the ASR and FTDA transformations stabilize variances more effectively and reduce the influence of extreme proportions.[9] Thus, while the logit transformation may be preferable when clinical interpretability and the linkage of the log-transformed diagnostic odds ratio [log(DOR)] with logit-transformed *Se* and logit-transformed *Sp* are prioritized, the arcsine-based transformations are more practically useful in situations where stability and estimation accuracy are the primary goals. Moreover, the inverse functions of the ASR and FTDA can be used to transform model estimates to the proportion scale, link the DOR with *Se* and *Sp*, and improve clinical interpretability of meta-analyses outputs.[9]

The strengths of this article include developing and validating a robust within-study variance calculation method for ADMA of DTA studies, comparing the new approach with several existing methods, demonstrating all methods using simulated and real-life meta-analytic data, and implementing the methods in open-source R software. Like the standard approaches for ADMA of DTA studies, assuming a single cut-off threshold for test positivity in each primary study and not accounting for the differences in the values of these cut-offs between studies can be considered a study limitation, which will be explored further in future work. Furthermore, although the within-study variances are accurately calculated using the proposed exact methods, statistical inference still relies on large-sample properties. Therefore, methods for small samples, such as the Kenward and Roger approach [40] could be explored in future work for meta-analyses with a small number of studies.

In conclusion, all methods performed similarly when there were no sparse primary studies with zero cell counts in TPs, TNs, FPs and FNs or when the number of diseased and non-diseased individuals in each study was large. However, the proposed exact approaches and the BGLMM outperformed the asymptotic BLMM-Logit method when there were primary studies with zero cell counts, regardless of other data characteristics or when there were no primary studies with zero cell counts and the sample sizes in each primary study were small. Therefore, researchers and practitioners should pay close attention to the characteristics of their data and choose an appropriate within-study variance calculation method before conducting an ADMA of DTA studies. However, the proposed exact methods or the BGLMM should be preferred to conduct ADMA of DTA data with sparse primary studies or for meta-analyses with small within-study sample sizes.

Acknowledgements. The authors are grateful to the two anonymous reviewers and the editors for their helpful comments, which improved the final version of the manuscript. The authors would also like to acknowledge the Faculty of Mathematics at the University of Waterloo for providing computing resources.

Author contributions. Investigation, methodology, and software: O.A.D. and Z.F.N.; Formal analysis, visualization, and writing—original draft: O.A.D.; Conceptualization, writing—review and editing, funding acquisition, and supervision: Z.F.N. Both authors approved the final submitted draft.

Competing interest statement. The authors declare that no competing interests exist.

Data availability statement. All software code and data generated or analyzed during this study is publicly available at https://doi.org/10.5281/zenodo.17990418.

Funding statement. This research was supported by funding from the Natural Sciences and Engineering Research Council of Canada (NSERC; Grant No. RGPIN-2024-04652).

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/rsm.2025.10070.

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

References

[1] Croft P, Altman DG, Deeks JJ, et al. The science of clinical practice: disease diagnosis or patient prognosis? Evidence about "what is likely to happen" should shape clinical practice. *BMC Med*. 2015;13: 20.

[2] Binney N, Hyde C, Bossuyt PM. On the origin of sensitivity and specificity. *Ann Intern Med*. 2021;174: 401–407.

[3] Borenstein M. Effect sizes for continuous data. In: Cooper H, Hedges LV, Valentine JC, eds. *The Handbook of Research Synthesis and Meta-Analysis*. 2nd ed. Russell Sage Foundation; 2009: 221–235.

[4] Zar LA, Alsharif FR, Zar A, et al. Methods of diagnostic meta-analysis: comparing the generalized linear mixed model and the split component synthesis model. *Curr Opin Epidemiol Public Health*. 2024;3: 56–64.

[5] Haidich AB. Meta-analysis in medical research. *Hippokratia*. 2010;14: 29–37.

[6] Zhao Y, Khan B, Negeri ZF. An evaluation of computational methods for aggregate data meta-analyses of diagnostic test accuracy studies. *BMC Med Res Methodol*. 2024;24: 111.

[7] 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: 982–990.

[8] Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med*. 2001;20: 2865–2884.

[9] Negeri ZF, Shaikh M, Beyene J. Bivariate random-effects meta-analysis models for diagnostic test accuracy studies using arcsine-based transformations. *Biom J*. 2018;60: 827–844.

[10] Trikalinos TA, Trow P, Schmid CH. *Simulation-Based Comparison of Methods for Meta-Analysis of Proportions and Rates: Methods Research Report 13(14)-EHC084-EF*. Tufts Medical Center Evidence-based Practice Center, Agency for Healthcare Research and Quality; 2013. Accessed 21/12/2024. https://effectivehealthcare.ahrq.gov/reports/final.cfm.

[11] 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.

[12] Li L, Lin L, Cappelleri JC, Chu H, Chen Y. ZIBGLMM: zero-inflated bivariate generalized linear mixed model for meta-analysis with double-zero-event studies. *Res Synth Methods*. 2025;16: 251–271.

[13] Haldane JBS. The estimation and significance of the logarithm of a ratio of frequencies. *Ann Hum Genet*. 1956;20: 309–311.

[14] Anscombe FJ. On estimating binomial response relations. *Biometrika*. 1956;43: 461–464.

[15] 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: 1331–1332.

[16] Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JAC. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics*. 2007;8: 239–251.

[17] Chu H, Guo H, Zhou Y. Bivariate random effects meta-analysis of diagnostic studies using generalized linear mixed models. *Med Decis Mak*. 2010;30: 499–508.

[18] Rosenberger KJ, Chu H, Lin L. Empirical comparisons of meta-analysis methods for diagnostic studies: a meta-epidemiological study. *BMJ Open*. 2022;12: e055336.

[19] Doria AS, Moineddin R, Kellenberger CJ, et al. US or CT for diagnosis of appendicitis in children and adults? A meta-analysis. *Radiology*. 2006;241: 83–94.

[20] Arevalo-Rodriguez I, Smailagic N, Roqué IF, et al. Mini-mental state examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database Syst Rev*. 2015;(3): CD010783.

[21] Cha S, Kim S, Lee E, et al. Saline-filled appendiceal ultrasonography in clinically equivocal acute appendicitis. *Abdom Imaging*. 1996;21: 525–529.

[22] Lowe LH, Penney MW, Stein SM, et al. Unenhanced limited CT of the abdomen in the diagnosis of appendicitis in children: comparison with sonography. *AJR Am J Roentgenol*. 2001;176: 31–35.

[23] Han TI. Improved sonographic visualization of the appendix with a saline enema in children with suspected appendicitis. *J Ultrasound Med*. 2002;21: 511–516.

[24] Siegel MJ, Carel C, Surratt S. Ultrasonography of acute abdominal pain in children. *JAMA*. 1991;266: 1987–1989.

[25] Rücker G, Schwarzer G, Carpenter J, Olkin I. Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Stat Med*. 2009;28: 721–738.

[26] Freeman MF, Tukey JW. Transformations related to the angular and the square root. *Ann Math Stat*. 1950;21: 607–611.

[27] Kasuya E. Angular transformation–another effect of different sample sizes. *Ecol Res*. 2004;19: 165–167.

[28] Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J Comput Graph Stat*. 1995;4: 12–35.

[29] Liu Q, Pierce DA. A note on Gauss-Hermite quadrature. *Biometrika*. 1994;81: 624–629.

[30] Jorgensen M. Iteratively reweighted least squares. In: *Encyclopedia of Environmetrics*. Wiley; 2006. https://doi.org/10.1002/9780470057339.

[31] Faraway J. Faraway: functions and datasets for Books by Julian Faraway. R package version 1.0.8; 2022. Accessed 09/01/2024. https://CRAN.R-project.org/package=faraway.

[32] Doebler P. mada: Meta-analysis of diagnostic accuracy. R package version 0.5.8; 2017. Accessed 09/06/2024. https://CRAN.R-project.org/package=mada.

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press

[33] Ripley BD. *Modern Applied Statistics with S*. 4th ed. Springer; 2002. https://doi.org/10.1007/978-0-387-21706-2.
[34] Ripley B, Venables B, Bates DM, et al. Package 'mass'. *Cran R*. 2013;538: 822.
[35] Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67: 1–48.
[36] Kriston L, Hölzel L, Weiser AK, Berner MM, Härter M. Meta-analysis: are 3 questions enough to detect unhealthy alcohol use? *Ann Intern Med*. 2008;149: 879–888.
[37] Doebler P, Holling H, Böhning D. A mixed model approach to meta-analysis of diagnostic studies with binary test outcome. *Psychol Methods*. 2012;17: 418.
[38] Takwoingi Y, Guo B, Riley RD, Deeks JJ. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. *Stat Methods Med Res*. 2015;26: 1896–1911.
[39] Guolo A. A double SIMEX approach for bivariate random-effects meta-analysis of diagnostic accuracy studies. *BMC Med Res Methodol*. 2017;17: 6.
[40] Kenward MG, Roger JH. An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Comput Stat Data Anal*. 2009;53: 2583–2595.

Cite this article: Dabi, OA, Negeri ZF. Diagnostic test accuracy meta-analysis based on an exact within-study variance calculation method. *Research Synthesis Methods*. 2026;17: 693–713. https://doi.org/10.1017/rsm.2025.10070

https://doi.org/10.1017/rsm.2025.10070 Published online by Cambridge University Press