Published in final edited form as:

Stat Methods Med Res. 2016 December ; 25(6): 3015–3037. doi:10.1177/0962280214536703.

A Hybrid Bayesian Hierarchical Model Combining Cohort and Case-control Studies for Meta-analysis of Diagnostic Tests: Accounting for Partial Verification Bias

Xiaoye Ma[1], **Yong Chen**[2], **Stephen R. Cole**[3], and **Haitao Chu**[1,*]

[1]Division of Biostatistics, School of Public Health, The University of Minnesota, Minneapolis, MN 55455

[2]Division of Biostatistics, School of Public Health, The University of Texas, Houston, TX 77030

[3]Department of Epidemiology, University of North Carolina, Chapel Hill, NC 27599

Abstract

To account for between-study heterogeneity in meta-analysis of diagnostic accuracy studies, bivariate random effects models have been recommended to jointly model the sensitivities and specificities. As study design and population vary, the definition of disease status or severity could differ across studies. Consequently, sensitivity and specificity may be correlated with disease prevalence. To account for this dependence, a trivariate random effects model had been proposed. However, the proposed approach can only include cohort studies with information estimating study-specific disease prevalence. In addition, some diagnostic accuracy studies only select a subset of samples to be verified by the reference test. It is known that ignoring unverified subjects may lead to partial verification bias in the estimation of prevalence, sensitivities and specificities in a single study. However, the impact of this bias on a meta-analysis has not been investigated. In this paper, we propose a novel hybrid Bayesian hierarchical model combining cohort and case-control studies and correcting partial verification bias at the same time. We investigate the performance of the proposed methods through a set of simulation studies. Two case studies on assessing the diagnostic accuracy of gadolinium-enhanced magnetic resonance imaging in detecting lymph node metastases and of adrenal fluorine-18 fluorodeoxyglucose positron emission tomography in characterizing adrenal masses are presented.

Keywords

Bayesian method; cohort and case-control studies; diagnostic test; partial verification bias; meta-analysis

1 Introduction

Accurate diagnosis of a disease is often the first step toward its treatment and prevention. The growing number of assessment instruments, as well as a rapid escalation in costs has

*Corresponding author: chux0051@umn.edu.

generated an increasing need for scientifically rigorous comparisons of the diagnostic tests in clinical practice. In the presence of a gold standard measure of disease status, the performance of a binary diagnostic test is often measured by paired indices, such as sensitivity (Se) and specificity (Sp), positive and negative predictive values (PPV and NPV), or positive and negative diagnostic likelihood ratios (LR+ and LR−).[1,2] Sensitivity and specificity are often regarded as intrinsic properties of a diagnostic test. However, it is well understood that Se and Sp may not reflect the clinical utility of a diagnostic test; such clinical utility depends on the prevalence of disease (π) in the population to which the instrument is applied.[3] In particular, high NPV or a low LR− is necessary for a diagnostic test to be useful at ruling out disease, and high PPV or a high LR+ is necessary for a diagnostic test to be useful at confirming disease.

Meta-analysis of diagnostic tests is a useful tool to combine evidence on diagnostic accuracies from multiple studies. Compared to conventional meta-analyses of controlled clinical trials, it has several additional statistical challenges. Specifically, the paired indices are typically correlated and heterogeneous across studies due to differences in study design, population selection, or laboratory methods.[4–13] Bivariate random effects models on sensitivities and specificities have been recommended to account for such correlation and heterogeneity in the literature and specifically by the Cochrane Diagnostic Methods group.[8,10,11] In addition, because the classification of disease status is typically based on a continuum of measurable traits, and such continuous traits not only determine disease prevalence, but also misclassification rates (subjects with true levels close to the cut-point are more likely to be misclassified), sensitivities and specificities can be correlated with study prevalences.[14] Trivariate random effects models on prevalence, sensitivities and specificities were proposed to account for such correlations.[15] However, many meta-analyses of diagnostic tests in practice contain both cohort and case-control study designs.[16] Using cohort design, a study first tests participants with the index test, next confirms disease status with the gold standard.[17] In case-control design studies, groups of patients with and without disease are identified before performing the index test.[18] Thus, case-control studies cannot be used to estimate disease prevalence and direct application of the trivariate random effects models has been restricted to a meta-analysis with cohort studies only. Under such situations, ignoring the information on prevalence to fit the bivariate random effects model[6,9–11] on Se and Sp, or excluding case-control studies to fit the trivariate random effects model[15] on prevalence can potentially lead to substantial loss of information contained in the data. For example, the former approach ignores disease prevalence information and the correlations between disease prevalence Se and Sp, which can lead to incorrect estimation of PPV and NPV.

Partial verification is a common and important potential source of bias that usually arises when the selection of samples to be verified by a reference standard test is affected by the results of a diagnostic test.[19,20] As stated in the quality assessment tool for diagnostic accuracy studies (QUADAS), partial verification bias occurs when not all of the study group receive confirmation of the diagnosis by the reference standard.[21] As an illustration, let us assume that the true Se and Sp of a diagnostic test are 0.8 and 0.9, respectively. A study with a population of 100 diseased (D+) and 200 non-diseased (D−) subjects is conducted to evaluate the diagnostic test performance. Assume 80% of the subjects with test positive

outcomes are verified, while only 20% of the subjects with test negative outcomes are verified by a reference standard. Let n_{td} denote the number of subjects with test results $T = t$ and disease status $D = d$ (t, d = 0, 1, m indicating negative, positive and missing results, respectively). Assuming no sampling variation, we will have n_{11} = 64, n_{01} = 4, n_{00} = 36, n_{10} = 16, n_{1m} = 20 and n_{0m} = 160. Now, if we only use verified samples, we overestimate Se as $\widehat{Se}=n_{11}/(n_{11}+n_{01})=0.94$ and underestimate Sp as $\widehat{Sp}=n_{00}/(n_{00}+n_{10})=0.69$. Moreover, the direction and magnitude of such bias depends on selection probabilities.[22] To avoid such bias, ideally, all subjects should be verified. However, due to some practical issues such as ethical and economic considerations, partial verification is prevalent. In a systematic review of bias and variation in meta-analysis of diagnostic accuracy studies, 15 out of 31 (48%) meta-analyses contain at least one study with partial verification.[22] Thus, it is important to adjust for partial verification bias in meta-analysis of diagnostic tests.[22,23]

Methods to adjust for verification bias in a single study are widely published. Most of the methods are built upon the missing at random (MAR) assumption, when the decision to ascertain disease status only depends on the observed index test result, T. Violations of this condition can happen when, for example, subjects with family disease history are more likely to get disease status verified.[1] Begg and Greenes[24] proposed a simple method based on Bayes theorem. Other methods such as multiple imputation, direct maximum likelihood, or Bayesian approaches have been proposed.[20,25–29] These methods give unbiased estimates of Se and Sp for individual studies instead of recovering missing counts of subjects. Thus we would not be able to apply the exact binomial likelihood assumption for a GLMM approach under meta-analysis settings. Few sensitivity analysis methods are available under the assumption of Missing Not At Random (MNAR), i.e., the probability of being verified by a reference standard depends on the unobserved data.[30,31]

On the other hand, only limited literature are available on methods to adjust verification bias in a meta-analysis setting. De Groot et al.[32] extended the Bayes theorem method to adjusting for this bias in meta-analysis of diagnostic tests with nominal outcomes. A two-stage Bayesian approach was described, where in the 1st stage the probability distribution of the index test was calculated and in the 2nd stage PPV and NPV are calculated using observed data based on their unbiasedness property under the MAR assumption.[1] Bayes theorem is then applied to achieve pooled sensitivity and specificity estimates. A few papers have discussed the missing data problem caused by imperfect reference standards, but these papers are not aimed at partial verification problems specifically. Chu et al.[33] discussed a latent class random effects model for such a scenario. The model allows variation in sensitivity, specificity and prevalence across different studies, and allows correlation among the parameters. Sadatsafavi et al.[34] proposed a random effects model which allows either sensitivity or specificity to vary across studies.

To the best of our knowledge, no one has considered methods to combine information from cohort and case-control studies, and to correct partial verification bias in meta-analyses of diagnostic tests simultaneously. In this paper we propose a hybrid generalized linear mixed model (hybrid GLMM) to solve the two problems together under the assumption of a gold standard reference test. The proposed method is described in Section 2. Simulation studies are carried out and reported in Section 3. Section 4 provides two motivating case studies.

The paper ends with a discussion in Section 5. The data sets for the two case studies are given in Appendix 1, Tables S1 and S2, and corresponding WinBUGS code are given in Appendix, respectively.

2 Bayesian Hierarchical Model

2.1 Notations

Suppose that we have a meta-analysis with N diagnostic accuracy studies, and the studies are indexed such that the N_1 cohort studies come first, followed by $N_2 = N - N_1$ case-control studies. To allow partial verification in some of the first N_1 cohort studies, let n_{itd} be the number of subjects with disease status $D = d$ and test results $T = t$ (d, $t = 0, 1, m$ indicating negative, positive and missing results, respectively) in the ith study ($i = 1, 2, \ldots, N_1$) and p_{itd} be the corresponding probability. As subjects with both D and T missing do not provide any information, we will not consider them. Let π_i, Se_i and Sp_i denote disease prevalence, sensitivity and specificity for study i such that $\pi_i = P(D = 1)$, $Se_i = P(T = 1|D = 1)$ and $Sp_i = P(T = 0|D = 0)$. Let $V = 1$ and $V = 0$ denote the subject is verified or not, respectively. Let ω_{itm} ($t = 0, 1$) and ω_{imd} ($d = 0, 1$) be the mutually exclusive probabilities of missing for subjects with test result $T = t$ and disease status $D = d$, respectively. Furthermore, given the nature of case-control studies, it is unnecessary to consider the influence of missing data in case-control studies: subjects with unverified disease status generally do not exist and subjects with missing diagnostic test outcomes can be ignored as prevalences in such studies are not well defined.

Table 1 presents the data structure and notation for the ith study when it is a cohort study or a case-control study. In each cell, the number of cell counts and the corresponding probabilities are presented. The left panel is for a cohort studies, which extends a standard 2 × 2 table to allow for partial verification. The sum of all cell probabilities is one. The right half is for a case-control studies with a typical 2 × 2 table. The cell probabilities sum up to one for diseased and non-diseased subjects respectively. Derivations of the cell probabilities for cohort studies are also provided at the footnote of Table 1.

2.2 The Likelihood with Random Effects Accounting for Heterogeneity

Let $\boldsymbol{\omega} = \{\boldsymbol{\omega}_i\}$ and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_i\}$, where $\boldsymbol{\omega}_i = (\omega_{i0m}, \omega_{i1m}, \omega_{im0}, \omega_{im1})$ and $\boldsymbol{\theta}_i = (\pi_i, Se_i, Sp_i)$ for study i. Assuming independence among subjects conditional on $\boldsymbol{\theta}_i$ and $\boldsymbol{\omega}_i$, the likelihood is the product of contribution from each study. Multinomial likelihoods are used for cohort studies and binomial likelihoods are used for case-control studies. In this paper we assume verification is MAR, where the missing probabilities $\boldsymbol{\omega}$ are independent of prevalence and test accuracy parameters, $\boldsymbol{\theta}$. Therefore, the likelihood can be factored as $L(\boldsymbol{\theta}, \boldsymbol{\omega}|\text{Data}) \propto L(\boldsymbol{\theta}|\text{Data}) \times L(\boldsymbol{\omega}|\text{Data})$. Specifically,

$$L(\boldsymbol{\omega}|\text{Data}) \propto \prod_{i=1}^{N_1} \left\{ \omega_{i1m}^{n_{i1m}} \omega_{im1}^{n_{im1}} \omega_{im0}^{n_{im0}} \omega_{i0m}^{n_{i0m}} \prod_{j,k=0,1} (1-\omega_{ijm}-\omega_{imk})^{n_{ijk}} \right\} \quad (1)$$

and

$$L(\boldsymbol{\theta}|\text{Data}) \propto \prod_{i=1}^{N} \{Se_i^{n_{i11}}(1-Sp_i)^{n_{i10}}(1-Se_i)^{n_{i01}}Sp_i^{n_{i00}}\} \prod_{i=1}^{N_1} \left\{ \pi_i^{\sum_j n_{ij1}} (1-\pi_i)^{\sum_j n_{ij0}} h_{i1}^{n_{i1m}} h_{i0}^{n_{i0m}} \right\}, \quad (2)$$

where $h_{i1} = \pi_i Se_i + (1 - \pi_i)(1 - Sp_i)$, $h_{i0} = \pi_i(1 - Se_i) + (1 - \pi_i)Sp_i$ and $j = 0, 1, m$.

To account for potential between-study heterogeneity, we consider a generalized linear mixed effects model (GLMM):

$$g(\pi_i)=\eta+\varepsilon_i; \quad g(Se_i)=\alpha+\mu_i; \quad g(Sp_i)=\beta+\nu_i, \quad (3)$$

where $g()$ is a link function such as logit or probit, and $(\varepsilon_i, \mu_i, \nu_i)^T$ is a vector of random effects. To account for potential correlation among π_i, Se_i and Sp_i, the random effects $(\varepsilon_i, \mu_i, \nu_i)^T$ are assumed to follow a multivariate normally distribution as $(\varepsilon_i, \mu_i, \nu_i)^T \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\sum = \begin{bmatrix} \sigma_\varepsilon^2 & \rho_{\varepsilon\mu}\sigma_\mu\sigma_\varepsilon & \rho_{\varepsilon\nu}\sigma_\nu\sigma_\varepsilon \\ & \sigma_\mu^2 & \rho_{\mu\nu}\sigma_\nu\sigma_\mu \\ & & \sigma_\nu^2 \end{bmatrix}.$$

The diagonal elements of the variance-covariance matrix $\boldsymbol{\Sigma}$, $(\sigma_\varepsilon^2, \sigma_\mu^2, \sigma_\nu^2)$, characterize the between-study heterogeneities of disease prevalence, test sensitivities and specificities, while the off diagonal elements $(\rho_{\varepsilon\mu}, \rho_{\varepsilon\nu}, \rho_{\mu\nu})$ capture the correlations between the corresponding random effects (π_i, Se_i), (π_i, Sp_i) and (Se_i, Sp_i) in the transformed scale, respectively. For simplicity, we assume the same correlation structure for sensitivities, specificities and prevalences for both case-control and cohort studies in this paper, which can be easily relaxed if necessary. However, for case-control studies, the study-specific prevalences are not contained in the likelihood and not directly estimable, and can be predicted using this correlation structure and study-specific sensitivity and specificity.

Study-level covariates, such as study quality, type of design (case-control versus cohort studies), race distribution and mean age, can be incorporated through meta-regression when necessary. For example, let $g(\pi_i) = \eta_0 + \boldsymbol{\eta}_1 \boldsymbol{X}_i + \varepsilon_i$, $g(Se_i) = \alpha_0 + \boldsymbol{\alpha}_1 \boldsymbol{W}_i + \mu_i$ and $g(Sp_i) = \beta_0 + \boldsymbol{\beta}_1 \boldsymbol{Z}_i + \nu_i$, where $\boldsymbol{X}_i$, $\boldsymbol{W}_i$ and $\boldsymbol{Z}_i$ denote the possibly overlapping study-level covariate vectors. Note that the hybrid GLMM accounts for different study designs in the construction of likelihood. Including type of study design as a covariate is helpful when there is a systematic difference between cohort and case-control studies, e. g., if the pooled sensitivity and specificity are believed to be different between the two designs.

The marginal likelihood integrated over random effects is:

$$L(\boldsymbol{\theta},\boldsymbol{\omega})=\iiint L(\boldsymbol{\theta},\boldsymbol{\omega}|\text{Data})\times p(\mu_i,\nu_i,\varepsilon_i|\sum)\,d\varepsilon_i\,d\mu_i\,d\nu_i \quad (4)$$

Frequentist methods (such as the maximum likelihood estimate) may converge slowly or have convergence problems due to the need to maximize the marginal likelihood with trivariate integrations, and the corresponding asymptotic approximations for standard errors of functions of parameters may not be sufficiently accurate.[35]

2.3 Bayesian Posterior Sampling approach

In this paper, we consider fully Bayesian approaches using Markov chain Monte Carlo (MCMC) methods for parameter estimation. In most instances, inferences obtained by Bayesian and classical frequentist methods are similar when the former uses non-informative or weakly informative prior distributions for all model parameters.[36] Compared to the frequentist methods, MCMC algorithms permit full posterior inference (e.g., credible intervals (CrIs)) even when the normality approximation based on large sample theory is insufficient, which is valuable here because the sampling distributions of $\boldsymbol{\pi}$, Se, Sp, PPV, NPV, LR+ and LR− are often skewed and the number of studies in the meta-analysis is typically small or moderate (e.g., $N < 30$). Specifically, we will draw posterior inference using Gibbs and Metropolis-Hastings sampling algorithms[37–40] with convergence assessed using trace plots, sample autocorrelations, and statistical convergence diagnostic tests.[41,42]

Let $p(\eta)$, $p(\alpha)$, $p(\beta)$ and $p(\boldsymbol{\Sigma})$ denote the prior distributions for η, α, β and $\boldsymbol{\Sigma}$. We take non-informative normal priors on η, α, β and a Wishart prior on the precision matrix $\boldsymbol{\Sigma}^{-1}$ (inverse Wishart prior on $\boldsymbol{\Sigma}$), denoted by

$$p(\eta)\sim N(0,10^2);\quad p(\alpha)\sim N(0,10^2);\quad p(\beta)\sim N(0,10^2);\quad p(\sum{}^{-1})\sim W(\mathbf{R},v), \quad (5)$$

where $\mathbf{R}$ is a 3 by 3 matrix, and a small number is chosen as the degrees of freedom v ($v \geq 3$). The posterior distribution of η, α, β and $\boldsymbol{\Sigma}$ can be written as:

$$p(\eta,\alpha,\beta,\sum|\text{Data})\propto L(\boldsymbol{\theta}|\text{Data})p(\eta)p(\alpha)p(\beta)p(\sum)\prod_{i=1}^{N}p(\varepsilon_i,\mu_i,\nu_i|\sum) \quad (6)$$

where $L(\boldsymbol{\theta}|data)$ depends on (η, α, β) through $\boldsymbol{\pi}_i = g^{-1}(\eta + \varepsilon_i)$, $Se_i = g^{-1}(\alpha + \mu_i)$ and $Sp_i = g^{-1}(\beta + \nu_i)$, and $g^{-1}(\cdot)$ is the inverse function of the link function $g(\cdot)$. When study-level covariates are included in the link functions, plug in $\boldsymbol{\pi}_i = g^{-1}(\eta_0 + \boldsymbol{\eta}_1\boldsymbol{X}_i + \varepsilon_i)$, $Se_i = g^{-1}(\alpha_0 + \boldsymbol{\alpha}_1\boldsymbol{W}_i + \mu_i)$ and $Sp_i = g^{-1}(\beta_0 + \boldsymbol{\beta}_1\boldsymbol{Z}_i + \nu_i)$ instead. Here we focus on the model without covariates for simplicity of the presentation.

Using the MCMC samples of η, α, and β, the posterior samples for population-averaged PPV, NPV, LR+, LR− can be approximated by the following formulas:

$$PPV=\frac{g^{-1}(\eta)g^{-1}(\alpha)}{g^{-1}(\eta)g^{-1}(\alpha)+\{1-g^{-1}(\eta)\}\{1-g^{-1}(\beta)\}}$$
$$NPV=\frac{\left\{1-g^{-1}(\eta)\right\}g^{-1}(\beta)}{\{1-g^{-1}(\eta)\}g^{-1}(\beta)+g^{-1}(\eta)\{1-g^{-1}(\alpha)\}}$$
$$LR+=g^{-1}(\alpha)/\left\{1-g^{-1}(\beta)\right\},\quad LR-=\left\{1-g^{-1}(\alpha)\right\}/g^{-1}(\beta)$$

3 Simulation

3.1 Simulation Design

We conduct 12 sets of simulations to compare the proposed Bayesian hybrid GLMM (model 1) to two alternative approaches which researchers are likely to apply in practice: 1) a complete case analysis approach in which subjects not verified are ignored (model 2); and 2) a trivariate GLMM approach in which case-control studies are excluded from the analysis (model 3). To fit model 2, case-control and cohort studies are combined as in the hybrid GLMM, while the missing counts are excluded. To fit model 3, the missing counts are accounted for as in the hybrid GLMM, while all case-control studies are excluded from the data. To investigate the performance of the proposed hybrid GLMM, for each generated dataset, we fit the hybrid GLMM, model 2 and model 3 separately using R package BRugs.[43] Each dataset contains equal numbers of case-control and cohort studies, where cohort studies are subject to partial verification. The probabilities of missing a reference test are 0.2 and 0.8, given diagnostic test results being positive and negative, respectively. The median prevalence is set to be 0.2 with the variances as $\sigma_\varepsilon^2=\sigma_\mu^2=\sigma_\nu^2=1$, and the number of subjects per study is chosen to be similar to the case studies in Section 4. Specifically, we consider 12 settings with small (10) or moderate (30) number of studies in a meta-analysis and high sensitivity (specificity) as 0.9 (0.95), or low sensitivity (specificity) as 0.7 (0.8), respectively. To evaluate the impact of the correlation structure, the correlation parameters ($\rho_{\varepsilon\mu}$,$\rho_{\varepsilon\nu}$, $\rho_{\mu\nu}$) are chosen as (0, 0, 0), (0.5, −0.5, −0.5) or (0.8, −0.8, −0.8) to correspond to no correlation, moderate or strong correlations among disease prevalence and test sensitivity and specificity (in logit scale). We assume a positive correlation between π_i and Se_i as it is likely to happen when population with higher prevalence may have more patients with clear-cut disease condition, leading to a higher sensitivity. However, a negative correlation was also observed in some studies.[14] For each setting, 2000 replicates are generated using the trivariate logit-normal random effects model. The posterior statistics (median and 95% equal tailed CrI) are summarized from 10000 posterior samples with 5000 burn-in iterations. Model performance is evaluated by comparing bias, relative efficiency (RE) and 95% equal tailed CrI coverage probability (CP) of the three models. The REs are calculated as the ratio of the variances of estimates from the hybrid model and the variances of the estimates from an alternative model. The larger RE, the more efficient the estimate from that alternative model.

3.2 Simulation Results

We summarized in Table 2 the bias, RE and CP of estimated overall Se, Sp, π, NPV and PPV for settings with 30 studies and median Se (Sp) as 0.7 (0.8). Simulation results under

other simulation settings are summarized in Appendix 2 Tables S3–S5. Under all settings, the hybrid GLMM gives nearly unbiased estimates and satisfactory CP of Se, Sp, π, PPV and NPV that are close to the nominal level of 95%.

As expected, when the partial verification is ignored as in model 2, some of the posterior estimates were considerably biased with grossly small CP. Under our simulation assumptions, specificities are under-estimated, and prevalences and sensitivities are overestimated, which agrees with the illustrative example described in the introduction. An intuitive explanation is that if we assume $\omega_{i1m} = 0$ and $\omega_{i0m} > 0$ such that partial verification would decrease n_{i10} and n_{i00} but n_{i11} and n_{i10} remain the same, leading to increased Se and decreased Sp estimates. From the simulations we also observe that the bias in π is larger when true Se (Sp) was 0.9 (0.95) (ranges from 0.13 to 0.2) than when true Se (Sp) was 0.7 (0.8) (ranges from 0.04 to 0.11), respectively. On the contrary, Sp and Se estimates are more biased when true Se (Sp) is 0.7 (0.8) (ranges from 0.04 to 0.14 and from 0.09 to 0.11 respectively) than when true Se (Sp) is 0.9 (0.95) (ranges from 0.01 to 0.04 and from 0.03 to 0.04 respectively). Because the estimates are biased, we do not calculate the RE of these estimates. Estimates of PPV and NPV from model 2 are nearly unbiased. Under the MAR assumption, we have $P(V = 1|D = 1, T = 1) = P(V = 1|T = 1)$, where $V = 1$ indicates verification of disease status, which would imply that $P(D = 1|V = 1, Y = 1) = P(D = 1|Y = 1)$.[1]

When only cohort studies are included as in model 3, the estimates are nearly unbiased and the CPs remain close to the nominal level. Specifically, for estimation of prevalence, when there is no correlation, model 3 performs as well as the hybrid GLMM because only the cohort studies have information of π. However, as the correlation becomes larger, the RE of model 3 becomes smaller indicating the hybrid GLMM is gaining efficiency. This is because information of estimating prevalence is borrowed from Se and Sp estimates from case-control studies. For estimations of Se and Sp, substantial loss of efficiency can be observed using model 3 with REs around 0.3 and 0.5. The reason is that half of the whole study set (the case-control studies) are discarded in model 3, which contains important information to estimate Se and Sp. For estimations of PPV and NPV, loss of efficiency can also be observed with REs ranging from 0.76 to 0.92, and from 0.44 to 0.69, respectively. Generally, the relative efficiencies indicate that estimates from the hybrid model are preferable. In summary, the hybrid model performs well in correcting partial verification bias and gaining efficiency by combining the information from cohort and case-control studies.

4 Case study

4.1 Meta-analysis of Gadolinium-enhanced Magnetic Resonance Imaging (MRI) in Detecting Lymph Node Metastases

We reanalyze the meta-analysis conducted by Klerkx et al.[16] using the proposed approach. Thirty-two studies were reported assessing diagnostic accuracy of gadolinium-enhanced MRI in detecting lymph node metastases, with histopathology test as the reference gold standard test. A bivariate random effects model[6] was applied by Klerkx et al.[16] Overall sensitivity and specificity were estimated as 0.72 with 95% confidence interval (CI) (0.66,

0.79) and 0.87 with 95% CI (0.82, 0.91), respectively. Data for each study is reported in the systematic review, as well as the QUADAS[21] quality assessment checklist.

The QUADAS criterion is used to classify case-control studies and studies with partial verification. The 1st QUADAS criterion is whether patients were representative of practice and six studies were reported as “No” or “Not Specified”. These studies are considered as case-control studies and the rest as cohort studies in our analysis. The 5th QUADAS criterion is whether all subjects were verified by the reference standard or not. Nine cohort studies reported as “No”. Among them, we failed to extract missing counts from two studies (study 13 and 25 in Table S1 of Appendix 1), thus are treated as having no partial verification in the analysis. The remaining seven studies are considered as having partial verification. Specific counts of n_{1m} and n_{0m} are extracted for studies 6, 11 and 20. However, four studies only indicated total numbers of patients not verified, while specific numbers of n_{1m} and n_{0m} are unclear. In practice, efforts should be made to recover missing values. Studies with missing values should be discarded to avoid bias. In the MRI study, the original papers from studies 10, 15, 16 and 22 were examined but failed to recover missing values. However, for purpose of illustration of our method, we assign all missing subjects as diagnostic test positive (n_{0m}=0) for simplicity.

4.1.1 Model Fitting via Bayesian Approach—We fit the data using the hybrid GLMM with logit link function using WinBUGS[44] to draw posterior samples. Model 2 and model 3 are also fitted for comparison. Non-informative normal priors $N(0, 10^2)$ are given to η, α and β and a Wishart prior $W(\mathbf{R}, \mathbf{v})$ is given to the precision matrix $\mathbf{\Sigma}^{-1}$ as in (5). The degrees of freedom v in the Wishart prior is set as $v = 4$, as pointed out by Tokuda et al. that when $v = k + 1$, where k is the dimension of $\mathbf{\Sigma}$, the correlation coefficient parameters in $\mathbf{\Sigma}$ will have an approximately Uniform (−1, 1) vague prior.[45,46] A scaled Wishart prior method is applied by setting $v = 4$ and $\mathbf{R}$ as a 3 by 3 identity matrix. Wishart prior is known as a conjugate prior for the precision matrix in a multivariate normal distribution. However, it is restricted in that it implies the same prior assumption on all of the variance components. The scaled Wishart prior method allows the flexibility of having separate priors on each of the precision parameter, while keeping the conjugacy property.[47] The same priors are applied to model 2 and model 3. After 100,000 burn-in samples, 1,000,000 posterior samples are collected. The median estimates and 95% CrI of interested parameters are presented in Table 3, where the estimates from hybrid GLMM are in bold.

The hybrid GLMM gives posterior median estimates of overall sensitivity as 0.76, which is 0.04 higher than the estimate reported by Klerkx et al.[16] and with a slightly narrower 95% CI, i.e., an interval of (0.70, 0.82) from the hybrid GLMM versus (0.66, 0.79) from the bivariate random effects method. The posterior median is 0.84 for the overall specificity, which is 0.03 lower than the bivariate model estimates. In addition, our approach allows the estimation of disease prevalence and possible correlations among prevalence, Se and Sp. We also presented posterior estimates of PPV, NPV, LR+ and LR− in Table 3. In this case-study, the estimates from hybrid GLMM and from model 3 are very similar as only 6 of the 32 studies are case-control studies, e.g., the median sensitivity is estimated as 0.762 in hybrid model and 0.770 in model 3. The quantile contours of posterior estimates Se versus π, Sp versus π, Se versus Sp and NPV versus PPV at quantile levels 0.25, 0.5, 0.75, 0.90 and 0.95

are presented in Figure 1 A–D, respectively. Figure 1A indicates slightly positive correlation between Se and π. Negative correlation can be observed between Sp and π and between Se and Sp in Figure 1B and 1C. This observation agrees with the posterior estimates of correlation coefficients in Table 3: posterior $\rho_{\varepsilon\mu}\rho_{\varepsilon\nu}$ and $\rho_{\mu\nu}$ has median estimates as 0.08, −0.42 and −0.47. Slightly negative correlation is shown in Figure 1D between NPV and PPV. The observed estimates of Se and Sp for each study and the posterior estimates from the hybrid GLMM and model 2 are ploted in Figure 2. The plot shows that different approaches can lead to different posterior estimates.

4.1.2 Sensitivity Analysis to Prior Distributions for Σ^{-1}—In addition to the scaled Wishart prior, an unscaled Wishart prior is commonly used in which no scale parameter is imposed on the precision matrix components. For the unscaled Wishart prior for $\mathbf{\Sigma}^{-1}$, there are several applicable selections of matrix **R**: the identity matrix,[48] or a diagonal matrix with diagonal entries chosen to be close to the diagonal elements of posterior precision matrix.[36] In the latter option, previous estimates of the precision matrix can serve as a prior for further estimations. As the scaled Wishart prior in Section 4.1.1 gives posterior variance parameter estimates close to (0.32^2, $0{,}55^2$, $0{,}91^2$), we choose the Wishart prior parameter **R** to have diagonal entries close to $(0.32^2, 0.55^2, 0.91^2)^{-1} \approx (9.8, 3.3, 1.2)$. Thus, to study whether the posterior estimates are sensitive to different prior assumptions, we fit the data via two unscaled Wishart priors: the identity matrix and a diagonal matrix with elements as (9,8, 3.3, 1.2). The fitted results are shown in Table 4 under unscaled methods. It shows that different priors have little impact on the posterior median Sp or π estimates.

To visually study the impact of different priors on posterior estimates, panel A of Figure 3 plots posterior densities of Se, Sp and π and panel B of Figure 3 plots posterior densities of PPV and NPV under different prior assumptions. Figure 3 shows that different priors have little impact on the posterior Sp or π estimates. The unscaled R = diag(9.8, 3.3, 1.2) prior gives negligibly larger Se, PPV and NPV posterior estimates than the other two priors. The small impact of prior assumption is consistent with intuition and the literature. For example, Lambert et al. pointed out that in a univariate setting that relatively large study sizes (15 or 30 in their simulation settings) would be less influenced by the prior of the scale parameter than small study size (5 in their simulation settings).[49]

4.1.3 An alternative Maximum Likelihood (MLE) approach—A referee has suggested considering a frequentist MLE approach as an alternative to obtain parameter estimates. Simulation studies comparing the Bayesian and MLE approaches are available in the literature.[33] We present here the estimates of MRI meta-analysis study via MLE approach, which was carried out by SAS NLMIXED procedure. The median estimate is 0.39 (95% CI: 0.30, 0.45) for disease prevalence, 0.77 (95% CI: 0.70, 0.83) for sensitiviy and 0.85 (95% CI: 0.80, 0.90) for specificity. The bivariate GLMM[8,10,11] ignoring partial verification was also fitted via SAS NLMIXED procedure, where sensitivity is estimated to be 0.72 (95% CI: 0.66, 0.79) and specificity is estimated to be 0.87 (95% CI: 0.82, 0.92). The estimates are close to our posterior estimates from model 1 and model 2 via the Bayesian approach (Table 3). The summary receiver operating characteristic (SROC) curves was first proposed by Moses et al.[50] to reflect the trade-off between sensitivity and

specificity caused by implicit thresholds and bigger area under curve (AUC) suggests better test performance. SROC curves using the MLE estimates from the hybrid GLMM and the bivariate GLMM approaches are plotted for comparison[6,12,51] (Figure 2). AUC are estimated to be 0.83 and 0.81 from the hybrid GLMM and the bivariate GLMM, respectively. The posterior Se and Sp estimates and AUC estimates from the hybrid GLMM and the bivariate GLMM ignoring partial verification are different, indicating that ignoring partial verification can lead to different conclusions on test accuracy. Thus, it is important to account for partial verification in a meta-analysis of diagnostic tests.

4.2 Meta-analysis of adrenal fluorine-18 fluorodeoxyglucose (FDG) positron emission tomography (PET) in Characterizing Adrenal Masses

Boland et al. conducted a systematic review and meta-analysis of 21 cohort studies about test accuracy of FDG-PET in characterizing adrenal masses.[52] The reference standard tests used in the 21 cohort studies include surgery, percutaneous biopsy and follow-up CT. FDG-PET is concluded to be highly accurate in detecting and differentiating malignant adrenal disease. The authors applied the bivarate random effects model and reported that the mean sensitivity, specificity of FDG-PET are estimated to be 0.97 (95% CI: 0.93, 0.98) and 0.91 (95% CI: 0.87, 0.94), respectively.[52] However, the authors evaluated the methodologic quality of the included studies by the QUADAS criterias and 18 out of the 21 studies were at risk of partial verification bias. Among the 18 studies with missing counts, we were able to extract the total missing counts for 8 studies from the original papers. The cell counts of each study are reported in Table S2 of Appendix A. Again, we impose a strong assumption on studies with only total missing counts available that the missing subjects were all tested negative by FDG-PET. We make this assumption here to creat a violation of the missing completely at random situation to show difference in estimates from the hybrid GLMM and from model 2. Under this assumption, sensitivity estimates will be conservative. Again, in practice, missing values should be recovered as much as possible and studies with missing values should be discarded to avoid bias.

We fit this data by the hybrid GLMM and model 2. In both models we use the same priors and number of posterior samples as in the meta-analysis of MRI data (section 4.1.1). We do not fit this example by model 3, because all the included studies in this meta-analysis are cohort studies. The estimates of interesting parameters are presented in Table 4. The hybrid GLMM estimates the overall median (95% CrI) sensitivity, specificity and prevalence as 0.94 (95% CrI: 0.91, 0.97), 0.93 (95% CrI: 0.90, 0.95) and 0.39 (95% CrI: 0.31, 0.47), respectively. The overall sensitivity, specificity and prevalence estimates from model 2 are 0.96 (95% CrI: 0.93, 0.98), 0.90 (95% CrI: 0.87, 0.94) and 0.45 (95% CrI: 0.37, 0.53), respectively. The trivariate GLMM ignoring partial verification overestimate sensitivity by 0.03, underestimate specificity by 0.03 and overestimate prevalence by 0.06. Again, this example shows that ignoring partial verification bias can give different estimates for the test accuracy parameters.

5 Discussion

In this paper we propose a hybrid Bayesian hierarchical model to combine cohort and case-control studies in meta-analysis of diagnostic tests to account for disease prevalence and to correct partial verification bias. In general, this approach improves the precision of the estimates of test accuracies and predictive values by using all available information, and can be easily applied in practice using free downloadable software R[53] and WinBUGS.[44] The WinBUGS code is provided in Appendix C.

Simulation studies are performed under a variety of settings to compare the performance of the proposed method with two practical alternative approaches of either ignoring unverified subjects or excluding case-control studies. We showed that ignoring unverified subjects can lead to substantial bias and excluding case-control studies can lead to substantial loss of efficiency. Overall the simulation results show that the hybrid approach gives nearly unbiased posterior medians under all settings considered. The coverage probabilities of posterior intervals are close to the nominal level. Thus in the presence of mixed study designs and partial verification bias in a meta-analysis, the hybrid GLMM should be preferred over the two common alternative approaches.

Two case studies are used to illustrate our method. The first case study evaluates the diagnostic accuracy of gadolinium-enhanced magnetic resonance imaging in detecting lymph node metastases. After combining the case-control and cohort studies and correcting for partial verification bias, compared to the original report, slightly higher sensitivity and lower specificity point estimates are obtained. The direction of bias on Se and Sp when ignoring the missing subjects is opposite of the simulation studies because we assume some studies have higher missing probability in MRI tested positives as $n_{i0m} = 0$. This can be intuitively explained under an extreme assumption that $\omega_{i0m} = 0$ and $\omega_{i1m} > 0$ such that partial verification would decrease n_{i11} and n_{i10} but keep n_{i10} and n_{i00} the same, leading to decreased Se and increased Sp estimates. In addition, our approach provides an overall estimate of disease prevalence, which is required for computing other clinical useful indices such as PPV and NPV. The second case study evaluates the diagnostic accuracy of FDG-PET in characterizing adrenal masses. After correcting partial verification bias, the hybrid GLMM provides lower sensitivity and prevalence estimates, and higher specificity estimates than the bivariate random effects model.

An important question is what is an approriate sample size for such meta-analysis? Our simulation settings assumed sample size of 10 and 30 studies and lead to nearly unbiased estimates. As we have taken a full Bayesian approach, this becomes an even more intriguing question as the needed sample size may depend on whether there are informative priors for some parameters to improve estimation. In practice, sample size of meta-analysis varies largely. Davey et al.[54] summarized that among 22,453 meta-analyses with at least two studies, the median number of studies is three and inter-quartile ranges from 2 to 6. As our hybrid GLMM is a random effects model, larger sample sizes may be needed.

In this article, we assume that the reference test is a gold standard. In practice, however, the reference test may be imperfect and subject to misclassification. Extensions to relax the

assumption of perfect reference test are currently under investigation. In such settings, every subjects true disease status is unknown and the imperfect tests may be correlated conditional on the latent disease status, inducing additional complexity for the estimation of test performance. Effort has been devoted in this regard. For example, Chu et al.[33] talked about adjusting for missing data with imperfect reference test. Dendukuri et al.[55] proposed a Bayesian approach to access overall sensitivity and specificity under absence of gold standard assumption, extending the hierarchical summary receiver operating characteristic method by Rutter and Gatsonis.[12] Both approaches included conditional dependence between the two tests through additional covariance terms. However, restrictions on the covariance terms have to be imposed to ensure well-defined probability models.

Another assumption to be relaxed in future research is the MAR assumption. We consider the MAR assumption to be practical because in many studies whether a subject is being tested by the reference test is merely dependent on the outcome of the diagnostic test and other observed characteristics. However, in some studies such as longitudinal studies the MNAR assumption may be more appropriate. Baker[31] discussed maximum likelihood estimates for the situation with multiple tests and Kosinski and Barnhart[30] presented a general likelihood-based regression approach, based on the conditional selection model by Little,[56] that can flexibly account for covariates and model different missing data mechanisms. Future development is needed to incorporate these approaches in meta-analysis settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the editor and two referees for their valuable comments during the improvement of this paper. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality. The authors also thank Dr. Bradley P. Carlin for his helpful comments on an earlier version of this paper.

Funding Xiaoye Ma and Haitao Chu were supported in part by the US NIAID AI103012, NCI P01CA142538, NCI P30CA077598, and U54-MD008620. Yong Chen was supported in part by grant number R03HS022900 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.

References

1. Pepe, MS. The statistical evaluation of medical tests for classification and prediction. Oxford University Press; 2003.
2. Zhou, XH.; Obuchowski, NA.; McClish, DK. Wiley Series in Probability and Statistics. 2011. Statistical methods in diagnostic medicine.
3. Li J, Fine JP, Safdar N. Prevalence-dependent diagnostic accuracy measures. Statistics in Medicine. 2007; 26(17):3258–3273. [PubMed: 17212380]
4. Song F, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. International Journal of Epidemiology. 2002; 31(1):88–95. [PubMed: 11914301]

5. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. Journal of Clinical Epidemiology. 2004; 57(9):925–932. [PubMed: 15504635]
6. 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. Journal of Clinical Epidemiology. 2005; 58(10):982–990. [PubMed: 16168343]
7. Mallett S, Deeks JJ, Halligan S, Hopewell S, Cornelius V, Altman DG. Systematic reviews of diagnostic tests in cancer: review of methods and reporting. British Medical Journal. 2006; 333(7565):413. [PubMed: 16849365]
8. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007; 8(2):239–251. [PubMed: 16698768]
9. Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. Statistics in Medicine. 2008; 27(5):687–697. [PubMed: 17611957]
10. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. Journal of clinical epidemiology. 2006; 59(12):1331–1332. [PubMed: 17098577]
11. Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Statistics in Medicine. 2002; 21(4):589–624. [PubMed: 11836738]
12. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Statistics in Medicine. 2001; 20(19):2865–2884. [PubMed: 11568945]
13. Ma X, Nie L, Cole SR, Chu H. Statistical methods for multivariate meta-analysis of diagnostic tests: An overview and tutorial. Statistical Methods in Medical Research. 2013 In press.
14. Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. Journal of Clinical Epidemiology. 2009; 62(1):5–12. [PubMed: 18778913]
15. Chu H, Nie L, Cole SR, Poole C. Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: Alternative parameterizations and model selection. Statistics in medicine. 2009; 28(18):2384–2399. [PubMed: 19499551]
16. Klerkx WM, Bax L, Veldhuis WB, Heintz APM, Mali WP, Peeters PH, et al. Detection of lymph node metastases by gadolinium-enhanced magnetic resonance imaging: systematic review and meta-analysis. Journal of the National Cancer Institute. 2010; 102(4):244–253. [PubMed: 20124189]
17. Liebeschuetz S, Bamber S, Ewer K, Deeks J, Pathan AA, Lalvani A. Diagnosis of tuberculosis in South African children with a T cell-based assay: a prospective cohort study. The Lancet. 2004; 364(9452):2196–2203.
18. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. Clinical Chemistry. 2005; 51(8):1335–1341. [PubMed: 15961549]
19. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. The New England Journal of Medicine. 1978; 299(17):926–930. [PubMed: 692598]
20. de Groot JA, Bossuyt PM, Reitsma JB, Rutjes AW, Dendukuri N, Janssen KJ, et al. Verification problems in diagnostic accuracy studies: consequences and solutions. BMJ. 2011; 343:d4770. [PubMed: 21810869]
21. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Medical Research Methodology. 2003; 3(1):25. [PubMed: 14606960]
22. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. Canadian Medical Association Journal. 2006; 174(4):469–476. [PubMed: 16477057]
23. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. Journal of the American Medical Association. 1999; 282(11):1061–1066. [PubMed: 10493205]

24. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics. 1983; 39(1):207–215. [PubMed: 6871349]
25. Zhou XH. Correcting for verification bias in studies of a diagnostic test's accuracy. Statistical Methods in Medical Research. 1998; 7(4):337–353. [PubMed: 9871951]
26. Roldán Nofuentes J, Luna del Castillo J. Comparing the likelihood ratios of two binary diagnostic tests in the presence of partial verification. Biometrical Journal. 2005; 47(4):442–457. [PubMed: 16161803]
27. Harel O, Zhou XH. Multiple imputation for correcting verification bias. Statistics in medicine. 2006; 25(22):3769–3786. [PubMed: 16435337]
28. De Groot J, Janssen K, Zwinderman A, Moons K, Reitsma J. Multiple imputation to correct for partial verification bias revisited. Statistics in medicine. 2008; 27(28):5880–5889. [PubMed: 18752256]
29. Donders ART, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. Journal of clinical epidemiology. 2006; 59(10):1087–1091. [PubMed: 16980149]
30. Kosinski AS, Barnhart HX. Accounting for nonignorable verification bias in assessment of diagnostic tests. Biometrics. 2003; 59(1):163–171. [PubMed: 12762453]
31. Baker SG. Evaluating multiple diagnostic tests with partial verification. Biometrics. 1995; 51(1): 330–337. [PubMed: 7539300]
32. de Groot JA, Dendukuri N, Janssen KJ, Reitsma JB, Brophy J, Joseph L, et al. Adjusting for partial verification or workup bias in meta-analyses of diagnostic accuracy studies. American journal of epidemiology. 2012; 175(8):847–853. [PubMed: 22422923]
33. Chu H, Chen S, Louis TA. Random effects models in a meta-analysis of the accuracy of two diagnostic tests without a gold standard. Journal of the American Statistical Association. 2009; 104(486):512–523. [PubMed: 19562044]
34. Sadatsafavi M, Shahidi N, Marra F, FitzGerald MJ, Elwood KR, Guo N, et al. A statistical method was used for the meta-analysis of tests for latent TB in the absence of a gold standard, combining random-effect and latent-class methods to estimate test accuracy. Journal of Clinical Epidemiology. 2010; 63(3):257–269. [PubMed: 19692208]
35. Hamza TH, Reitsma JB, Stijnen T. Meta-analysis of diagnostic studies: a comparison of random intercept, normal-normal, and binomial-normal bivariate summary ROC approaches. Medical Decision Making. 2008; 28(5):639–649. [PubMed: 18753684]
36. Carlin, BP.; Louis, TA. Bayesian methods for data analysis. CRC Press; 2011.
37. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970; 57(1):97–109.
38. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. The Journal of Chemical Physics. 1953; 21:1087.
39. Gelfand AE, Smith AF. Sampling-based approaches to calculating marginal densities. Journal of the American statistical association. 1990; 85(410):398–409.
40. Gilks WR, Best N, Tan K. Adaptive rejection Metropolis sampling within Gibbs sampling. Applied Statistics. 1995; 44:455–472.
41. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statistical science. 1992; 7:457–472.
42. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. Journal of computational and graphical statistics. 1998; 7(4):434–455.
43. Thomas A, OHara B, Ligges U, Sturtz S. Making BUGS open. R news. 2006; 6(1):12–17.
44. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing. 2000; 10(4):325–337.
45. Barnard J, McCulloch R, Meng XL. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. Statistica Sinica. 2000; 10(4):1281–1312.
46. Tokuda, T.; Goodrich, B.; Van Mechelen, I.; Gelman, A.; Tuerlinckx, F. Technical report. University of Leuwen, Belgium and Columbia University; USA: 2011. Visualizing distributions of

covariance matrices. URL http://www.stat.columbia.edu/~gel-man/research/unpublished/Visualization.pdf

47. Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. Bayesian data analysis. CRC press; 2003.
48. Browne W, Draper D. Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. Computational statistics. 2000; 15:391–420.
49. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Statistics in Medicine. 2005; 24(15):2401–2428. [PubMed: 16015676]
50. 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. Statistics in Medicine. 1993; 12(14):1293–1316. [PubMed: 8210827]
51. Chu H, Guo H, Zhou Y. Bivariate random effects meta-analysis of diagnostic studies using generalized linear mixed models. Medical Decision Making. 2010; 30(4):499–508. [PubMed: 19959794]
52. Boland GW, Dwamena BA, Jagtiani Sangwaiya M, Goehler AG, Blake MA, Hahn PF, et al. Characterization of adrenal masses by using FDG PET: a systematic review and meta-analysis of diagnostic test performance. Radiology. 2011; 259(1):117–126. [PubMed: 21330566]
53. Ihaka R, Gentleman R. R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics. 1996; 5(3):299–314.
54. Davey J, Turner RM, Clarke MJ, Higgins JP. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. BMC medical research methodology. 2011; 11(1):160. [PubMed: 22114982]
55. 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. [PubMed: 22568612]
56. Little RJ. Pattern-mixture models for multivariate incomplete data. Journal of the American Statistical Association. 1993; 88(421):125–134.

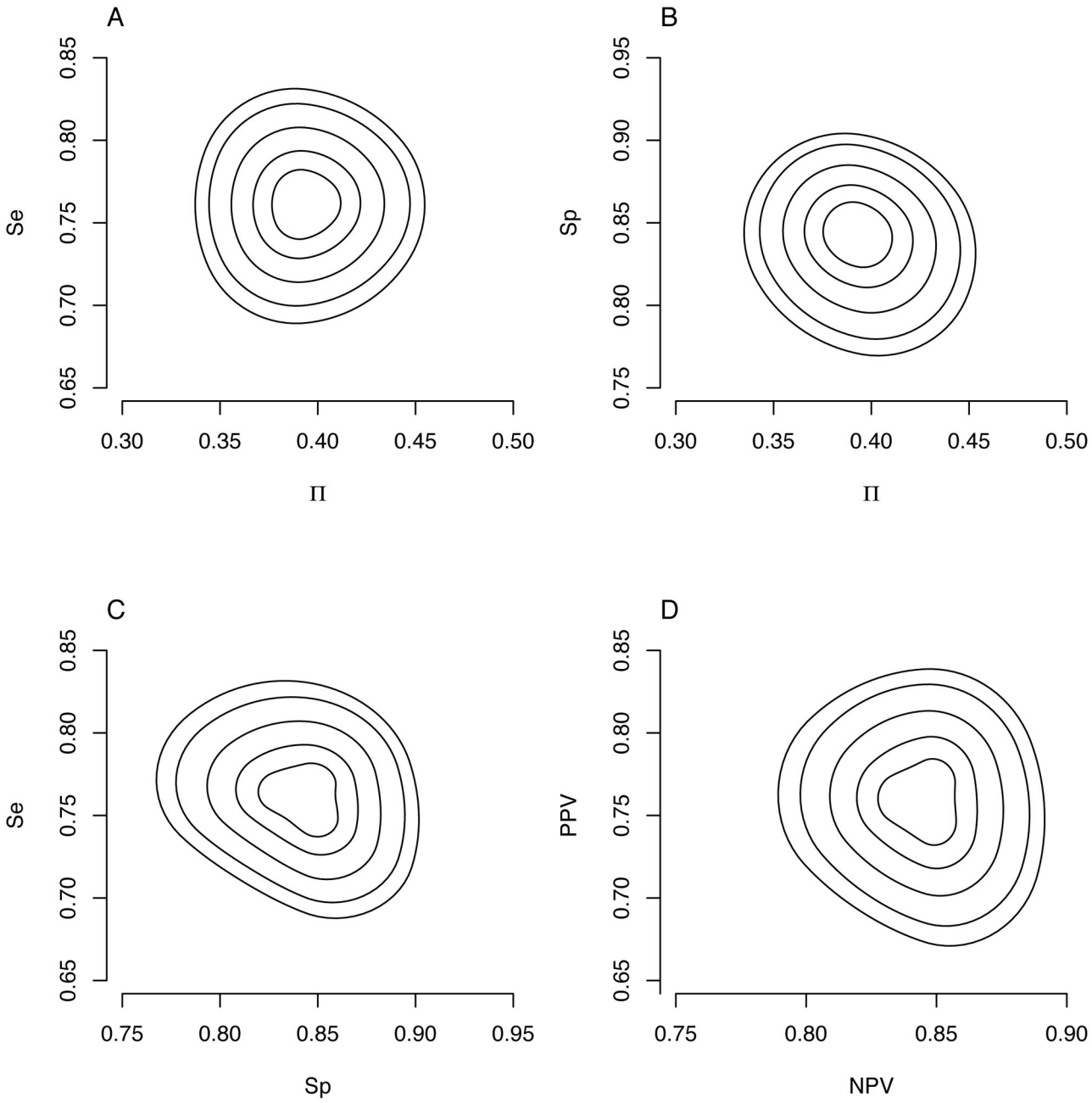

Figure 1.

Quantile contours of posterior densities from estimates of the meta-analysis of gadolinium-enhanced MRI in detecting lymph node metastases assuming scaled Wishart prior. A–D plot posterior Se versus prevalence (π), Sp versus π, Se versus Sp and PPV versus NPV, respectively, at quantile levels 0.25, 0.5, 0.75, 0.9 and 0.95.

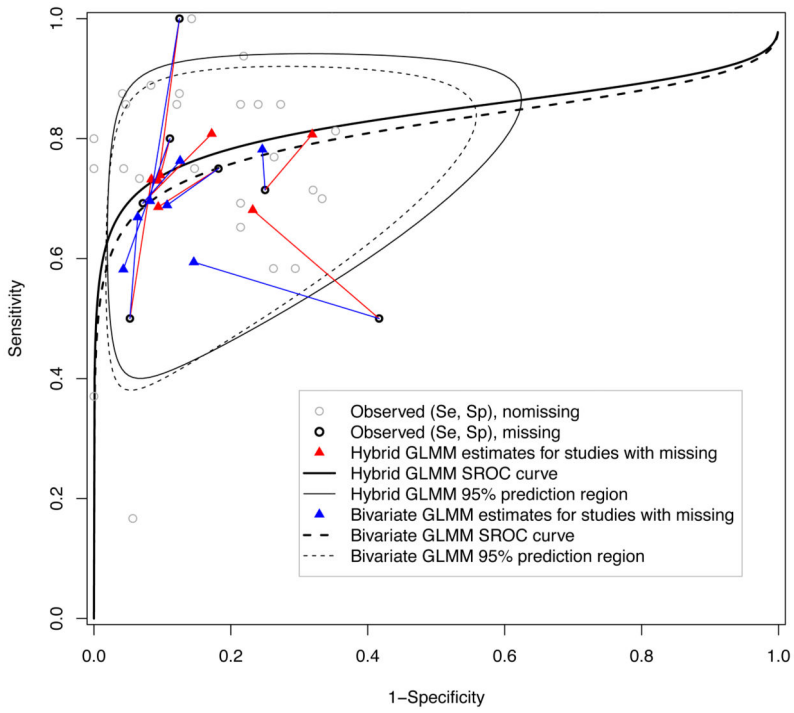

Figure 2.

SROC curves from the Hybrid GLMM and the bivariate GLMM using MLE approach. Solid lines are the SROC curve from the hybrid GLMM estimates and the 95% prediction region for the summary point estimates of Se and Sp. Dashed lines are the SROC curve from the bivaraite estimates and the 95% prediction region for the summary point estimates of Se and Sp. Black and gray circles are the observed Se and Sp from studies with and without missing counts, respectively. Red and blue triangles are the posterior estimates of Se and Sp from the Hybrid GLMM and the Bivariate GLMM ignoring partial verification, respectively.

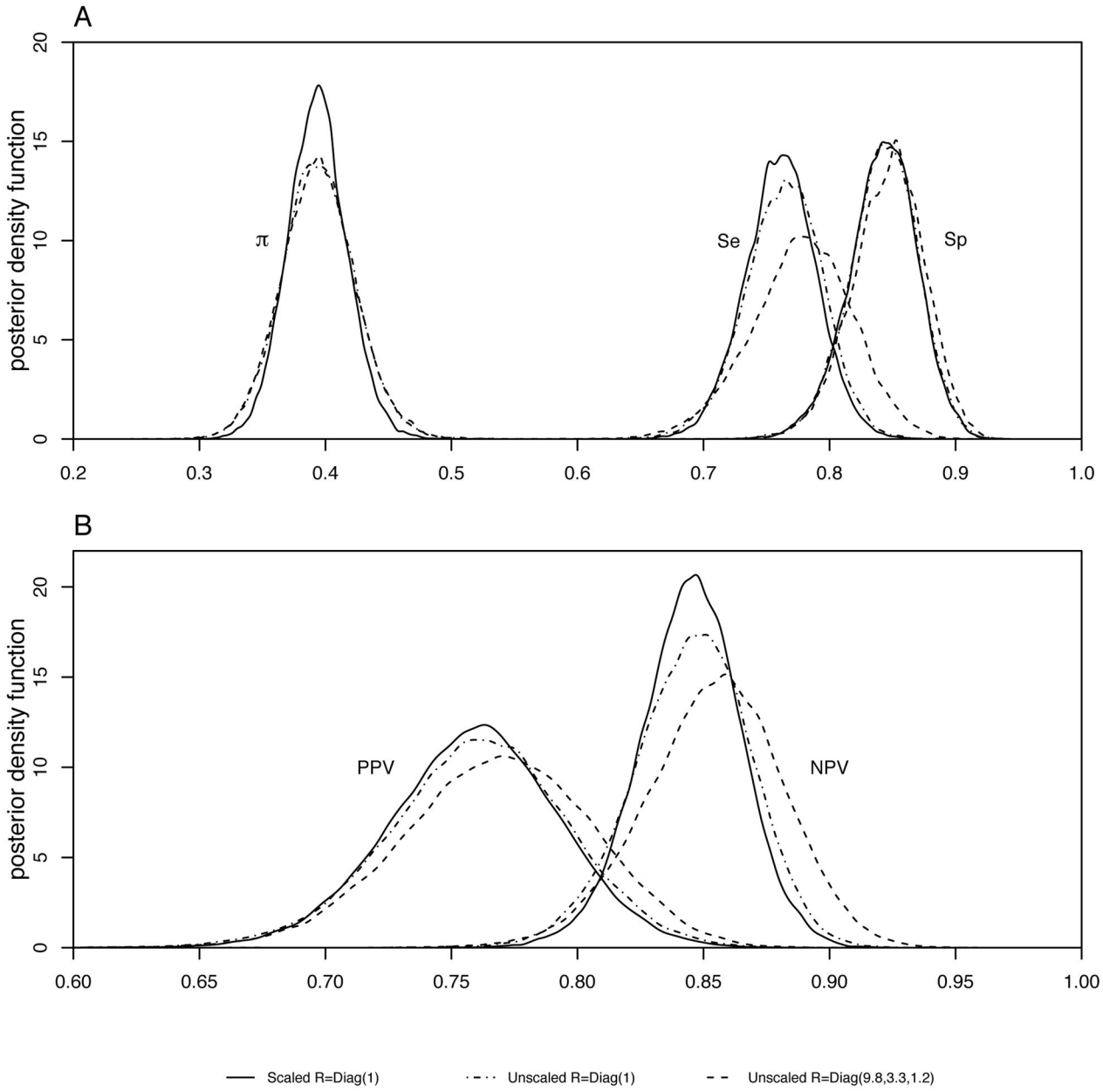

Figure 3.

Density plots of posterior estimates of the meta-analysis of gadoliniumenhanced MRI in detecting lymph node metastases under different prior assumptions. Panel A plots posterior densities of Se, Sp and prevalence (π). Panel B plots posteriors densities of PPV and NPV.

Table 1

Data display for the ith study when it is a cohort study and when it is a case-control study. In each cell, the number of cell counts are presented and the probabilities corresponding to the cell counts are presented below the cell counts.

	Gold standard				
	Cohort ($i = 1, \ldots, N_1$)			**Case-control ($i = N_1 + 1, \ldots, N$)**	
Index test	**+**	**−**	**Missing**	**+**	**−**
+	n_{i11} $(1 - \omega_{i1m} - \omega_{im1})\pi_i Se_i$	n_{i10} $(1 - \omega_{i1m} - \omega_{im0})(1 - \pi_i)(1 - Sp_i)$	n_{i1m} $\omega_{i1m}\{\pi_i Se_i + (1 - \pi_i)(1 - Sp_i)\}$	n_{i11} Se_i	n_{i10} $1 - Sp_i$
−	n_{i01} $(1 - \omega_{i0m} - \omega_{im1})\pi_i(1 - Se_i)$	n_{i00} $(1 - \omega_{i0m} - \omega_{im0})(1 - \pi_i)Sp_i$	n_{i0m} $\omega_{i0m}\{\pi_i(1 - Se_i) + (1 - \pi_i)Sp_i\}$	n_{i01} $1 - Se_i$	n_{i00} Sp_i
Missing	n_{im1} $\omega_{im1}\pi_i$	n_{im0} $\omega_{im0}(1 - \pi_i)$			

Probabilities for subjects with $V = 1$ in the ith cohort study:

$p_{i11} = P(V = 1|D = 1, T = 1)P(D = 1)P(T = 1|D = 1) = (1 - \omega_{i1m} - \omega_{im1})\pi_i Se_i$, $p_{i10} = P(V = 1|D = 0, T = 1)P(D = 0)P(T = 1|D = 0) = (1 - \omega_{i1m} - \omega_{im1})(1 - \pi_i)(1 - Sp_i)$, $p_{i01} = P(V = 1|D = 1, T = 0)P(D = 1)P(T = 0|D = 1) = (1 - \omega_{i0m} - \omega_{im1})\pi_i(1 - Se_i)$, $p_{i00} = P(T = 0|D = 0)P(V = 1|D = 0, T = 0)P(D = 0) = (1 - \omega_{i0m} - \omega_{im0})(1 - \pi_i)Sp_i$.

Probabilities for subjects with $V = 0$ in the ith cohort study:

$p_{i1m} = P(V = 0|T = 1)P(T = 1) = P(V = 0|T = 1)\{P(T = 1, D = 1) + P(T = 1, D = 0)\} = P(V = 0|T = 1)\{P(D = 1)P(T = 1|D = 1) + P(D = 0)P(T = 1|D = 0)\} = \omega_{i1m}\{\pi_i Se_i + (1 - \pi_i)(1 - Sp_i)\}$, $p_{i0m} = P(V = 0|T = 0)P(T = 0) = P(V = 0|T = 0)\{P(T = 0, D = 1) + P(T = 0, D = 0)\} = P(V = 0|T = 0)\{P(D = 1)P(T = 0|D = 1) + P(D = 0)P(T = 0|D = 0)\} = \omega_{i0m}\{\pi_i(1 - Se_i) + (1 - \pi_i)Sp_i\}$, $p_{im1} = P(V = 0|D = 1)P(D = 1) = \omega_{im1}\pi_i$, and $p_{im0} = P(V = 0|D = 0)P(D = 0) = \omega_{im0}(1 - \pi_i)$.

Table 2

Summary of 2000 simulations with data generated from settings with 30 studies, true Se (Sp)=0.7 (0.8) and different correlation assumptions. Three models are fitted: model 1 stands for the hybrid GLMM, model 2 stands for a complete-case analysis where case-control and cohort studies are combined while partial verification are ignored and model 3 stands for a trivariate GLMM where partial verificaiton bias is adjusted while case-control studies are excluded. Bias, relative efficiency and 95% CP are collected for Se, Sp, π, NPV and PPV estimates from the three models.

		Sp			Se			π			PPV			NPV		
Corr[a]	Model[b]	Bias	RE	CP	Bias	RE	CP	Bias	RE	CP	Bias	RE	CP	Bias	RE	CP
0	1	0	1	0.93	0	1	0.94	0.01	1	0.92	0.01	1	0.93	−0.01	1	0.93
0	2	−0.13	NA	0.29	0.1	NA	0.42	0.09	NA	0.71	0.02	NA	0.92	−0.02	NA	0.88
0	3	−0.01	0.47	0.93	0	0.29	0.93	0.01	1.07	0.93	0	0.83	0.94	−0.01	0.62	0.93
0.5	1	0	1	0.94	0	1	0.95	0.01	1	0.93	0	1	0.94	0	1	0.94
0.5	2	−0.13	NA	0.26	0.11	NA	0.34	0.06	NA	0.84	−0.01	NA	0.93	−0.01	NA	0.94
0.5	3	−0.01	0.5	0.94	0.02	0.32	0.94	0.01	0.95	0.93	0	0.89	0.95	0	0.68	0.95
0.8	1	0	1	0.93	0	1	0.94	0	1	0.95	0	1	0.95	0	1	0.96
0.8	2	−0.13	NA	0.29	0.11	NA	0.3	0.04	NA	0.88	−0.03	NA	0.94	0.01	NA	0.96
0.8	3	0	0.48	0.93	0.03	0.33	0.94	0	0.75	0.94	0.01	0.89	0.97	0.01	0.56	0.96

[a] Corr = 0: $(\rho_{e\mu}, \rho_{e\nu}, \rho_{\mu\nu}) = (0, 0, 0)$, Corr = 0.5: $(\rho_{e\mu}, \rho_{e\nu}, \rho_{\mu\nu}) = (0.5, -0.5, -0.5)$, Corr = 0.8: $(\rho_{e\mu}, \rho_{e\nu}, \rho_{\mu\nu}) = (0.8, -0.8, -0.8)$.

[b] Model = 1: Hybrid GLMM, Model = 2: Model2, Model = 3: Model3.

Table 3

Median estimates and 95% CrI for meta-analysis of MRI: comparing two prior families (the scaled and unscaled inverse Wishart prior) and comparing different choices of the Wishart prior parameter ***R***. Model 2 stands for a complete-case analysis where case-control and cohort studies are combined while partial verification are ignored and model 3 stands for a trivariate GLMM where partial verificaiton bias is adjusted while case-control studies are excluded.

	Scaled Method			Unscaled Method	
Parameters	**Hybrid GLMM R=I**	**Model 2 R=I**	**Model 3 R=I**	**Hybrid GLMM R=I**	**Hybrid GLMM diag(R)=(9.8,3.3,2.2)**
π	**0.39 (0.35,0.44)**	0.37 (0.32,0.42)	0.39 (0.35,0.44)	0.39 (0.34, 0.45)	0.39 (0.34,0.45)
σ_e	**0.32 (0.08,0.57)**	0.45 (0.22,0.73)	0.32 (0.10,0.57)	0.46 (0.31,0.69)	0.47 (0.33,0.69)
Se	**0.76 (0.70,0.82)**	0.73 (0.66,0.78)	0.77 (0.71,0.83)	0.77 (0.70,0.82)	0.78 (0.69,0.85)
σ_μ	**0.55 (0.21,0.99)**	0.47 (0.17,0.92)	0.47 (0.10,0.99)	0.64 (0.41,1.00)	1.03 (0.76,1.45)
Sp	**0.84 (0.79,0.89)**	0.87 (0.82,0.91)	0.85 (0.79,0.90)	0.84 (0.79,0.89)	0.85 (0.79,0.90)
σ_v	**0.92 (0.62,1.33)**	0.89 (0.62,1.31)	0.74 (0.41,1.22)	0.88 (0.61,1.27)	0.97 (0.69,1.37)
$\rho_{\mu v}$	**−0.47 (−0.92,0.15)**	−0.56 (−0.96,0.11)	−0.60 (−0.97,0.31)	−0.39 (−0.76,0.17)	−0.49 (−0.83,0.12)
$\rho_{e\mu}$	**0.08 (−0.74,0.85)**	0.37 (−0.50,0.94)	0.16 (−0.71,0.89)	−0.01 (−0.55,0.56)	0.04 (−0.56,0.61)
ρ_{ev}	**−0.42 (−0.92,0.40)**	−0.57 (−0.93,0.09)	−0.41 (−0.91,0.36)	−0.30 (−0.72,0.31)	−0.34 (−0.76,0.31)
NPV	**0.85 (0.80,0.88)**	0.85 (0.80,0.88)	0.85 (0.81,0.89)	0.85 (0.80,0.89)	0.86 (0.80,0.91)
PPV	**0.76 (0.69,0.83)**	0.76 (0.69,0.82)	0.76 (0.69,0.83)	0.76 (0.69,0.83)	0.77 (0.69,0.84)
LR+	**3.22 (2.37,4.52)**	2.65 (1.98,3.61)	3.34 (2.45,4.82)	3.25 (2.35,4.63)	3.53 (2.25,5.76)
LR−	**0.31 (0.22,0.42)**	0.38 (0.28,0.51)	0.30 (0.21,0.41)	0.31 (0.22,0.43)	0.28 (0.17,0.44)

Table 4

Median estimates and 95% CrI for meta-analysis of FDG PET: comparing the hybrid GLMM and model 2 where partial verification is ignored

Parameter	Hybrid GLMM	Model 2
π	0.39 (0.31, 0.47)	0.45 (0.37, 0.53)
σ_e	0.68 (0.47, 1.01)	0.63 (0.43, 0.95)
Se	0.94 (0.91, 0.97)	0.96 (0.93, 0.98)
σ_μ	0.68 (0.23, 1.51)	0.71 (0.23, 1.54)
Sp	0.93 (0.90, 0.95)	0.90 (0.87, 0.94)
σ_ν	0.54 (0.22, 1.08)	0.51 (0.22, 1)
$\rho_{\mu\nu}$	0.78 (−0.37, 0.97)	0.80 (−0.28, 0.97)
$\rho_{e\mu}$	−0.07 (−0.76, 0.74)	−0.05 (−0.80, 0.73)
$\rho_{e\nu}$	−0.46 (−0.89, 0.37)	−0.31 (−0.85, 0.49)
NPV	0.96 (0.93, 0.98)	096 (0.93, 0.98)
PPV	0.89 (0.84, 0.93)	0.89 (0.84, 0.93)
LR+	16.83 (9.94, 37.97)	21.77 (12.85, 49.75)
LR−	0.06 (0.03, 0.10)	0.05 (0.02, 0.08)