A Bayesian hierarchical model for network meta-analysis of multiple diagnostic tests

XIAOYE MA^{*,†}, QINSHU LIAN[†], HAITAO CHU^{*}

*Division of Biostatistics, School of Public Health, University of Minnesota,
420 Delaware St, Minneapolis, MN 55455, USA
xiaoye1043@gmail.com or chux0051@umn.edu*

JOSEPH G. IBRAHIM

*Department of Biostatistic, Gillings School of Global Public Health, University of North Carolina at
Chapel Hill, 135 Dauer Dr, Chapel Hill, NC 27599, USA*

YONG CHEN

*Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of
Pennsylvania 210 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021, USA*

SUMMARY

To compare the accuracy of multiple diagnostic tests in a single study, three designs are commonly used (i) the multiple test comparison design; (ii) the randomized design, and (iii) the non-comparative design. Existing meta-analysis methods of diagnostic tests (MA-DT) have been focused on evaluating the performance of a single test by comparing it with a reference test. The increasing number of available diagnostic instruments for a disease condition and the different study designs being used have generated the need to develop efficient and flexible meta-analysis framework to combine all designs for simultaneous inference. In this article, we develop a missing data framework and a Bayesian hierarchical model for network MA-DT (NMA-DT) and offer important promises over traditional MA-DT: (i) It combines studies using all three designs; (ii) It pools both studies with or without a gold standard; (iii) it combines studies with different sets of candidate tests; and (iv) it accounts for heterogeneity across studies and complex correlation structure among multiple tests. We illustrate our method through a case study: network meta-analysis of deep vein thrombosis tests.

Keywords: Diagnostic test; Hierarchical model; Missing data; Multiple test comparison; Network meta-analysis.

1. INTRODUCTION

Comparative effectiveness research relies fundamentally on accurate assessment of clinical outcomes. The growing number of assessment instruments, as well as the rapid escalation in the cost, have generated the increasing need for scientifically rigorous comparisons of multiple diagnostic tests in clinical practice.

^{*}To whom correspondence should be addressed.

[†]Xiaoye Ma and Qinshu Lian contributed equally to this work.

To compare the accuracy of multiple diagnostic tests in a single study, three designs are commonly used (Takwoingi, Leeflang and Deeks, 2013): (i) The multiple test comparison design where all subjects are diagnosed by all candidate tests and verified by a gold standard; (ii) The randomized design where subjects are randomly assigned to one of candidate tests, and all subjects are verified by a gold standard; and (iii) The non-comparative design where different sets of subjects are used to compare a candidate test to a gold standard or to another candidate test. Systematic reviews and meta-analysis methods have been developed as useful tools to improve the estimation of diagnostic test accuracy by combining information from multiple studies (Rutter and Gatsonis, 2001; Reitsma and others, 2005). Thus, a flexible meta-analysis framework is needed to combine information from all three designs for effectively ranking all candidate tests.

However, in the methodology literature of meta-analysis of diagnostic tests, a great deal of attention has been devoted to developing methods to estimate the performance of one candidate test compared to a reference test. When the reference test is a gold standard, multivariate random effects models are developed to account for the heterogeneity of test performance across studies and correlations among test accuracy indices (such as sensitivity and specificity) (Rutter and Gatsonis, 2001; Reitsma and others, 2005; Chu and Cole, 2006; Harbord and others, 2007; Chu, Chen and Louis, 2009; Ma and others, 2016b; Ma and others, 2016a; Chen and others, 2015). When the reference test cannot perfectly distinguish diseased and non-diseased subjects (i.e., non-gold standard), latent class random effects models (Chu, Chen and Louis, 2009; Dendukuri and others, 2012; Liu, Chen and Chu, 2015) are proposed to estimate diagnostic accuracy of both candidate and reference tests.

Very few papers have discussed how to simultaneously compare multiple candidate tests in meta-analysis. A naive procedure is to conduct separate meta-analysis methods of diagnostic tests (MA-DT) of each candidate test then compare their summary estimates, which is valid only under the missing completely random (MCAR) assumption. However, there are some important drawbacks of this procedure. First, for studies that compared multiple diagnostic tests, the accuracy estimates of each candidate test from separate MA-DT are typically correlated, as multiple test comparison design may be used for some studies and some subjects may be evaluated by multiple tests. Ignoring such correlations can lead to efficiency loss. Secondly, current methods are not able to combine studies comparing a candidate test with a gold standard and studies comparing a candidate test with a non-gold standard reference; Thirdly, when candidate tests are evaluated one at a time, the number of studies is typically small, which can potentially lead to issues of model fitting (Hamza and others, 2008) and difficulty in estimating between-study heterogeneity. In addition, as different studies that represent heterogeneous populations are synthesized, the candidate tests are not directly comparable without certain strong assumptions, thus limiting the generalizability of results. At last, separate MA-DT does not allow for “borrowing of information,” which can potentially lead to statistical efficiency loss.

To address these limitations, we develop a network MA-DT (NMA-DT) framework from the perspective of missing data analysis to simultaneously compare multiple tests. The proposed framework is motivated from the literature on network meta-analysis of randomized clinical trials, which extends the scope of traditional pairwise meta-analysis by synthesizing both direct and indirect comparisons of multiple treatments across randomized controlled trials (Lu and Ades, 2004; Salanti and others, 2011; Zhang and others, 2014). Specifically, we view studies using the randomized design and non-comparative design as if they were designed using the multiple test comparison design such that all subjects in all studies were evaluated by all candidate tests and a gold-standard test. However, most of the studies only include a subset of the whole set of tests of interest. The test outcomes from non-included tests are considered as missing data. By simultaneously comparing all candidate tests and a gold standard, the proposed approach can make use of all available information, allow for borrowing of information across studies and rank diagnostic tests through full posterior inferences. It effectively handles three critical challenges in the traditional MA-DT by (i) combining information from studies with all three designs; (ii) pooling both studies

with or without a gold standard; and (iii) allowing different sets of candidate tests in different studies or different subsets of subjects within a study. This model also accounts for potential heterogeneity across studies (due to difference of study population, design and lab technical issues) as in conventional MA-DT models, as well as the complex correlation structure among multiple diagnostic tests.

The rest of this article is organized as follows. In Section 2, we describe our motivating case study: NMA of deep vein thrombosis (DVT) tests. We present the proposed NMA-DT model and the Bayesian inference method in Section 3 and apply the proposed method to the motivating study in Section 4. Simulation studies are conducted in Section 5, and Section 6 provides a brief discussion. A *directed graphical model* of the proposed model, an additional case study, data and some additional results are provided in the supplementary material available at *Biostatistics* online.

2. MOTIVATING STUDY: NMA OF DVT TESTS

DVT is developed when blood clots form in one or more deep veins of the human body. If DVT is left untreated, the blood clot can cause pulmonary embolus and result in death (Venta and Venta, 1987). The gold standard diagnostic test for DVT, contrast venography, is an invasive procedure and can introduce allergic reaction. Therefore, ultrasonography is a commonly used surrogate test because it is noninvasive and has good accuracy. Alternatively, D-dimer is a small protein fragment present in the blood when there is a blood clot and thus testing its concentration in a screening blood test can also be used to diagnose DVT.

A recent paper by Kang and others (2013) presented a meta-analysis that included 12 studies comparing the accuracy of diagnostic tests for DVT. Among the 12 studies, 4 studies compared D-dimer test to venography, three studies compared ultrasonography to venography and 5 studies compared the D-dimer to ultrasonography (Kang and others, 2013). None of the studies compared the three tests together. A mixed-effects log-linear model was applied and random effects were incorporated to account for the heterogeneity in test accuracies of D-dimer but not for ultrasonography. In addition, the log linear model for test accuracies made it difficult to interpret the model parameters, and hard to generalize when comparing more diagnostic tests.

3. A UNIFIED STATISTICAL FRAMEWORK

We present a Bayesian hierarchical NMA-DT model to compare multiple tests simultaneously. In this article, we focus on modeling a commonly used pair of test accuracy indices, sensitivity (Se) and specificity (Sp), where sensitivity is the probability of a candidate test being positive given a diseased subject and specificity is the probability of a candidate test being negative given non-diseased (Pepe, 2003). In addition, other test accuracy indices such as positive and negative likelihood ratios (LR+ and LR−), positive and negative predictive values (PPV and NPV) can be useful in practice. LR+ (LR−) is the likelihood that a positive (negative) test result would be expected on a patient with the target disease compared to the likelihood that the same result would be expected on a patient without the disease. And PPV (NPV) describes the chance of truly diseased (non-diseased) given a positive (negative) test result. However, these test indices (PPV and NPV) are closely related to disease prevalence and estimation of these quantities requires information on prevalence. Furthermore, disease prevalence has been argued to be potentially correlated with Se and Sp and meta-analysis models accounting for such correlation are proposed (Chu and others, 2009; Leeflang and others, 2009). Therefore, the Bayesian hierarchical NMA-DT approach also models disease prevalence to account for these correlations and to provide the inference of other test accuracy indices. In this section, we first present the hierarchical model with random effects, then describe the prior distributions of the parameters. Next we provide the likelihood and posterior estimates.

3.1. Hierarchical model

We view different studies as if they were all designed to adopt a multiple test comparison design such that all studies should undergo a whole set of tests containing all candidate tests and a gold-standard. However, each of the studies includes a subset of the whole set, and the test outcomes from non-included tests are considered as missing data (Little and Rubin, 2002). We assume that the missing test outcomes are missing at random (MAR). Under MAR, the presence of a test does not depend on any unobserved characteristics, which means in our case missingness is independent of its sensitivity and specificity. In Section 3.4, we will provide a method for sensitivity analysis under missing not at random assumption (MNAR).

Let $T = \{T_0, T_1, \dots, T_K\}$ be a set of $K + 1$ binary diagnostic tests, where T_0 denotes a gold standard and $T_1, \dots, T_K$ stand for candidate tests under evaluation. Suppose we have a collection of $i = 1, \dots, N$ studies, where each of them reports outcomes of tests in a subset of T . In the i th study, for $k = 0, 1, \dots, K$, let y_{ijk} be the test outcome of T_k on subject j ($y_{ijk} = 1$ if positive and 0 if negative) and let δ_{ijk} be the missing data indicator ($\delta_{ijk} = 1$ if T_k is conducted to the j th subject and 0 if not). Let π_i be the study-specific disease prevalence: $\pi_i = P(y_{ij0} = 1)$, $i = 1, \dots, N$. For $k = 1, \dots, K$, let Se_{ik} and Sp_{ik} denote the study specific sensitivity and specificity for the k th test, respectively: $Se_{ik} = P(y_{ijk} = 1 | y_{ij0} = 1)$ and $Sp_{ik} = P(y_{ijk} = 0 | y_{ij0} = 0)$. Denote K_{ij} as the set of *candidate tests* conducted on subject j ($j = 1, \dots, J_i$) in the i th study, and $\mathbf{y}_{ij} = \{y_{ijk} : k \in K_{ij}\}$ as the collection of candidate test outcomes for this subject.

Multivariate random effects are used to account for potential across study heterogeneities in prevalence, sensitivities and specificities and correlations among them. Specifically, we write

$$\pi_i = \Phi(\eta + \varepsilon_i), Se_{ik} = \Phi(\alpha_k + \mu_{ik}), Sp_{ik} = \Phi(\beta_k + v_{ik}), i = 1, \dots, N, k = 1, \dots, K, \quad (3.1)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function for probit transformations. We note that other link functions can be specified as well. The parameter η is the fixed effect for prevalence. The parameters α_k and β_k are the fixed effects for sensitivity and specificity of T_k , respectively. The random effects ε_i , μ_{ik} and v_{ik} are the study specific effects for prevalence, sensitivity and specificity of T_k , respectively. It is straightforward to incorporate meta-regression covariates in (3.1) by

$$\pi_i = \Phi(\eta + \tilde{\boldsymbol{\eta}}\mathbf{X}_i + \varepsilon_i), Se_{ik} = \Phi(\alpha_k + \tilde{\boldsymbol{\alpha}}\mathbf{W}_i + \mu_{ik}), Sp_{ik} = \Phi(\beta_k + \tilde{\boldsymbol{\beta}}\mathbf{Z}_i + v_{ik}),$$

for $i = 1, \dots, N$ and $k = 1, \dots, K$, where $\mathbf{X}_i$, $\mathbf{W}_i$ and $\mathbf{Z}_i$ are study-level covariates such as study population characteristics, and $\tilde{\boldsymbol{\eta}}$, $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ are the corresponding coefficient vectors. In this paper, we focus on models without covariates for simplicity.

We introduce the within-study dependency structure of multiple test parameters by assuming that the random effect vector follows a multivariate normal distribution. Furthermore, this distribution also accounts for potential correlations between prevalence and the test accuracy parameters (Leeflang and others, 2009, Chu, Chen and Louis, 2009):

$$\boldsymbol{\theta}_i = (\varepsilon_i, \mu_{i1}, v_{i1}, \dots, \mu_{iK}, v_{iK})^T \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}), i = 1, \dots, N. \quad (3.2)$$

The covariance matrix $\boldsymbol{\Sigma}$ can be written as $\boldsymbol{\Sigma} = S\Omega S$, where S is a $(2K + 1) \times (2K + 1)$ diagonal matrix with diagonal elements $(\sigma_\pi, \sigma_{Se_1}, \sigma_{Sp_1}, \dots, \sigma_{Se_K}, \sigma_{Sp_K})$ capturing the between study heterogeneities and Ω is a positive definite correlation matrix whose diagonal elements are 1 and the off-diagonal elements measure potential correlations among disease prevalence and the test accuracy parameters. We assume the same correlation structure for all studies. Therefore, studies reporting all test outcomes of T contribute to estimating $\boldsymbol{\Sigma}$ and studies with missing test outcomes directly contribute to estimating a submatrix of $\boldsymbol{\Sigma}$. By assuming MAR and the same covariance matrix across all studies, which is equivalent to assuming

all studies apply the multiple test comparison design, the NMA-DT model can combine studies reporting different sets of candidate tests and make inferences on the relative test performances.

3.2. Likelihood specification

To derive the likelihood for the j th subject in the i th study, we first consider a subject that is tested by the gold standard test ($\delta_{ij0} = 1$) such that the true disease status is known. Conditional independence is assumed such that candidate test results are independent given the disease status. This assumption has been used in latent class models assessing accuracy of diagnostic tests without a gold standard (Chu, Chen and Louis, 2009). Denote $\alpha = (\alpha_1, \dots, \alpha_K)^T$, $\beta = (\beta_1, \dots, \beta_K)^T$, $\mu_i = (\mu_{i1}, \dots, \mu_{iK})^T$ and $\mathbf{v}_i = (v_{i1}, \dots, v_{iK})^T$, the probability of the test outcomes for a diseased subject, given the random effects, is calculated as,

$$\begin{aligned} P(\mathbf{y}_{ij}, y_{ij0} = 1 | \eta, \alpha, \varepsilon_i, \mu_i) &= P(y_{ij0} = 1)P(\mathbf{y}_{ij} | y_{ij0} = 1, \eta, \alpha, \varepsilon_i, \mu_i) \\ &= \pi_i \prod_{k \in K_{ij}} (Se_{ik})^{y_{ijk}} (1 - Se_{ik})^{(1-y_{ijk})} = \pi_i h_{ij1}, \end{aligned} \quad (3.3)$$

where $h_{ij1} = \prod_{k \in K_{ij}} (Se_{ik})^{y_{ijk}} (1 - Se_{ik})^{(1-y_{ijk})}$. Similarly, the probability for a non-diseased subject j in study i is given by

$$\begin{aligned} P(\mathbf{y}_{ij}, y_{ij0} = 0 | \eta, \beta, \varepsilon_i, \mathbf{v}_i) &= P(y_{ij0} = 0)P(\mathbf{y}_{ij} | y_{ij0} = 0, \eta, \beta, \varepsilon_i, \mathbf{v}_i) \\ &= (1 - \pi_i) \prod_{k \in K_{ij}} (Sp_{ik})^{(1-y_{ijk})} (1 - Sp_{ik})^{y_{ijk}} = (1 - \pi_i) h_{ij0}, \end{aligned} \quad (3.4)$$

where $h_{ij0} = \prod_{k \in K_{ij}} (Sp_{ik})^{(1-y_{ijk})} (1 - Sp_{ik})^{y_{ijk}}$.

Now we consider the setting where the subject j has not been tested by the gold standard T_0 (i.e., $\delta_{ij0} = 0$). By the law of total probability, the probability of the test outcomes $\mathbf{y}_{ij}$, given the random effects, can be written as

$$\begin{aligned} P(\mathbf{y}_{ij}, y_{ij0} = \text{missing} | \eta, \alpha, \beta, \varepsilon_i, \mu_i, \mathbf{v}_i) &= P(\mathbf{y}_{ij}, y_{ij0} = 1 | \eta, \alpha, \varepsilon_i, \mu_i) + P(\mathbf{y}_{ij}, y_{ij0} = 0 | \eta, \beta, \varepsilon_i, \mathbf{v}_i) \\ &= \pi_i h_{ij1} + (1 - \pi_i) h_{ij0}. \end{aligned} \quad (3.5)$$

In general, the probability of test outcomes for subject j (with or without the gold standard) can be written in the following unified form

$$(\pi_i h_{ij1})^{\delta_{ij0} y_{ij0}} [(1 - \pi_i) h_{ij0}]^{\delta_{ij0} (1 - y_{ij0})} [\pi_i h_{ij1} + (1 - \pi_i) h_{ij0}]^{(1 - \delta_{ij0})}. \quad (3.6)$$

3.3. Prior specifications and posterior estimations

In this subsection, we describe specifications of prior distributions of η, α, β and Σ . Conjugate Wishart prior can be assumed for the precision matrix: $\Sigma^{-1} \sim \text{Wishart}(\mathbf{R}, \nu)$. Taking the degrees of freedom ν equal to the dimension of Σ , $2K + 1$, will have approximately uniform prior in the correlation coefficients. Different choices of $\mathbf{R}$ can give relatively informative or non-informative priors on the variance parameters. Specific choices of $\mathbf{R}$ are discussed in the case studies.

Vague normal priors with mean 0 and variance 10 are assumed for η, α_k and β_k ($k = 1, \dots, K$), which correspond to equal tail 95% prior credible intervals (CI) of approximately (0,1) for π, Se_k and Sp_k , $k = 1, \dots, K$.

Denote the above prior distributions for α_k, β_k, η and Σ by $f_{\alpha_k}(\alpha_k), f_{\beta_k}(\beta_k), f_{\eta}(\eta)$, and $f(\Sigma)$, respectively. With these specifications, the joint posterior distribution is:

$$\begin{aligned} Pr(\eta, \alpha, \beta, \varepsilon_i, \mu_i, \mathbf{v}_i | \mathbf{y}) &\propto \prod_{i=1}^N \left[\prod_{j=1}^{J_i} P(y_{ij}, y_{ij0} | \eta, \alpha, \beta, \varepsilon_i, \mu_i, \mathbf{v}_i) P(\varepsilon_i, \mu_i, \mathbf{v}_i | \Sigma) \right] \\ &\times \left\{ \prod_{k=1}^K \left[f_{\alpha_k}(\alpha_k) f_{\beta_k}(\beta_k) \right] \right\} f_{\eta}(\eta) f(\Sigma). \end{aligned} \quad (3.7)$$

Given the likelihood specification in Section 3.2, the posterior can be written as

$$\begin{aligned} Pr(\eta, \alpha, \beta, \varepsilon_i, \mu_i, \mathbf{v}_i | \mathbf{y}) &\propto \\ &\prod_{i=1}^N \left[\prod_{j=1}^{J_i} (\pi_i h_{ij1})^{\delta_{ij0} y_{ij0}} [(1 - \pi_i) h_{ij0}]^{\delta_{ij0} (1 - y_{ij0})} [\pi_i h_{ij1} + (1 - \pi_i) h_{ij0}]^{(1 - \delta_{ij0})} |\Sigma|^{-1/2} e^{-\frac{1}{2} \theta_i' \Sigma^{-1} \theta_i} \right] \\ &\times \left\{ \prod_{k=1}^K \left[f_{\alpha_k}(\alpha_k) f_{\beta_k}(\beta_k) \right] \right\} f_{\eta}(\eta) f(\Sigma), \end{aligned} \quad (3.8)$$

where h_{ij1} and h_{ij0} are defined after equations (3.3) and (3.4), respectively, and θ_i is defined in equation (3.2).

We use JAGS software via the rjags package in R to sample from the joint posterior distribution using Markov Chain Monte Carlo (MCMC) methods (Lunn and others, 2000; Plummer and others, 2003). The posterior samples are drawn by Gibbs and Metropolis-Hasting's algorithms. Posterior estimates are similar to the maximum likelihood estimates when the priors are non-informative, and Bayesian approach allows for full posterior inference, so that the asymptotic approximations are not required. Convergence is assessed using trace plots, sample autocorrelation and Gelman-Rubin statistic (Gelman and Rubin, 1992).

Posterior samples of median disease prevalence, sensitivity and specificity of T_k can be achieved by MCMC sampling: $\pi = \Phi(\eta)$, $Se_k = \Phi(\alpha_k)$ and $Sp_k = \Phi(\beta_k)$. Posterior medians of other measures of clinical interest such as PPV, NPV, positive likelihood (LR+) and negative likelihood (LR-) of T_k can also be achieved:

$$\begin{aligned} PPV_k &= \frac{Se_k \pi}{Se_k \pi + (1 - Sp_k)(1 - \pi)}, \quad NPV_k = \frac{Sp_k(1 - \pi)}{Sp_k(1 - \pi) + (1 - Se_k)\pi}, \\ LR+_k &= \frac{Se_k}{1 - Sp_k}, \quad LR-_k = \frac{1 - Se_k}{Sp_k}. \end{aligned}$$

The surface under the cumulative ranking (SUCRA) can be used to rank test performance, accounting for the uncertainty in ranking (Salanti and others, 2011). The SUCRA for each candidate test can be calculated as follows.

$$SUCRA_k = \frac{\sum_{C=1}^{K-1} cum_{k,C}}{K - 1},$$

where $cum_{k,C}$ is the cumulative probability of T_k ranking as the C th best test: $cum_{k,C} = \sum_{c=1}^C P_{kc}$, where $P_{kc} = Pr(\text{rank}(Se_k) \leq c, \text{rank}(Sp_k) \leq c | \text{Data})$ can be calculated from posterior samples. Larger value

of SUCRA indicates better test performance, and SUCRA=1 (SUCRA=0) if the test always ranks first (last). Therefore, we can pick the best test by comparing their SUCRA value. This approach is most useful when the difference in preference between successive ranks is the same across the entire ranking scale, otherwise it can be misleading.

3.4. A sensitivity analysis for missingness not at random

The NMA-DT model is built upon the assumption of MAR. However, this assumption may be questionable in some applications. For example, researchers select candidate tests that are believed to have better performance, and hence missing test outcomes are related to unknown test accuracy parameters (which is MNAR). In this subsection, we present a model of missingness to be incorporated into the NMA-DT model to account for known MNAR mechanism. However, in practice, the MAR assumption is hardly known and is not testable. Thus, different models of missingness can be used in sensitivity analyses to evaluate the impact on parameter estimates if the MAR assumption is violated.

Let $N \times K$ matrix $\mathbf{M}$ denote the study-level missingness of a NMA-DT dataset containing N studies and K candidate tests. The entries of $\mathbf{M}$ are m_{ik} , $i = 1, \dots, N$ and $k = 1, \dots, K$, such that $m_{ik} = 1$ if T_k is missing in study i and $m_{ik} = 0$ otherwise. $m_{ik} \sim \text{Ber}(p_{ik})$, where p_{ik} is the probability of missing T_k in study i . We specify a model of missingness for p_{ik} as

$$\text{logit}(p_{ik}) = \gamma_k + \gamma_{1k} \times \text{logit}(Se_{ik}) + \gamma_{0k} \times \text{logit}(Sp_{ik}),$$

where γ_{1k} (γ_{0k}) controls the degree of association between the missing outcomes and the study-specific sensitivity (specificity). We assume non-positive γ_{1k} and γ_{0k} such that T_k is prone to be missing when it's accuracy is low. When $\gamma_{1k} = \gamma_{0k} = 0$, the outcomes of T_k are MAR with respect to its sensitivity and specificity. The model of missingness can then be incorporated in the likelihood in (3.6) under different pre-specified values of γ_{1k} and γ_{0k} . It should be noticed that the model of missingness described here is not for general MNAR scenarios, but is specific for the NMA-DT problem, and we only consider the scenarios when missingness is related to the test accuracy indices.

4. CASE STUDY RESULTS AND SENSITIVITY ANALYSES: NMA OF DVT TESTS

We analyze the NMA of DVT tests in Section 2 by the proposed NMA-DT model. In this study, we have $K = 2$. We adopt a relatively moderate informative Wishart prior with $\nu = 5$ and $\mathbf{R}$ with diagonal elements equal to 5 and off-diagonal elements equal to 0.05. This Wishart prior corresponds to a 95% prior CI of (0.2, 15) for the standard deviation components ($\sigma_\pi, \sigma_{Se1}, \sigma_{Sp1}, \sigma_{Se2}, \sigma_{Sp2}$). We fit the model by assuming vague $N(0, 10)$ priors for η, α_k and β_k . Here the assumed conditional independence between candidate tests given disease status implies that any agreement between D-dimer and ultrasound test results for a specific subject is only a result of the subject's disease status.

After 10 000 burn-in samples, 1 000 000 posterior samples are obtained. Table 1 shows the results from the proposed NMA-DT model. Figure 1 plots joint posterior distributions and study-specific posterior medians and 95% CIs for prevalence, sensitivity and specificity parameters. We write posterior medians followed by 95% CI in brackets for the rest of this article. The NMA-DT model concludes that ultrasonography has median Se of 0.90 (0.77, 0.96) and median Sp of 0.80 (0.54, 0.97). The D-dimer test is estimated to have moderate ability in diagnosing DVT with median Se 0.83 (0.68, 0.92) and median Sp 0.88 (0.75, 0.97). The SUCRA values for ultrasonography is 0.2 and for D-dimer is 0.13. Overall, ultrasonography is favored in detecting the diseased with higher sensitivity, whereas D-dimer performs better in ruling out the non-diseased with higher specificity.

Table 1. *Meta-analysis of DVT tests: Posterior median estimates and 95% CIs*

	Ultrasonography	D-dimer
Sensitivity	0.90 (0.77, 0.96)	0.83 (0.68, 0.92)
Specificity	0.80 (0.54, 0.97)	0.88 (0.75, 0.97)
PPV	0.84 (0.68, 0.96)	0.84 (0.68, 0.96)
NPV	0.91 (0.80, 0.97)	0.87 (0.77, 0.94)
LR+	4.39 (1.89, 27.90)	7.00 (3.10, 33.49)
LR−	0.13 (0.05, 0.33)	0.20 (0.09, 0.38)
Prevalence	0.43 (0.36, 0.50)

Fig. 1. Meta-analysis of DVT tests: forest plots and contour plots. (a) is the forest plot for prevalence, (b), (c) are forest plots for sensitivity and specificity of D-dimer, respectively, and (e), (f) are forest plots for sensitivity and specificity of ultrasound, respectively. The solid (dashed) lines denote the corresponding 95% credible intervals when the test is included (not included) in the study, respectively. (d) and (g) are the quantile countours of posterior sensitivity versus specificity at quantile levels 0.25, 0.5, 0.75, 0.9 and 0.95

4.1. Sensitivity analyses to prior distribution of Σ^{-1}

Sensitivity analyses to the prior distributions of Σ^{-1} are conducted to evaluate the effect of the prior distribution on the posterior prevalence, sensitivity and specificity. A relatively more informative Wishart

prior with $\nu = 5$ and $\mathbf{R}$ with diagonal elements equal to 20 and off-diagonal elements equal to 0.05 is used to repeat the analysis. This Wishart prior corresponds to a 95% prior CI of (0.1, 7.5) for the standard deviation components ($\sigma_\pi, \sigma_{Se_1}, \sigma_{Sp_1}, \sigma_{Se_2}, \sigma_{Sp_2}$). The posterior median disease prevalence is estimated to be 0.43 (0.37, 0.49). Ultrasonography has posterior median sensitivity of 0.89 (0.78, 0.96) and specificity of 0.79 (0.54, 0.96). The D-dimer test has posterior median sensitivity of 0.82 (0.67, 0.92) and specificity of 0.88 (0.75, 0.97). Similar posterior medians and 95% CIs compared to Table 1 are derived using a more informative prior.

A vague prior taking $\nu = 5$ and $\mathbf{R}$ with diagonal elements equal to 1 and off-diagonal elements equal to 0.05 is also used to repeat the analysis. This prior distribution corresponds to a 95% prior CI of (0.4, 35) for the standard deviation components. The posterior median of disease prevalence is 0.43 (0.33, 0.53). Ultrasonography has posterior median sensitivity 0.90 (0.74, 0.97) and specificity 0.82 (0.56, 0.98). D-dimer has posterior median sensitivity 0.83 (0.65, 0.93) and specificity of 0.89 (0.74, 0.98). Compared to Table 1, this prior leads to wider CIs for all parameters and slightly higher posterior medians of ultrasonography Sp.

Overall, different choices of the Wishart prior for Σ^{-1} have little effect on the posterior medians of prevalence, sensitivity and specificity but have slight influences on the width of their CIs.

4.2. Sensitivity analysis to the MAR assumption

The MAR assumption is untestable, but looking at the observed data might inform us the validity of this assumption. For example, in Figure 1(b) and (c), the Se and Sp estimates for D-dimer test are generally higher in studies 1-4 which includes D-dimer test and the gold standard, than other studies, due to some reasons leading to MNAR.

In this section, we conduct sensitivity analyses to explore the influence on parameter estimates when the MAR assumption is violated. We incorporate the model of missingness in Section 3.4 under different values of γ_{1k} and γ_{0k} : 0, -0.5, -1, and -2, which corresponds to MAR, an odds ratio of missingness 0.61, 0.37, and 0.13 (with respect to 1 unit increase in the logit scale of accuracy parameters), respectively.

The posterior medians of prevalence, sensitivities, and specificities are presented in Table 2 under different missingness assumptions: MAR, missingness related to accuracy of ultrasonography or D-dimer test only, missingness related to sensitivities of both tests or specificities only and missingness related to sensitivities and specificities of both tests. Compared to MAR, the estimates of π are barely affected under different assumptions. When missingness is negatively correlated with one of the tests, assuming MAR overestimates its Se and Sp while underestimates the other test's performance. When the missingness probabilities are negatively correlated with specificities, assuming MAR overestimates specificities but similar phenomenon is only observed for $\gamma_{1k} = -2$ when missingness is related to sensitivities. When the missing probabilities are negatively correlated with all parameters, assuming MAR generally overestimates all test accuracies (except for sensitivity estimates when the relation is weak). The differences between the estimates under the MAR and MNAR assumptions are generally enlarged when the bond between missing and test accuracy becomes stronger. In general, when missingness is negatively correlated with the test accuracy parameters, ignoring the model of missingness will overestimate the test performance. Note that, as shown in this example, due to the complex dependency structure of multiple test parameters, it is hard to tell whether the other tests will be over or underestimated when one of the tests is MNAR.

In summary, the estimates of diagnostic test accuracies were fairly robust under various MNAR models that we considered. The relative rank in test accuracies among these tests also preserved under different MNAR models. Additional model fitting results when the γ parameters are treated as random are presented in the supplementary material available at *Biostatistics* online, showing that the γ s are weakly identified from this model.

Table 2. Meta-analysis of DVT tests: median parameter estimates and 95% CIs under different missingness assumptions. MNAR=“None” is equivalent to MAR; MNAR=“D-dimer” (“Ultrasonography”) means missingness related to sensitivity and specificity of D-dimer test (ultrasonography); MNAR=“Se” (“Sp”) means missingness related to the sensitivities (specificities) of both the D-dimer test and ultrasonography; MNAR=“All” means missingness related to sensitivities and specificities of both tests. Bold numbers indicate parameters directly related to missingness

MNAR	γ_{11}	γ_{12}	γ_{01}	γ_{02}	π	Se: D-dimer	Se: ultrasonography	Sp: D-dimer	Sp: ultrasonography
None	0	0	0	0	0.43 (0.36,0.50)	0.83 (0.68,0.92)	0.90 (0.77,0.96)	0.88 (0.75,0.97)	0.80 (0.54,0.97)
D-dimer	-0.5	0	-0.5	0	0.44 (0.37,0.51)	0.81 (0.58,0.95)	0.94 (0.84,1)	0.84 (0.61,0.96)	0.8 (0.56,0.98)
Ultrasonography	0	-0.5	0	-0.5	0.43 (0.36,0.51)	0.89 (0.75,0.99)	0.89 (0.66,0.99)	0.91 (0.78,1)	0.61 (0.15,0.91)
Se	-0.5	-0.5	0	0	0.44 (0.37, 0.52)	0.86 (0.72,0.96)	0.93 (0.83,0.99)	0.88 (0.68,0.99)	0.73 (0.38,0.96)
Sp	0	0	-0.5	-0.5	0.43 (0.37,0.51)	0.85 (0.67,0.97)	0.91 (0.7,0.99)	0.88 (0.74,0.98)	0.76 (0.43,0.96)
All	-0.5	-0.5	-0.5	-0.5	0.43 (0.36,0.51)	0.85 (0.68,0.97)	0.92 (0.79,0.99)	0.87 (0.71,0.97)	0.72 (0.4,0.93)
D-dimer	-1	0	-1	0	0.44 (0.37,0.51)	0.78 (0.49,0.94)	0.95 (0.85,1)	0.80 (0.45,0.96)	0.83 (0.60, 1)
Ultrasonography	0	-1	0	-1	0.43 (0.36,0.50)	0.91 (0.78,1)	0.88 (0.64,0.98)	0.92 (0.79,1)	0.54 (0.11,0.89)
Se	-1	-1	0	0	0.43 (0.37,0.51)	0.84 (0.63,0.96)	0.89 (0.63,0.99)	0.90 (0.77,0.99)	0.79 (0.51,0.99)
Sp	0	0	-1	-1	0.43 (0.36,0.51)	0.86 (0.69,0.98)	0.93 (0.83,0.99)	0.86 (0.63,0.98)	0.70 (0.37,0.92)
All	-1	-1	-1	-1	0.44 (0.37,0.51)	0.85 (0.68,0.96)	0.90 (0.78,0.98)	0.87 (0.72,0.98)	0.71 (0.38,0.91)
D-dimer	-2	0	-2	0	0.44 (0.37,0.52)	0.77 (0.46,0.93)	0.95 (0.85,1)	0.81 (0.49,0.96)	0.85 (0.62,1)
Ultrasonography	0	-2	0	-2	0.43 (0.36,0.50)	0.91 (0.77,1)	0.87 (0.56,0.99)	0.92 (0.79,1)	0.53 (0.11, 0.87)
Se	-2	-2	0	0	0.44 (0.37, 0.52)	0.81 (0.56, 0.94)	0.84 (0.53, 0.97)	0.93 (0.81, 0.99)	0.83 (0.57, 0.98)
Sp	0	0	-2	-2	0.43 (0.36,0.51)	0.88 (0.74,0.98)	0.96 (0.86,1)	0.83 (0.61,0.96)	0.68 (0.38,0.89)
All	-2	-2	-2	-2	0.44 (0.37,0.51)	0.82 (0.55,0.96)	0.88 (0.65,0.98)	0.86 (0.58,0.99)	0.69 (0.25,0.92)

5. SIMULATION

5.1. Simulation setups

Simulation studies were conducted to test how the NMA-DT model performs under different assumptions. As the case study, we assume $K=2$, i.e., the whole test set contains two candidate tests (T_1 and T_2) and a gold standard (T_0). The Se (Sp) of T_1 is 0.8 (0.9) and the Se (Sp) of T_2 is 0.6 (0.7). The overall true disease prevalence is 0.4. We assume the random effects have standard deviations of 0.3: $(\sigma_\pi, \sigma_{Se_1}, \sigma_{Sp_1}, \dots, \sigma_{Se_K}, \sigma_{Sp_K}) = 0.3$. The correlations between prevalence and sensitivities are set to be 0.5. The correlations between prevalence and specificities, and the correlation between sensitivities and specificities are all set to be -0.5 .

We create partially missing data under MCAR, MAR, and MNAR assumptions, respectively. Under each scenario, we simulate 1000 replicates of NMA-DT datasets. Each dataset comprises 20 studies where 100 subjects are tested by both candidate tests and the gold standard. To generate test outcomes in each study, study specific prevalences, sensitivities and specificities are sampled from the multivariate normal distribution of the probit transformed parameters in Section 3.2, with the mean vector and covariance matrix specified above. Under the MCAR assumption, the missing indicators for all studies are prespecified such that the first five studies do not have a missing test outcome, the next five studies are missing T_1 , the next five are missing T_2 , and the last five studies are missing T_0 . In addition to the moderate correlation (0.5) assumption, we also investigate the model performance under weak (0.3) and strong (0.8) correlation assumptions for the probit transformed parameters. To generate the partially missing data under the MAR and MNAR assumptions, we only prespecify the missing indicator for the gold standard test such that T_0 is missing in five studies. We then apply a logit model with the observed (unobserved) data as covariates to calculate the missing probabilities of T_1 and T_2 for MAR (MNAR) scenario, which can be further used to generate the random missing indicators to drop the candidate tests in each study. The model for the missing indicators can be described as below:

For both MAR and MNAR: $m_{ik} \sim \text{Ber}(p_{ik})$

MAR: $\text{logit}(p_{ik}) = \beta_{0k}^{\text{MAR}} + \beta_{1k}^{\text{MAR}} \times \text{logit}(\pi_{ik})$

MNAR: $\text{logit}(p_{ik}) = \beta_{0k}^{\text{MNAR}} + \beta_{1k}^{\text{MNAR}} \times \text{logit}(Se_{ik}) + \beta_{2k}^{\text{MNAR}} \times \text{logit}(Sp_{ik}),$

where $k=1, 2$. Since each study performs at least 2 tests, if 1 test is missing, the missing indicators for the other 2 tests are automatically equal to 1. The intercept terms β_{0k}^{MAR} and β_{0k}^{MNAR} are chosen to control the average number of studies missing test k so that it is approximately equal to 5 in each scenario (as the MCAR scenario): $\beta_{01}^{\text{MAR}} = -20.97$, $\beta_{02}^{\text{MAR}} = -4.05$, $\beta_{01}^{\text{MNAR}} = 2.08$, $\beta_{02}^{\text{MNAR}} = 0.81$. To demonstrate the robustness of our model under the MAR assumption, we create a case where the missing probabilities highly depend on the observed prevalences: $\beta_{11}^{\text{MAR}} = 50$. Under the MNAR, both methods are biased, we choose $\beta_{11}^{\text{MNAR}} = \beta_{12}^{\text{MNAR}} = 2$ and $\beta_{21}^{\text{MNAR}} = \beta_{22}^{\text{MNAR}} = 0$ as an example to investigate the impacts of MNAR on the estimates from each method. Cross-classified cell counts can be collected for each study to present the observed data as in the case studies.

We compare the performance of the NMA-DT model with the “naive” approach. The “naive” method applies the trivariate generalized linear mixed model (TGLMM) (Chu and others, 2009) to studies reporting both T_1 and T_0 , accounting for potential correlations between disease prevalence and test accuracy parameters. Specifically, studies reporting T_1 and T_2 and studies reporting T_2 and T_0 are excluded from the naive analysis. The “naive” analysis is not applied to T_2 because T_1 and T_2 are exchangeable. Test outcomes of T_2 in studies reporting all three tests are ignored and only 2×2 tables cross-classifying outcomes of T_1 and T_0 are used to fit the trivariate GLMM. In total, 10 out of the 20 studies in each dataset are used to evaluate the performance of T_1 in the “naive” approach. The estimates of the fixed effects for prevalence, sensitivity and specificity of T_1 are compared with the estimates from the NMA-DT model.

Table 3. *Simulation results: bias, mean square error (MSE) and 95% CI coverage probabilities (CP) of the estimates for fixed effects $\eta, \alpha_1, \beta_1, \alpha_2, \beta_2$. Estimates from the proposed NMA-DT model and the “naive” method are compared for T_1*

Parameter (true)	NMA-DT			Naive		
	Bias	MSE	CP	Bias	MSE	CP
MCAR						
Weak Correlation (0.3)						
η (-0.25)	0.001	0.008	0.957	0.001	0.011	0.976
α_1 (0.84)	0.005	0.015	0.965	0.008	0.017	0.975
β_1 (1.28)	-0.001	0.013	0.966	0.003	0.015	0.978
α_2 (0.25)	0.006	0.012	0.963			
β_2 (0.52)	0.003	0.01	0.966			
Moderate Correlation (0.5)						
η (-0.25)	0.001	0.005	0.967	0.001	0.011	0.962
α_1 (0.84)	0.008	0.014	0.961	0.011	0.017	0.956
β_1 (1.28)	0.008	0.014	0.957	0.013	0.017	0.958
α_2 (0.25)	0.007	0.01	0.955			
β_2 (0.52)	0.007	0.009	0.959			
Strong Correlation (0.8)						
η (-0.25)	-0.006	0.007	0.964	-0.005	0.011	0.972
α_1 (0.84)	0.01	0.013	0.972	0.014	0.016	0.973
β_1 (1.28)	0.021	0.014	0.972	0.023	0.016	0.971
α_2 (0.25)	0.007	0.011	0.971			
β_2 (0.52)	0.01	0.01	0.969			
MAR						
Medium Correlation (0.5)						
η (-0.25)	-0.009	0.006	0.962	0.230	0.059	0.534
α_1 (0.84)	0.055	0.021	0.972	0.116	0.033	0.920
β_1 (1.28)	-0.040	0.020	0.967	-0.105	0.030	0.928
α_2 (0.25)	0.012	0.007	0.967			
β_2 (0.52)	0.005	0.007	0.955			
MNAR						
Medium Correlation (0.5)						
η (-0.25)	-0.009	0.006	0.954	0.047	0.012	0.959
α_1 (0.84)	0.073	0.018	0.932	0.104	0.024	0.926
β_1 (1.28)	-0.015	0.013	0.966	-0.036	0.016	0.968
α_2 (0.25)	0.032	0.010	0.959			
β_2 (0.52)	0.007	0.009	0.955			

5.2. Simulation results

Table 3 summarizes the bias, mean squared error (MSE) and 95% CI coverage probability (CP) of the fixed effects estimates using the proposed NMA-DT model (in column “NMA-DT”). Under different assumptions, the NMA-DT model is shown to provide nearly unbiased estimates for all parameters with small MSE. Generally, the estimates are more biased under MAR and MNAR assumptions, or as the correlation becomes stronger. The coverage probabilities remain close to the nominal level of 0.95 in all scenarios, and decrease under the MNAR assumption.

The fixed effects estimates for prevalence, sensitivity, and specificity for T_1 from the ‘naive’ approach are also summarized in Table 3, column “Naive.” Under the MCAR assumption, both methods perform reasonably well, although NMA-DT is slightly more efficient since it borrows strength from the indirect evidence. However, under MAR and MNAR assumptions, the estimates from the naive method are significantly more biased than those from the NMA-DT model, and the coverage probabilities may be substantially lower than 0.95 when the missingness is strongly associated with the observed or unobserved data. These observations are expected because the underlying assumption is MAR for NMA-DT, while it is MCAR for the naive approach.

6. DISCUSSION

There is a growing interest in simultaneously comparing the performance of multiple diagnostic tests in a network meta-analysis setting. However, due to mixture study designs, various reported test outcomes, heterogeneity in a meta-analysis, and complex correlation structure of multiple test outcomes, the methodological development for NMA-DT remains challenging. In this article, we presented a Bayesian hierarchical NMA-DT framework that unifies all three types of study designs into the multiple test comparison design using a missing data framework. In addition, it can provide ranks of diagnostic tests to guide clinical decision making. Through simulation studies, we have shown that the proposed method can provide unbiased estimates for prevalence and test accuracy. In addition, it is more efficient than a commonly used “naive” approach doing separate meta-analyses for each candidate test.

The NMA-DT model relies on an “consistency” assumption, which assumes that candidate tests would have been performed consistently on subjects assigned and not assigned to the test. However, inconsistency could happen, when studies that do not include T_1 may include a population for whom T_1 is inappropriate, and hence their performance may differ systematically from studies that do include T_1 . In this situation, the MAR assumption is questionable and borrowing information from studies must be done with caution. The concern of inconsistency is also discussed in contrast based NMA methods (Lu and Ades, 2006) that indirect evidence may be inconsistent with direct evidence. White and others (2012) proposed frequentist ways to estimate consistency and inconsistency models by expressing them as multivariate random-effects meta-regressions. Lu and Ades (2006) proposed to use inconsistency degrees of freedom to estimate the degree of inconsistency in evidence cycles. However, this method cannot be directly applied in NMA-DT because it is restricted to relative effects (e.g., log odds ratio) while NMA-DT is estimating marginal test accuracies (e.g., Se and Sp). On the other hand, researchers have been working on developing methods to detect inconsistency in arm-based NMA models (Hong, chu and others, 2016a; Zhao and others, 2016). Discussions have been elaborated around the comparison of contrast based and arm-based NMA methods (Dias and Ades, 2016; Hong and others, 2016b). Further research is needed to develop a formal test of inconsistency in NMA-DT.

An assumption made in the proposed model is the conditional independent test results given the true disease status and all study-specific diagnostic accuracy parameters. Such an assumption may be violated when two candidate tests are based on a similar biological mechanisms (Vacek, 1985). Attempts were made to account for this dependence through a correlation parameter (Chu, Chen and others, 2009), an additional latent class random effect (Qu and others, 1996), or multivariate probit models (Xu and Craig, 2009). However, they cannot be directly applied to NMA-DT model, because correlation parameters are only suitable for pairwise comparisons and only a small portion of the studies in NMA-DT may be subject to conditional dependence. Specifically, for studies adopting the randomized design, each candidate test is compared to the gold standard, thus conditional independence assumption is not required. For studies adopting the multiple test comparison design, conditional dependence may become a concern, since several candidate tests are compared simultaneously. Similarly, non-comparative designs may also suffer conditional dependence, but only when gold-standard test is not involved and subjects are tested by two

candidate tests. As a result, how to adjust for conditional dependence in NMA-DT is subject to future studies.

A concern brought by combining studies in a systematic review is how to correctly measure between-study heterogeneity. In this article, generalized linear mixed models are used to account for heterogeneity in a Bayesian framework, where posterior random effects covariance estimate measures the extent of heterogeneity. Inverse Wishart prior is used for the covariance matrix, but is limited in that the variance components are always positive. Another limitation is that when the correlation matrix grows, it imposes an unstructured covariance matrix while a structured correlation assumption may be more efficient. Applications of alternative priors for the covariance matrix (Daniels and Kass, 1999) in NMA-DT deserve further research.

Finally, we note that although multivariate normal distribution for random effects is assumed in the proposed model, it is straightforward to extend the proposed model to other multivariate distributions including distributions generated from copulas.

SUPPLEMENTARY MATERIAL

Supplementary material is available at <http://biostatistics.oxfordjournals.org>.

ACKNOWLEDGMENTS

We thank the associate editor and an anonymous reviewer for many constructive comments. *Conflict of Interest*: None declared.

FUNDING

Research reported in this publication was supported in part by NIAID R21 AI103012 (H.C., X.M.), NIDCR R03 DE024750 (H.C.), NLM R21 LM012197 (H.C.), NIDDK U01 DK106786 (H.C.), and NHLBI T32HL129956 (Q.L). The content is solely the responsibility of the authors and does not necessarily represent official views of the National Institutes of Health.

REFERENCES

- CHEN, Y., LIU, Y., NING, J., CORMIER, J. AND CHU, H. (2015). A hybrid model for combining case-control and cohort studies in systematic reviews of diagnostic tests. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **64**, 469–489.
- CHU, H., CHEN, S. AND LOUIS, T. A. (2009). Random effects models in a meta-analysis of the accuracy of two diagnostic tests without a gold standard. *Journal of the American Statistical Association* **104**, 512–523.
- CHU, H. AND COLE, S. R. (2006). Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology* **59**, 1331–1332.
- CHU, H., NIE, L., COLE, S. R. AND POOLE, C. (2009). Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: Alternative parameterizations and model selection. *Statistics in Medicine* **28**, 2384–2399.
- DANIELS, M. J. AND KASS, R. E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association* **94**, 1254–1263.
- DENDUKURI, N., SCHILLER, I., JOSEPH, L. AND PAI, M. (2012). Bayesian meta-analysis of the accuracy of a test for tuberculous pleuritis in the absence of a gold standard reference, *Biometrics* **68**, 1285–1293.
- DIAS, S. AND ADES, A. E. (2016). Absolute or relative effects? Arm-based synthesis of trial data, *Research Synthesis Methods* **7**, 23–28.

- GELMAN, A. AND RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences, *Statistical Science* **7**, 457–472.
- HAMZA, T. H., REITSMA, J. B. AND STIJNEN, T. (2008). Meta-analysis of diagnostic studies: a comparison of random intercept, normal-normal, and binomial-normal bivariate summary ROC approaches, *Medical Decision Making* **28**, 639–649.
- HARBORD, R. M., DEEKS, J. J., EGGER, M., WHITING, P. AND STERNE, J. A. (2007). A unification of models for meta-analysis of diagnostic accuracy studies, *Biostatistics* **8**, 239–251.
- HONG, H., CHU, H., ZHANG, J. AND CARLIN, B. P. (2016a). A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons, *Research Synthesis Methods* **7**, 6–22.
- HONG, H., CHU, H., ZHANG, J. AND CARLIN, B. P. (2016b). Rejoinder to the discussion of a Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons, by S. Dias and AE Ades, *Research Synthesis Methods* **7**, 29–33.
- KANG, J., BRANT, R. AND GHALI, W. A. (2013). Statistical methods for the meta-analysis of diagnostic tests must take into account the use of surrogate standards, *Journal of Clinical Epidemiology* **66**, 566–574.
- LEEFLANG, M. M., BOSSUYT, P. M. AND IRWIG, L. (2009). Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis, *Journal of Clinical Epidemiology* **62**, 5–12.
- LITTLE, R. J. AND RUBIN, D. (2002). *Statistical Analysis with Missing Data*, 2nd edition. Hoboken, NJ: John Wiley & Sons.
- LIU, Y., CHEN, Y. AND CHU, H. (2015). A unification of models for meta-analysis of diagnostic accuracy studies without a gold standard, *Biometrics* **71**, 538–547.
- LU, G. AND ADES, A. (2004). Combination of direct and indirect evidence in mixed treatment comparisons, *Statistics in Medicine* **23**, 3105–3124.
- LU, G. AND ADES, A. (2006). Assessing evidence inconsistency in mixed treatment comparisons, *Journal of the American Statistical Association* **101**, 447–459.
- LUNN, D. J., THOMAS, A., BEST, N. AND SPIEGELHALTER, D. (2000). WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility, *Statistics and Computing* **10**, 325–337.
- MA, X., CHEN, Y., COLE, S. R. AND CHU, H. (2016a). A hybrid Bayesian hierarchical model combining cohort and case-control studies for meta-analysis of diagnostic tests: accounting for partial verification bias, *Statistical Methods in Medical Research* **25**, 3015–3037.
- MA, X., NIE, L., COLE, S. R. AND CHU, H. (2016b). Statistical methods for multivariate meta-analysis of diagnostic tests: an overview and tutorial, *Statistical Methods in Medical Research* **25**, 1596–1619.
- PEPE, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*, chapter 2. Oxford: Oxford University Press.
- PLUMMER, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. pp. 20–22.
- QU, Y., TAN, M. AND KUTNER, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests, *Biometrics* **52**, 797–810.
- REITSMA, J. B., GLAS, A. S., RUTJES, A. W., SCHOLTEN, R. J., BOSSUYT, P. M. and ZWINDERMAN, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews, *Journal of Clinical Epidemiology* **58**, 982–990.
- RUTTER, C. M. AND GATSONIS, C. A. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations, *Statistics in Medicine* **20**, 2865–2884.

- SALANTI, G., ADES, A. AND IOANNIDIS, J. (2011). Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial, *Journal of Clinical Epidemiology* **64**, 163–171.
- TAKWOINGI, Y., LEEFLANG, M. M. AND DEEKS, J. J. (2013). Empirical evidence of the importance of comparative studies of diagnostic test accuracy, *Annals of Internal Medicine* **158**, 544–554.
- VACEK, P. M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests, *Biometrics* **41**, 959–968.
- VENTA, E. R. AND VENTA, L. A. (1987). The diagnosis of deep-vein thrombosis: an application of decision analysis, *Journal of the Operational Research Society* **38**, 615–624.
- WHITE, I. R., BARRETT, J. K., JACKSON, D. AND HIGGINS, J. (2012). Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression, *Research Synthesis Methods* **3**, 111–125.
- XU, H. AND CRAIG, B. A. (2009). A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests, *Biometrics* **65**, 1145–1155.
- ZHANG, J., CARLIN, B. P., NEATON, J. D., SOON, G. G., NIE, L., KANE, R., VIRNIG, B. A. AND CHU, H. (2014). Network meta-analysis of randomized clinical trials: reporting the proper summaries, *Clinical Trials* **11**, 246–262.
- ZHAO, H., HODGES, J. S., MA, H., JIANG, Q. AND CARLIN, B. P. (2016). Hierarchical Bayesian approaches for detecting inconsistency in network meta-analysis, *Statistics in Medicine* **35**, 3524–3536.

[Received April 27, 2016; revised February 24, 2017; accepted for publication March 18, 2017]