Published in final edited form as:

J Am Stat Assoc. 2019 ; 114(527): 949–961. doi:10.1080/01621459.2018.1476239.

A Bayesian Hierarchical Summary Receiver Operating Characteristic Model for Network Meta-analysis of Diagnostic Tests

Qinshu Lian¹, James S. Hodges¹, Haitao Chu^{1,*}

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

Abstract

In studies evaluating the accuracy of diagnostic tests, three designs are commonly used, crossover, randomized, and non-comparative. Existing methods for meta-analysis of diagnostic tests mainly consider the simple cases in which the reference test in all or none of the studies can be considered a gold standard test, and in which all studies use either a randomized or non-comparative design. The proliferation of diagnostic instruments and the diversity of study designs create a need for more general methods to combine studies that include or do not include a gold standard test and that use various designs. This paper extends the Bayesian hierarchical summary receiver operating characteristic model to network meta-analysis of diagnostic tests to simultaneously compare multiple tests within a missing data framework. The method accounts for correlations between multiple tests and for heterogeneity between studies. It also allows different studies to include different subsets of diagnostic tests and provides flexibility in the choice of summary statistics. The model is evaluated using simulations and illustrated using real data on tests for deep vein thrombosis, with sensitivity analyses.

Keywords

Diagnostic tests; Bayesian hierarchical model; Missing data; Multiple tests comparison; Network meta-analysis

1 Introduction

Statistical methods for meta-analysis of diagnostic tests have focused on combining and contrasting measures of test performance across multiple studies by parsing differences as real, likely explained by chance, or explicable by known differences in study characteristics (Macaskill et al. 2010). For meta-analysis of a single diagnostic test, several models have been proposed to jointly model the test's sensitivity and specificity, where sensitivity is defined as the probability that the test correctly identifies a patient who has the disease and specificity as the probability that the test correctly identifies a patient without disease. Such

*Corresponding author: chux0051@umn.edu.

Conflict of Interest: None declared.

meta-analyses have important characteristics: (1) a diagnostic test's estimated sensitivity and specificity are typically negatively correlated due to the trade-off between these accuracy measurements; (2) substantial between-study heterogeneity may be present due to clinical or methodological variation between studies, e.g., differences in positivity thresholds or variation in participants; (3) the continuum of measurable traits used to determine disease status may impose correlations between sensitivity, specificity, and disease prevalence (Leeftang et al. 2009; Chu, Nie, Cole and Poole 2009); and (4) if the reference test is not a gold standard test, its imperfect nature needs to be considered. These characteristics should be incorporated into meta-analyses of diagnostic tests.

To assess the diagnostic accuracy of *one* test and evaluate the trade-off between sensitivity and specificity in a meta-analysis, several fixed- and random-effect models have been proposed. For meta-analyses in which each study's reference test is a gold standard, the summary receiver operating characteristic (SROC) curve has been developed (Moses et al. 1993; Irwig et al. 1995; Dukic and Gatsonis 2003). It models the relationship between sensitivity and specificity across studies using fixed-effect regression models. Walter (2002) discussed properties of the resulting SROC curve, describing it as a function of the overall odd ratios. To capture heterogeneity between studies, Rutter and Gatsonis (2001) presented a hierarchical summary receiver operating characteristic (HSROC) model that combines study-specific estimates of sensitivity and specificity using a random effect model. In the absence of a gold standard test, latent class models based on the SROC or HSROC models have been proposed (Walter et al. 1999; Dendukuri et al. 2012). Besides these models, bivariate and multivariate random effect models (BGLMM or MGLMM) provide another important approach to meta-analysis of diagnostic tests (van Houwelingen et al. 2002; Chu and Cole 2006; Chu, Chen and Louis 2009; Reitsma et al. 2005; Sadatsafavi et al. 2010) in the presence or absence of a gold standard reference test. The BGLMM and MGLMM are easily fit using common software such as SAS, Stata, or R. However, the HSROC parameterization provides more flexibility by modeling the positivity thresholds and naturally leads to a SROC curve (Harbord et al. 2007). These two frameworks are equivalent under certain circumstances (Harbord et al. 2007; Liu et al. 2015). However, none of these meta-analysis models can combine studies that do and do not include a gold standard test.

While the above methods are mainly used when only one diagnostic test is of interest, some studies simultaneously evaluate *multiple* diagnostic tests for a given disease. Three designs are commonly used in studies comparing multiple tests (Takwoingi et al. 2013): crossover (also called paired or multiple tests design), randomized, and non-comparative. In a crossover design, all patients undergo all tests including a reference standard test; in a randomized design, all patients undergo a reference standard test and one randomly assigned index test. We use the term "index test" for any diagnostic test other than the gold standard. These two designs are recommended because tests are compared in the same or similar populations. Many studies use a non-comparative design in which different sets of patients undergo each index test and a reference test. The rapid evolution of diagnostic strategies and the diversity of study designs lead to five challenges in multiple-test comparison. First, estimates of accuracy indices for index tests are typically correlated because multiple tests are included in the same study. Efficiency can be lost by ignoring these correlations. Second, few studies in a meta-analysis compare all diagnostic tests of interest; usually, different

studies include different subsets of tests. Third, usually some studies use a gold standard reference test while others use an error-prone reference. Fourth, different studies use different designs. Finally, different tests may not be equally heterogeneous across studies.

So far, most available meta-analysis methods for comparing multiple tests either undertake separate meta-analyses for each index test or conduct a meta-regression with the type of diagnostic test as a covariate (Rutter and Gatsonis 2001; Reitsma et al. 2005). The former approach is essentially a complete case analysis that implicitly assumes unavailable test results are missing completely at random. The latter approach assumes the different tests have homogeneous variances. Both methods do not account for the correlations between tests applied to the same studies. Trikalinos et al. (2014) proposed models for joint meta-analysis of studies comparing multiple index tests in crossover designs, which can be extended to incorporate randomized and non-comparative designs. However, all these methods require each patient's true disease status to be known, i.e., that each study's reference test is a gold standard. To the best of our knowledge, no existing meta-analysis method for comparing multiple diagnostic tests can simultaneously incorporate studies with different designs and studies with or without a gold standard test while properly accounting for correlations between multiple index tests and heterogeneity between studies.

To address these limitations and challenges, and motivated by the literature on network meta-analysis of randomized clinical trials (Caldwell et al. 2005; Lu and Ades 2006; Mills et al. 2012; Zhang et al. 2014), we present a Bayesian hierarchical summary receiver operating characteristic model for network meta-analysis of diagnostic tests (HSROC-NMADT). Network meta-analysis borrows strength from indirect evidence, which can improve statistical efficiency and reduce bias. We treat all studies as if they *could have* adopted a crossover design in which all patients underwent the whole set of index tests and a gold standard test. If a study's reference test is not a gold standard, it is treated as an index test. If a test was not in fact evaluated, its results are treated as missing data in a missing data framework. We consider only dichotomous diagnostic tests.

Section 2 describes a motivating example, tests for deep vein thrombosis. Section 3 presents the proposed HSROC-NMADT model and a Bayesian analysis of it. Section 4 demonstrates the method by analyzing data from the motivating example. Section 5 describes simulation studies illustrating our approach's performance under various conditions. Finally, Section 6 discusses our findings and implications for future work.

2 A Motivating Study

Deep vein thrombosis (DVT), or deep venous thrombosis, is a blood clot that forms in a vein deep in the body. It is imperative to diagnose DVT correctly and early: an untreated thrombus can cause a fatal pulmonary embolism, while anticoagulation in the absence of thrombosis is unethical (Kyrle and Eichinger 2005). Clinical assessments of DVT based on physical examination and medical history are not reliable; in practice, laboratory studies and imaging techniques are used for diagnosis. Contrast venography is regarded as the gold standard test but is not always feasible because it is invasive and has potential contraindications (Tovey and Wyatt 2003; Kyrle and Eichinger 2005). Ultrasonography is

considered the best non-invasive alternative, while another safe and feasible choice is measurement of D-dimer, a small protein fragment present in the blood after a clot is degraded by fibrinolysis (Perone et al. 2001; Wells et al. 2003; Scarvelis and Wells 2006).

Several studies have used comparative or non-comparative designs to measure the accuracy of these tests. Kang et al. (2013) did a meta-analysis of 12 studies evaluating the accuracy of the D-dimer test, using a semi-quantitative latex (SL) assay for D-dimer. Among the 12 studies, four compared only the D-dimer test to venography, three compared only ultrasonography to venography, and five compared only D-dimer to ultrasonography. Thus seven of the 12 studies included the gold standard test while the rest did not. Kang et al. proposed a mixed-effects log-linear model to combine studies with or without a gold standard test and took account of the reference test's imperfect nature and between-study heterogeneity in disease prevalence and the performance of D-dimer testing. However, the analysis did not account for possible heterogeneity in the performance of ultrasonography, which may bias the estimates and conclusions. Also, the model was suitable only because each study compared two tests. If all three diagnostic tests had been applied to the patients in a single study, this approach could not account for correlations among the three tests included in the same study.

3 Bayesian Network Meta-analysis of Diagnostic Tests

This section presents a Bayesian hierarchical summary receiver operating characteristic model for network meta-analysis of diagnostic tests (HSROC-NMADT), which compares multiple tests simultaneously. Suppose we wish to evaluate the performance of K index tests, denoted as $T_1, T_2, \dots, T_K$, with T_0 denoting a gold standard test. Our main goal is to estimate the overall disease prevalence and the sensitivity and specificity of each index test, with secondary goals to estimate other accuracy indices such as positive predictive value (PPV) and negative predictive value (NPV). We show how to estimate an SROC curve for each index test using samples from the posterior of the model parameters.

3.1 The model for network meta-analysis of diagnostic tests

We consider a network meta-analysis of diagnostic tests with a collection of N studies. Each study reports results only for a subset of the complete collection of $K + 1$ diagnostic tests ($T_0, T_1, T_2, \dots, T_K$). We view studies of different designs as if they all could have adopted the crossover design in which each subject is diagnosed by the whole set of index tests and verified by a gold standard test. In each study, unavailable test outcomes are considered missing. For now, we assume they are missing at random (MAR) (Little and Rubin 2002); data is called MAR if, given the observed data, failure to observe a test result does not depend on unobserved data. Section 4.2 gives a sensitivity analysis of the MAR assumption.

All tests are dichotomous, taking the value 1 when positive and 0 when negative. Let y_{ijk} denote the outcome of T_k for subject j in study i ($i = 1, 2, \dots, N, j = 1, 2, \dots, N_i$). Let y_{ij0} be the outcome of the gold standard test, which is patient j 's true disease status D_{ij} , i.e., $y_{ij0} = 1$ if D_{ij} is positive, and 0 otherwise. Let $\mathcal{K}_i$ be the subset of index tests included in study i ; K_i is the number of tests in $\mathcal{K}_i$. Let δ_{ik} be the missingness indicator for T_k in study i , taking the

value 1 when the test result is available and 0 otherwise. Disease prevalence in study i is π_i and let Se_{ik} and Sp_{ik} denote the sensitivity and specificity respectively of test k in study i , so $Se_{ik} = P(y_{ijk} = 1 | y_{ij0} = 1)$, $Sp_{ik} = P(y_{ijk} = 0 | y_{ij0} = 0)$, and $\pi_i = P(y_{ij0} = 1)$.

Extending along the lines of the pioneering work by Rutter and Gatsonis (2001) and Liu et al. (2015) on a single diagnostic test, for multiple tests we construct a hierarchical model with three levels to capture variability within a study, heterogeneity between studies, and correlations among tests in the same study. We make a crucial assumption of conditional independence, i.e., the y_{ijk} s are independent given the true disease status y_{ij0} and unknown parameters α_i , θ_i and β defined below. Section 3.2.3 discusses the implications of this assumption and Section 4.3 gives a sensitivity analysis allowing a form of conditional dependence.

3.1.1 Level I (within-study)—We assume the dichotomous outcome of index test k applied to patient j in study i arises from an underlying continuous latent variable Z_{ijk} . If Z_{ijk} is greater than the study-specific cutoff θ_{ik} , then T_k gives a positive result; otherwise, it gives a negative result. Under the conditional independence assumption, we assume Z_{ijk} independently follows a normal distribution given the true disease status and the unknown parameters α_i , θ_i and β , specifically

$$Z_{ijk} \sim \begin{cases} N(-\alpha_{ik}/2, \exp(-\beta_k/2)) & \text{for } y_{ij0} = 0, \\ N(\alpha_{ik}/2, \exp(\beta_k/2)) & \text{for } y_{ij0} = 1. \end{cases} \quad (1)$$

The study-specific parameter α_{ik} is called the “accuracy value” of test k in study i because it measures the distance between the distributions of Z_{ijk} when disease is present versus absent. The parameter β_k is called a “shape parameter” because it allows difference in the variance in the outcomes of diseased and disease-free populations. We assume the β_k s are the same in all studies for identifiability; Rutter and Gatsonis (2001), Dendukuri et al. (2012), and Liu et al. (2015) made a similar assumption. With these assumptions, T_k 's study-specific sensitivity and specificity are

$$Se_{ik} = \Phi\left(-\frac{\theta_{ik} - \alpha_{ik}/2}{\exp(\beta_k/2)}\right), \quad Sp_{ik} = \Phi\left(\frac{\theta_{ik} + \alpha_{ik}/2}{\exp(-\beta_k/2)}\right) \quad (2)$$

respectively, where $\Phi(\cdot)$ denotes the probit link (the standard normal cumulative distribution function). To account for heterogeneity between studies in disease prevalence, we assume disease status D_{ij} is positive ($y_{ij0} = 1$) if a latent variable Z_{ij0} with a standard normal distribution is greater than a study-specific cutoff θ_0 and negative ($y_{ij0} = 0$) otherwise. Therefore, the study-specific prevalence is $\pi_i = \Phi(-\theta_0)$.

3.1.2 Level II (between-study)—Let $\alpha_i = (\alpha_{i1}, \dots, \alpha_{iK})$ and $\theta_i = (\theta_0, \theta_{i1}, \dots, \theta_{iK})$, $i = 1, 2, \dots, N$ and assume α_i and θ_i are mutually independent. To account for heterogeneity between studies and correlations between multiple index tests, we assume the study-specific cutoff and accuracy values follow multivariate normal distributions

$$\begin{aligned}
 & \begin{pmatrix} \theta_{i0} \\ \theta_{i1} \\ \vdots \\ \theta_{iK} \end{pmatrix} \stackrel{iid}{\sim} \text{MVN} \left(\Theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_K \end{pmatrix}, \Sigma_{\Theta} = \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \cdots & \sigma_{0K} \\ \sigma_{01} & \sigma_1^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{0K} & \cdots & \cdots & \sigma_K^2 \end{pmatrix} \right) \text{ and} \\
 & \begin{pmatrix} \alpha_{i1} \\ \alpha_{i2} \\ \vdots \\ \alpha_{iK} \end{pmatrix} \stackrel{iid}{\sim} \text{MVN} \left(\Lambda = \begin{pmatrix} \Lambda_1 \\ \Lambda_2 \\ \vdots \\ \Lambda_K \end{pmatrix}, \Sigma_{\Lambda} = \begin{pmatrix} \tau_1^2 & \tau_{12} & \cdots & \tau_{1K} \\ \tau_{12} & \tau_2^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{1K} & \cdots & \cdots & \tau_K^2 \end{pmatrix} \right), i = 1, \dots, N,
 \end{aligned} \tag{3}$$

where Σ_{Θ} and Σ_{Λ} are positive definite covariance matrices with diagonals describing between-study variability in the cutoff and accuracy values respectively, and off-diagonals describing between-test covariance between pairs of cutoff and accuracy values, respectively.

In the HSROC framework, the cutoff and accuracy values are independent characteristics that jointly induce correlation between an index test's sensitivity and specificity (Rutter and Gatsonis 2001). The sensitivities and specificities of tests may also be associated with disease prevalence because the definition and severity of disease may vary between studies due to study designs and populations (Li et al. 2007; Leeflang et al. 2009; Chu, Nie, Cole and Poole 2009). By assuming the cutoff value that determines disease prevalence, θ_0 , and the cutoff values of the index tests, $(\theta_1, \theta_2, \dots, \theta_K)$, have a joint normal distribution, we implicitly allow the sensitivities and specificities to be associated with disease prevalence through $\sigma_{01}, \dots, \sigma_{0K}$. By assuming that the covariance matrices Σ_{Θ} and Σ_{Λ} are unstructured, we let the data determine inferences about these correlations. These dependence structures may be simplified if prior knowledge is available; Section 4 considers reduced models.

3.1.3 Level III (prior specifications)—The Bayesian specification is completed with vague normal priors with large variances for Θ_0, Θ_k and $\Lambda_k, k = 1, 2, \dots, K$. The prior for the shape parameter β_k is $\text{Uniform}(b_1, b_2)$. The choices of b_1 and b_2 depend on prior knowledge about the diagnostic tests and should cover all possible β_k .

As Rutter and Gatsonis (2001) noted, the posterior distributions are potentially sensitive to the priors on the covariance matrices of the cutoff and accuracy values; in particular, the prior affects the width of posterior credible intervals. In general, these priors should not assign too much probability to large diagonal elements (i.e., variance parameters), while still placing a diffuse distribution on the correlations. Therefore, the usual inverse-Wishart prior for a covariance is not recommended because of its restrictive form. To model these covariance matrices, we adopt a separation strategy (Barnard et al. 2000; O'Malley and Zaslavsky 2008), decomposing the covariance matrix of the cutoff values as $\Sigma_{\Theta} = \text{diag}(S)R \text{diag}(S)$, where $\text{diag}(S)$ is a diagonal matrix with diagonal values $S = (s_0, s_1, \dots, s_K)$ and R is a $(K+1) \times (K+1)$ positive definite matrix that determines the correlations between the

cutoff values, but is not a correlation matrix. We use this non-identified parameterization to simplify computing; all functionals of interest are identified. Specifically, the cutoff values have standard deviations $\sigma_k = \sqrt{s_k^2 R_{k,k}}$, $k = 0, 1, \dots, K$, and the correlation matrix is $\text{Corr}_\Theta = \text{diag}(R)^{-1/2} R \text{diag}(R)^{-1/2}$. Similarly, the covariance matrix of the accuracy values, Σ_Λ , is decomposed as $\Sigma_\Lambda = \text{diag}(P\Omega) \text{diag}(P)$, where $P = (p_1, p_1, \dots, p_K)$, and the standard deviations and correlation matrix are $\tau_k = \sqrt{p_k^2 \Omega_{k,k}}$ and $\text{Corr}_\Lambda = \text{diag}(\Omega)^{-1/2} \text{diag}(\Omega)^{-1/2}$ respectively. We assign a $N(\eta, \zeta^2)$ prior to each element of $\log(P)$ and $\log(S)$, with η and ζ^2 reflecting prior knowledge about these standard deviations. Inverse-Wishart priors $IW(I_{K+1}, K+2)$ and $IW(I_K, K+1)$ are placed on R and Ω respectively, where I_K is the identity matrix of dimension K , so the marginal priors of all correlation parameters are approximately uniform.

Under the MAR assumption, studies with all test outcomes contribute to estimating the full matrices Σ_Θ and Σ_Λ , while studies with partial test outcomes contribute to estimating only submatrices of Σ_Θ and Σ_Λ .

3.2 Estimation

This section derives the likelihood and posterior distribution and shows how to obtain posterior distributions for summary statistics.

3.2.1 The Likelihood—For the above model, we first express the likelihood of the observed data in terms of the study-specific prevalence, sensitivity, and specificity. In the proposed model, we assume $Y_{ijk} \sim \text{Bernoulli}(Se_{ik})$, $k > 0$, when $y_{ij0} = 1$, and $Y_{ijk} \sim \text{Bernoulli}(1 - Sp_{ik})$, $k > 0$, when $y_{ij0} = 0$. Let $\mathbf{y}_{ij}^c = (\mathbf{y}_{ij}^o, \mathbf{y}_{ij}^m)$ be the outcomes of the complete test set for subject j in study i , where the superscripts o and m denote observed and missing outcomes respectively. Let $\mathbf{y}^o = \{\mathbf{y}_{ij}^o; \text{all } i, j\}$ and $\mathbf{y}^m = \{\mathbf{y}_{ij}^m; \text{all } i, j\}$. The vector $\boldsymbol{\xi}$ denotes the unknown parameters in the HSROC-NMADT model above and $\boldsymbol{\gamma}$ denotes the parameters of the mechanism that determines missingness. (We have not specified this mechanism because under the MAR assumption, we do not need to.) Let $\boldsymbol{\delta} = \{\delta_{ik}; \text{all } i, k\}$ be the missingness indicators. Assume $\boldsymbol{\xi}$ and $\boldsymbol{\gamma}$ are functionally independent, which is a reasonable and common assumption in practice (Higgins et al. 2008; Ibrahim and Molenberghs 2009). Assuming MAR, the distribution of the observed data $\{\mathbf{y}^o, \boldsymbol{\delta}\}$ can be factored into a marginal density for observed test outcomes and a conditional density for the missingness indicators given the observed data, i.e., $f(\mathbf{y}^o, \boldsymbol{\delta} | \boldsymbol{\xi}, \boldsymbol{\gamma}) = f(\mathbf{y}^o | \boldsymbol{\xi}, \boldsymbol{\gamma}) f(\boldsymbol{\delta} | \mathbf{y}^o, \boldsymbol{\xi})$ (see Appendix A). As a result, the missing data mechanism $f(\boldsymbol{\delta} | \mathbf{y}^o, \boldsymbol{\gamma})$ is ignorable and inference about $\boldsymbol{\xi}$ is based solely on $f(\mathbf{y}^o | \boldsymbol{\xi})$.

If study i included the gold standard test, the probability of subject j 's observed test outcomes is

$$\begin{aligned}
 l_{ij1} &= P(y_{ij-}^o, y_{ij0}^o \mid \theta_i, \alpha_i, \beta) = P(y_{ij0}^o \mid \theta_i) P(y_{ij-}^o \mid \theta_i, \alpha_i, \beta, y_{ij0}^o) \\
 &= P(y_{ij0}^o \mid \theta_i) \prod_{k \in \mathcal{K}_i} P(y_{ijk}^o \mid \theta_i, \alpha_i, \beta, y_{ij0}^o),
 \end{aligned} \tag{4}$$

where y_{ij-}^o is the vector of observed *index test* outcomes and $\mathcal{K}_i$ is the subset of index tests included in study i . Consequently, study i 's likelihood contribution is

$$\begin{aligned}
 l_{i1} &= \prod_j P(y_{ij0}^o \mid \theta_i, \alpha_i, \beta) P(y_{ij-}^o \mid \theta_i, \alpha_i, \beta, y_{ij0}^o) \quad \text{by conditional independence} \\
 &= \prod_j \left\{ (\pi_i)^{y_{ij0}^o} (1 - \pi_i)^{1 - y_{ij0}^o} \prod_{k \in \mathcal{K}_i} \left[(Se_{ik})^{y_{ijk}^o} (1 - Se_{ik})^{(1 - y_{ijk}^o)} \right]^{y_{ij0}^o} \left[(1 - Sp_{ik})^{y_{ijk}^o} (Sp_{ik})^{(1 - y_{ijk}^o)} \right]^{1 - y_{ij0}^o} \right\} \\
 &= (\pi_i)^{\sum_j y_{ij0}^o} (1 - \pi_i)^{n_i - \sum_j y_{ij0}^o} \times \\
 &\quad \prod_{k \in \mathcal{K}_i} Se_{ik}^{\sum_j y_{ijk}^o y_{ij0}^o} (1 - Se_{ik})^{\sum_j (1 - y_{ijk}^o) y_{ij0}^o} (1 - Sp_{ik})^{\sum_j y_{ijk}^o (1 - y_{ij0}^o)} Sp_{ik}^{\sum_j (1 - y_{ijk}^o) (1 - y_{ij0}^o)},
 \end{aligned} \tag{5}$$

where Se_{ik} and Sp_{ik} are given in Equation (2). Note that $\sum_j y_{ijk}^o y_{ij0}^o$ is the number of diseased subjects with a positive result for test k , $\sum_j (1 - y_{ijk}^o) y_{ij0}^o$ is the number of diseased subjects with a negative result for test k , $\sum_j y_{ijk}^o (1 - y_{ij0}^o)$ is the number of healthy subjects with a positive result for test k , and $\sum_j (1 - y_{ijk}^o) (1 - y_{ij0}^o)$ is the number of healthy subjects with a negative result for test k . These four sums are the four cells in the marginal 2×2 cross-tabulation of index test k and the gold standard test.

If study i did not include the gold standard test, the probability of subject j 's observed outcomes is

$$\begin{aligned}
 l_{ij0} &= P(y_{ij}^o \mid \theta_i, \alpha_i, \beta) = P(y_{ij}^o, y_{ij0}^m = 1 \mid \theta_i, \alpha_i, \beta) + P(y_{ij}^o, y_{ij0}^m = 0 \mid \theta_i, \alpha_i, \beta) \\
 &= P(y_{ij0}^m = 1 \mid \theta_i, \alpha_i, \beta) \prod_{k \in \mathcal{K}_i} P(y_{ijk}^o \mid y_{ij0}^m = 1, \theta_i, \alpha_i, \beta) + \\
 &\quad P(y_{ij0}^m = 0 \mid \theta_i, \alpha_i, \beta) \prod_{k \in \mathcal{K}_i} P(y_{ijk}^o \mid y_{ij0}^m = 0, \theta_i, \alpha_i, \beta) \\
 &= \left(\pi_i \right) \prod_{k \in \mathcal{K}_i} (Se_{ik})^{y_{ijk}^o} (1 - Se_{ik})^{(1 - y_{ijk}^o)} + \left(1 - \pi_i \right) \prod_{k \in \mathcal{K}_i} (1 - Sp_{ik})^{y_{ijk}^o} (Sp_{ik})^{(1 - y_{ijk}^o)},
 \end{aligned} \tag{6}$$

where the last line also follows by conditional independence. Therefore, the likelihood contribution from a study without the gold standard test is

$$\begin{aligned}
l_{i0} &= \prod_j \left[\left(\pi_i \prod_{k \in \mathcal{K}_i} (Se_{ik})^{y_{ijk}^o} (1 - Se_{ik})^{(1 - y_{ijk}^o)} + (1 - \pi_i) \prod_{k \in \mathcal{K}_i} (1 - Sp_{ik})^{y_{ijk}^o} (Sp_{ik})^{(1 - y_{ijk}^o)} \right) \right] \\
&= \prod_{\mathcal{S}_i} \left[\left(\pi_i \prod_{k \in \mathcal{K}_i} (Se_{ik})^{t_k} (1 - Se_{ik})^{(1 - t_k)} + (1 - \pi_i) \prod_{k \in \mathcal{K}_i} (1 - Sp_{ik})^{t_k} (Sp_{ik})^{(1 - t_k)} \right) \right]^{n_{i1} t_1 \dots t_K}, \quad (7)
\end{aligned}$$

In (7), $\mathcal{S}_i$ is the set of distinct possible test results, with members having the form $\{T_1 = t_1, \dots, T_K = t_K\}$ for $t_k \in \{0, 1, m\}$ indicating test k has a positive, negative, or unobserved result respectively. $\mathcal{S}_i$ has 2^{K_i} members; the test results of person j in study i are one of $\mathcal{S}_i$ members. The count of subjects having results equal to $\{T_1 = t_1, \dots, T_K = t_K\}$ is $n_{i1} t_1 \dots t_K$. For example, if $K = 4$ and $\mathcal{K}_i = \{T_1, T_2, T_4\}$, then $K_i = 3$ and

$\mathcal{S}_i = \{n_{i11m1}, n_{i11m0}, n_{i10m1}, n_{i10m0}, n_{i01m1}, n_{i00m1}, n_{i01m0}, n_{i00m0}\}$. As before, Se_{ik} and Sp_{ik} are given by Equation (2).

Equations (5) and (7) show that our approach requires these data: from studies including the gold-standard test, the 2×2 cross-tabulation of each index test with the gold standard test; and from study i omitting the gold-standard test and including K_i index tests, the K_i -way cross-tabulation describing the results of all K_i tests for each of study i 's subjects. Equations (4) and (5) prove that assuming conditional independence, the $(K_i + 1)$ -way cross-tabulation for a study with a gold standard test provides no more information than the $K_i 2 \times 2$ cross-tabulations of the index tests with the gold standard test.

The likelihood function of the observed data can be summarized as

$$P(y^o | \theta_i, \alpha_i, \beta) = \prod_i (l_{i1})^{\delta_{i0}} (l_{i0})^{1 - \delta_{i0}}. \quad (8)$$

where δ_{i0} is the missingness indicator for the gold standard test. The total number of degrees of freedom available depends on the designs of the included studies. A study with a gold standard test contributes $2^{(K_i+1)} - 1$ degrees of freedom; a study without a gold standard test contributes $2^{K_i} - 1$. Therefore, depending on the designs of the N included studies, the total degrees of freedom is between $3N$ and $N \times (2^{K+1} - 1)$. The total number of parameters to be estimated is at least $K^2 + 5K + 2$. Therefore, the minimum number of studies N required to estimate this model without informative prior distributions is between $\frac{K^2 + 5K + 2}{2^{K+1} - 1}$ and

$\frac{K^2 + 5K + 2}{3}$, depending on the study types.

3.2.2 The Posterior Distribution—For the priors specified in Section 3.1.3, the joint posterior distribution is:

$$\begin{aligned}
& L(\boldsymbol{\theta}_i, \boldsymbol{\alpha}_i, \boldsymbol{\beta}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\theta}}, \boldsymbol{\Lambda}, \Sigma_{\boldsymbol{\Lambda}} | \mathbf{y}^o) \\
& \propto \prod_i \left[\prod_j \left((l_{ij1})^{\delta_{i0}} (l_{ij0})^{1-\delta_{i0}} \right) | \Sigma_{\boldsymbol{\theta}} |^{-1/2} e^{-\frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\Theta})' \Sigma_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\Theta})} | \Sigma_{\boldsymbol{\Lambda}} |^{-1/2} e^{-\frac{1}{2}(\boldsymbol{\alpha}_i - \boldsymbol{\Lambda})' \Sigma_{\boldsymbol{\Lambda}}^{-1} (\boldsymbol{\alpha}_i - \boldsymbol{\Lambda})} \right] \quad (9) \\
& f_{\boldsymbol{\theta}}(\boldsymbol{\Theta}) f_{\boldsymbol{\Lambda}}(\boldsymbol{\Lambda}) f_P(P) f_S(S) f_R(R) f_{\Omega}(\Omega) f_{\boldsymbol{\beta}}(\boldsymbol{\beta}).
\end{aligned}$$

We sample from the joint posterior using Markov Chain Monte Carlo (MCMC) as implemented in JAGS v.4.2.0 via the rjags package in R v.3.3.2 (Plummer 2003; Plummer 2016). The overall prevalence (Π) of the disease and the sensitivity (Se_k) and specificity (Sp_k) of each index test can be summarized as

$$\Pi = \Phi\left(-\theta_0\right), \quad Se_k = \Phi\left(-\frac{\theta_k - \Lambda_k / 2}{\exp(\beta_k / 2)}\right), \quad Sp_k = \Phi\left(\frac{\theta_k + \Lambda_k / 2}{\exp(-\beta_k / 2)}\right) \quad (10)$$

respectively. In each MCMC iteration, draws of Π , Se_k , and Sp_k are calculated from the MCMC draw using Equation (10). We use medians and equal-tailed credible intervals (CIs) of these posterior samples to make inferences from the HSROC-NMADT model.

Other accuracy indices can also be computed from the MCMC samples. A diagnostic test's PPV and NPV are defined as the proportions of the test's positive and negative results that are in fact true positives and true negatives and can be written as $PPV_k = Se_k \Pi / (Se_k + (1 - Sp_k)(1 - \Pi))$ and $NPV_k = Sp_k(1 - \Pi) / ((1 - Se_k)\Pi + Sp_k(1 - \Pi))$. Appendix D gives equations for the population-averaged test accuracy indices $E(Se_{ik})$ and $E(Sp_{ik})$. All these quantities can be calculated for each MCMC iteration and summarized to estimate posterior quantities. The SROC plot is computed from posterior samples of $\boldsymbol{\Lambda}$, $\boldsymbol{\Theta}$ and $\boldsymbol{\beta}$. As in Rutter and Gatsonis (2001), we can use Equation (10) to express a test's model-based sensitivity in terms of its specificity:

$$Se_k = \Phi\left\{-\frac{\Phi^{-1}(Sp_k)\exp(-\beta_k / 2) - \Lambda_k}{\exp(\beta_k / 2)}\right\}. \quad (11)$$

To derive a test's SROC curve, we first obtain posterior samples of its sensitivity by calculating Equation (11) for each MCMC iteration for a series of Sp_k 's, then plot the posterior median of sensitivity against 1 - specificity. The posterior samples also give a 95% pointwise credible band for the estimated SROC curve to show uncertainty of the SROC curve. Note that θ_k is eliminated to obtain Equation (11). Implicit in this construction is that θ_k is selected to achieve given Sp_k and corresponding Se_k .

3.2.3 The Conditional Independence Assumption—Derivation of the likelihood function (8) relies on the conditional independence assumption. This subsection discusses some implications of this assumption.

Like all other meta-analyses with multiple outcomes, the correlation between tests arises from two sources, between-study correlation and within-study correlation. Between-study correlation describes how test accuracies are correlated across studies because of study-level

characteristics such as study-specific cutoff values or aggregated covariates such as average age. Between-study correlation is modeled by Equation (3). Within-study correlation is association between accuracies of different tests arising from patient-level characteristics. The conditional independence assumption means that *given true disease status*, within-study correlations between test results are zero. However, test results are still correlated within studies *unconditionally*.

The conditional independence assumption is questionable when the correlation between test outcomes on a subject cannot be fully explained by the binary disease status. Section 4.3 shows one way to do a sensitivity analysis for this assumption.

4 The Case Study

This section illustrates our approach by analyzing the dataset from the motivating study. The model described in previous sections allows general specifications for the covariance matrices in Equation (3); we also show how to incorporate extra information to simplify the model.

The 12 included studies have sample sizes ranging from 14 to 171. Cross-tabular data for each study, classified by all the available test results, is available in the Web Appendix. All studies used the same nominal cutoff for the D-dimer test but substantial variation in *effective* cutoff values can still be present due to differences between studies in instruments and readers, so the random-effect model described in previous sections should still be considered, although Σ_{Θ} 's structure may be simplified. We consider the fits of 6 models for the cutoff values in the proposed HSROC-NMADT model, specifically:

- Model 1: The cutoff value for the latent probit score of D-dimer is the same for all studies, i.e., $\theta_{i1} = \Theta_1$ for all i . The cutoff values for the latent probit scores of prevalence and ultrasound are still treated as random effects and are allowed to be correlated.
- Model 2: The cutoff value for the latent probit scores of all index tests is the same for all studies, i.e., $\theta_{ik} = \Theta_k$ for all i and k .
- Model 3: The cutoff values for the latent probit scores of D-dimer, ultrasound, and prevalence (i.e., the gold standard test) are all random effects but are independent of each other, i.e., Σ_{Θ} 's off-diagonal elements are all 0.
- Model 4: Disease prevalence is independent of the test accuracy indices, i.e., in Σ_{Θ} , the elements σ_{12} , σ_{13} , σ_{21} , and σ_{31} are all 0.
- Model 5: The study-specific cutoff value θ_{i1} for the latent probit score of D-dimer is independent from those of all other tests, i.e., in Σ_{Θ} , the elements σ_{12} , σ_{23} , σ_{21} , and σ_{32} are all 0.
- Model 6: The full model with unstructured covariance for cutoff values as in Section 3.1.2.

In all models, we allow the different tests to have different random-effect variances. We assign slightly informative priors to the variance parameters but vague priors to other model

parameters. For instance, for the full model (Model 6), the prior distributions for the unknown parameters are $\Lambda_k \sim N(0, 100^2)$, $\Theta_m \sim N(0, 100^2)$, $\log(s_m) \sim N(0, 0.8)$, $\log(p_k) \sim N(0, 0.8)$, $R \sim IW(\mathbf{I}_3, 4)$, $\Omega \sim IW(\mathbf{I}_2, 3)$, for $m = 0, 1, 2$ and $k = 1, 2$. These specifications correspond to a 95% prior CI of (0.120, 9.375) for the standard deviation parameters and a 95% prior CI of (-0.955, 0.948) for the correlation parameters (computed by simulation). Appendix B gives details of the prior specifications of all other models.

MCMC algorithms were implemented using JAGS (Plummer 2003). For each model, we ran five independent MCMC chains with over-dispersed starting values. Convergence to the stationary distribution was assessed using trace plots, sample autocorrelation, and the Gelman-Rubin statistic (Gelman and Rubin 1992). We discarded 5,000 burn-in samples and kept 1,000,000 posterior samples from each chain. Markov Chain standard error is 0.001 for the specificity of ultrasonography and otherwise affects the 4th significant digit (Brooks et al. 2011). Models were compared using the Deviance Information Criterion (DIC) (Spiegelhalter et al. 2002). Appendix B gives results for all six models. Models 1 and 2 have DIC clearly worse than Models 3-6, which have DICs in a range of 0.65 DIC units and give similar estimates and intervals (Appendix B). To incorporate all sources of correlation across studies, we present results and sensitivity analyses based on the full model (Model 6), which has the most flexible covariance structure.

For Model 6, the posterior medians (95% equal-tailed credible intervals (CIs)) of sensitivity and specificity for the D-dimer test using the SL assay are 0.86 (0.67, 0.99) and 0.88 (0.69, 1.00) respectively. For ultrasonography, the posterior medians (95% CI) of sensitivity and specificity are 0.96 (0.80, 1.00) and 0.82 (0.39, 1.00) respectively. The overall prevalence of DVT in this collection of studies is estimated to be 0.43 (0.37, 0.49). Table 1 gives estimates of other diagnostic indices, Figure 1 shows forest plots of study-specific results, and Figure 2 shows estimated SROC curves for the index tests.

The proposed model's goodness of fit can be assessed using the posterior predictive method (Gelman et al. 1996). We use the Chi-square discrepancy statistic:

$$D^2 = \sum_i \sum_{\mathcal{S}_i} \frac{\left(n_{it_0 t_1 t_2 \dots t_K} - E(n_{it_0 t_1 t_2 \dots t_K} \mid \text{model, data}) \right)^2}{E(n_{it_0 t_1 t_2 \dots t_K} \mid \text{model, data})}, \quad (12)$$

Similar to the notation in Section 3.2.1, $\mathcal{S}_i$ is the set of distinct possible test results, with members having the form $\{T_0 = t_0, T_1 = t_1, \dots, T_K = t_K\}$ for $t_k \in \{0, 1, m\}$ indicating test k has a positive, negative, or unobserved result respectively, and $n_{it_0 t_1 t_2 \dots t_K}$ is the count of subjects having results equal to $\{T_0 = t_0, T_1 = t_1, \dots, T_K = t_K\}$. D^2 has null distribution χ_{df}^2 , where df depends on the designs of included studies as in Section 3.2.1. To assess the model's goodness of fit, we can calculate the posterior predictive p-value $p = \int \left(\chi_{df}^2 \geq D^2 \mid \xi \right) Pr(\xi \mid y) d\xi$. This goodness of fit test requires the $(K_j + 1)$ -way cross-tabulation for a study including the gold standard test, and the K_j -way cross-tabulation for a study omitting the gold standard test. In the case study, $K_j = 1$ for all studies that include the

gold standard test. The posterior predictive p-value of the full model is 0.44, i.e., no indication of lack of fit.

The proposed HSROC-NMADT model makes some unverifiable assumptions. The following sections show sensitivity analyses evaluating their impact on the results. We did sensitivity analyses on the full model (Model 6); analyses for simpler models are similar.

4.1 Sensitivity to prior distributions

This section evaluates the impact of prior distributions on the results. We consider two alternative priors for $\log(P)$ and $\log(S)$, $N(0, 0.2)$ and $N(0, 2)$, which give 95% prior CIs of (0.246, 5.086) and (0.046, 21.514) for standard deviations and 95% prior CIs of (−0.951, 0.944) and (−0.948, 0.951) for correlations, respectively. Table 2 shows the results for Model 6 under these alternatives. The posterior medians are not sensitive to these prior choices, nor is the credible interval for overall prevalence. For sensitivity and specificity, the prior with larger variance gives credible intervals that include lower values. Otherwise, inferences are driven mostly by the data, not the prior.

We also consider a logistic distribution for the latent variable Z_{ijk} instead of a normal distribution. The estimates and 95% CIs are very close to the results in Table 1 (data not shown). Therefore, the analysis is reasonably robust to these specification choices.

4.2 Sensitivity to the MAR assumption

Tests may be missing from studies not at random (MNAR) if, for example, clinicians or researchers include tests with better performance based on their belief and experience. The joint distribution of the observed data $(y^o, \cdot)$ is $f(y^o, \cdot | \xi, \gamma) = \int f(y^o, y^m, \gamma) \times f(y^o, y^m | \xi) d\gamma^m$. Under the MNAR assumption, the missing-data mechanism is non-ignorable and a model $f(y^o, y^m, \gamma)$ must be specified to make inferences about ξ . This section assumes the gold standard test is missing at random and that missingness of an index test is related only to its own accuracy indices. Other missingness mechanisms could be considered in a similar manner. We assume $\delta_{ik} \sim \text{Bernoulli}(1 - p_{ik})$ with tests independent given p_{ik} , i.e., test T_k in study i is missing with probability p_{ik} . For the present purpose, we use the following model for p_{ik} and thus for missingness:

$$\text{logit}(p_{ik}) = \gamma_{0k} + \gamma_{1k} \times \text{logit}(Se_{ik}) + \gamma_{2k} \times \text{logit}(Sp_{ik}), \quad k = 1, 2, 3, \dots, K, \quad (13)$$

where γ_{0k} is the logit of the probability T_k is missing when $Se_{ik} = Sp_{ik} = 0.5$, and γ_{1k} (γ_{2k}) describes the strength of association between missingness and the study-specific sensitivity (specificity), with $\gamma_{1k} = 0$ ($\gamma_{2k} = 0$) indicating T_k is MAR with respect to its sensitivity (specificity). The joint posterior under MNAR is (Appendix B gives details): Let $\gamma_1 = (\gamma_{11}, \dots, \gamma_{1K})$ and $\gamma_2 = (\gamma_{21}, \dots, \gamma_{2K})$. Without external information, γ_1 and γ_2 are only weakly identified by the data (Appendix B shows this for the case study). Thus, for the purpose of this sensitivity analysis, we simply specify values for γ_1 and γ_2 to see how estimates of sensitivity and specificity are affected. We assume γ_{1k} and γ_{2k} are non-positive based on the belief that outcomes tend to be missing for tests with low accuracy. To study the impact of MNAR, we consider five missingness mechanisms (sets of values of γ_1 and γ_2): (1)

missingness is related to test sensitivity only; (2) missingness is related to test specificity only; (3) missingness of D-dimer is related to D-dimer results only; (4) missingness of ultrasonography is related to ultrasonography only; and (5) missingness is related to sensitivities and specificities of both tests. These five situations correspond to these γ_1 and γ_2 values: (1) $\gamma_1 = (-a, -a)$, $\gamma_2 = (0, 0)$; (2) $\gamma_1 = (0, 0)$, $\gamma_2 = (-a, -a)$; (3) $\gamma_1 = (-a, 0)$, $\gamma_2 = (-a, 0)$; (4) $\gamma_1 = (0, -a)$, $\gamma_2 = (0, -a)$; and (5) $\gamma_1 = (-a, -a)$, $\gamma_2 = (-a, -a)$, where $a > 0$ is a real number. For all scenarios, we assign γ_{0k} a $N(0,1)$ prior, $k = 1, 2$.

We present results for $a = 0.5, 1$, and 2.5 , which imply odds ratios of missingness of 0.61 , 0.37 , and 0.08 for a 1 unit increase in logit sensitivity or specificity. Table 3 shows the results. As expected, these missingness mechanisms have little effect on inferences for disease prevalence because the gold standard test is assumed to be MAR. When missingness is related to one of the tests, the posterior median (95% CI) of the accuracy indices for that test are lower (wider) than under MAR, while the posterior median (95% CI) of the accuracy indices for the other test are higher (narrower). When missingness is related to test sensitivity, the posterior median of sensitivity decreases and its 95% CI is notably wider than under MAR. Due to the correlations between the test indices, estimates of specificities change as well. Analogous phenomena occur when missingness is related to specificity. When missingness is related to both accuracy indices of both tests, all estimates of test accuracy indices decrease. In summary, a misspecified missingness mechanism may result in over-estimating a test's performance, but the ranking of tests by their accuracy indices is relatively stable.

In general, the mechanism of MNAR is unobservable. The sensitivity analysis above was intended to examine the risk of bias under a few MNAR mechanisms. Missingness may depend on other unobserved characteristics of the study population and even if missingness is only related to test accuracy indices, the dependency may differ from what we considered.

4.3 Sensitivity to the conditional independence assumption

This section shows an analysis of sensitivity to the conditional independence assumption similar to those of Chu, Chen and Louis (2009) and Dendukuri et al. (2012). We model conditional dependence by adding covariance terms between the sensitivities of the two index tests in study i and between their specificities. This analysis is not completely general; it applies only to meta-analyses in which each study includes no more than 2 index tests, i.e., $K_i = 2$ (K can be any number, and studies can include different index tests). Under this conditional dependence assumption, we need the full $(K_i + 1)$ -way cross-tabulation for a study that includes the gold standard test and the K_i -way cross tabulation for a study without the gold standard test.

In a study evaluating two index tests T_p and T_q and the gold standard test T_0 , each subject contributes this likelihood term:

$$\left\{ \prod_i \int \left[\prod_j (\pi_i)^{y_{ij0}} (1 - \pi_i)^{1 - y_{ij0}} \prod_{k=1}^K \left((se_{ik})^{y_{ijk}} (1 - se_{ik})^{1 - y_{ijk}} \right)^{y_{ij0}} \left((1 - sp_{ik})^{y_{ijk}} (sp_{ik})^{1 - y_{ijk}} \right)^{1 - y_{ij0}} \right]^{\delta_{i0}} \right. \\ \left. \left[\prod_j l_{ij0} \right]^{1 - \delta_{i0}} \prod_{k=1}^K (p_{ik})^{1 - \delta_{ik}} (1 - p_{ik})^{1 - \delta_{ik}} \right] d\mathbf{y}_i^m \times f(\boldsymbol{\theta}_i | \boldsymbol{\Theta}, \boldsymbol{\Sigma}_{\boldsymbol{\Theta}}) f(\boldsymbol{\alpha}_i | \boldsymbol{\Lambda}, \boldsymbol{\Sigma}_{\boldsymbol{\Lambda}}) \left| f(\boldsymbol{\Theta}; \boldsymbol{\Lambda}; P; S; R; \boldsymbol{\Omega}; \boldsymbol{\beta}; \gamma_0) \right\}$$

b_{0pqi} and b_{1pqi} are defined below. In a study evaluating only two index tests T_p and T_q , each subject contributes this likelihood term:

$$\begin{aligned} P(y_{ij0}^o, y_{ijp}^o, y_{ijq}^o | \boldsymbol{\theta}_i, \boldsymbol{\alpha}_i, \boldsymbol{\beta}) &= P(y_{ij0}^o | \boldsymbol{\theta}_i, \boldsymbol{\alpha}_i, \boldsymbol{\beta}) P(y_{ijp}^o, y_{ijq}^o | \boldsymbol{\theta}_i, \boldsymbol{\alpha}_i, \boldsymbol{\beta}, y_{ij0}^o) \\ &= (\pi_i)^{y_{ij0}^o} (1 - \pi_i)^{1 - y_{ij0}^o} \left[(se_{ip})^{y_{ijp}^o} (1 - se_{ip})^{1 - y_{ijp}^o} (se_{iq})^{y_{ijq}^o} (1 - se_{iq})^{1 - y_{ijq}^o} + (-1)^{1 - y_{ijp}^o - y_{ijq}^o} b_{1pqi} \right]^{y_{ij0}^o} \times \quad (14) \\ &\quad \left[(1 - sp_{ip})^{y_{ijp}^o} (sp_{ip})^{1 - y_{ijp}^o} (1 - sp_{iq})^{y_{ijq}^o} (sp_{iq})^{1 - y_{ijq}^o} + (-1)^{1 - y_{ijp}^o - y_{ijq}^o} b_{0pqi} \right]^{1 - y_{ij0}^o}; \end{aligned}$$

where $b_{1pqi} = \varrho_{1pq} \sqrt{se_{ip} se_{iq} (1 - se_{ip}) (1 - se_{iq})}$ and $b_{0pqi} = \varrho_{0pq} \sqrt{sp_{ip} sp_{iq} (1 - sp_{ip}) (1 - sp_{iq})}$ are the covariances between tests p and q in study i 's diseased and non-diseased subjects respectively. Although negative dependence is possible, in practice positive dependence between index tests is more plausible, i.e., $0 < b_{1pqi} \leq \min(se_{ip}, se_{iq}) - se_{ip} se_{iq}$ and $0 < b_{0pqi} \leq \min(sp_{ip}, sp_{iq}) - sp_{ip} sp_{iq}$. Thus, we assign these prior distributions to ϱ_{1pq} and ϱ_{0pq} : $\varrho_{1pq} \sim U(0, \min(\sqrt{se_{ip}(1 - se_{ip}) / ((1 - se_{ip}) se_{iq})}, \sqrt{(1 - se_{ip}) se_{iq} / (se_{ip}(1 - se_{iq}))}))$ and $\varrho_{0pq} \sim U(0, \min(\sqrt{sp_{ip}(1 - sp_{ip}) / ((1 - sp_{ip}) sp_{iq})}, \sqrt{(1 - sp_{ip}) sp_{iq} / (sp_{ip}(1 - sp_{iq}))}))$. Table 4 shows the results, which are very close to those in Table 1, so that the conditional independence assumption does not appear to be a concern in this particular example. Note that without full cross-tabulations, conditional dependence cannot be detected.

5 Simulation Studies

In this section, we fit the proposed model to simulated datasets. We assume 4 diagnostic tests are under evaluation, one gold standard test (T_0), and three index tests (T_1 , T_2 , T_3), i.e., $K = 3$. The true overall sensitivities of the three index tests are 0.90, 0.80, and 0.70 and their true overall specificities are 0.85, 0.70, and 0.75, respectively. The overall true disease prevalence is 0.55. We evaluate the proposed HSROC-NMADT method by comparing it with the HSROC meta-regression method (Rutter and Gatsonis 2001). In the latter, each patient's true disease status must be known, so only studies that included T_0 can be included in that analysis. We consider two common scenarios for which Figure 3 shows network graphs.

In scenario I, we simulate meta-analyses having 32 studies, of which only 16 include T_0 . Each index test is directly compared with T_0 in some studies, so both the HSROC-NMADT and meta-regression methods can provide inferences about all three index tests. To focus on the impacts of different covariance structures, the true β_k s are all set to 0.25 in simulating

data. We consider three covariance assumptions, as follows. (1) Independence and homogeneous variance: diagonal and off-diagonal elements of Σ_{Θ} are 0.5 and 0 respectively, while diagonal and off-diagonal elements of Σ_{Λ} are 0.25 and 0 respectively. (2) Dependence and homogeneous variance: diagonal and off-diagonal elements of Σ_{Θ} are 0.5 and 0.2 respectively, while diagonal and off-diagonal elements of Σ_{Λ} are 0.25 and 0.1 respectively. (3) Dependence and heterogeneous variance: diagonal elements are (0.25, 0.5, 0.75, 1) in Σ_{Θ} and (0.5, 0.75, 1) in Σ_{Λ} . Off-diagonal elements of these two matrices are 0.2 and 0.1, respectively.

In scenario II, we simulate meta-analyses having 28 studies, of which 12 include T_0 while the other 16 evaluate only the index tests. No study compares T_3 with T_0 , i.e., there is no direct evidence about T_3 versus T_0 . Σ_{Θ} and Σ_{Λ} are assigned using covariance assumption (2) in scenario I. We assume different β_k s for the three index tests, i.e., $\beta_k = -1, 0.25, \text{ or } 1$. In this scenario, the proposed HSROC-NMADT model provides inferences for all three index tests, while the meta-regression method can estimate the diagnostic accuracy of only T_1 and T_2 .

For each scenario, we simulated 1000 datasets assuming each study had 150 patients. For each artificial dataset, the true study-specific cutoff and accuracy values were sampled from the multivariate normal distributions described in Section 3.1.2. True study-specific sensitivities and specificities were computed using Equation (2), and each patient's test outcomes were generated using Equation (8). To implement HSROC-NMADT, we used unstructured covariance matrices for all scenarios regardless of the true data-generating covariances. To implement the meta-regression, we added study-specific covariates X_{ik} to the traditional HSROC model to indicate whether study i included T_k . Appendix C gives the priors used to analyze the simulated data.

Table 5 presents results from the simulation studies, comparing the two methods in terms of absolute bias, mean squared error (MSE), 95% CI coverage probability (CP) and relative efficiency (RE), defined as MSE from the meta-regression divided by MSE from HSROC-NMADT. The HSROC-NMADT model had nearly unbiased estimates and small MSEs under all conditions simulated, with coverage probabilities near the 95% nominal level. Under covariance assumption (1) of scenario I, both methods performed reasonably well; the meta-regression assumed the correct covariance matrices, but it did not use studies without a gold standard test, so it was less efficient than the HSROC-NMADT model. Under the second and the third covariance assumptions, the HSROC-NMADT model had less biased results and smaller MSEs than the meta-regression method, which assumed the index tests have the same variances and ignores correlation between the tests. Overall, the proposed method outperformed the HSROC meta-regression.

We would like to highlight the proposed model's advantage in situations when no study directly compares an index test with the gold standard test, as in scenario II. In this situation, existing methods cannot provide posterior distributions for the sensitivities and specificities of all index tests simultaneously. Doing a separate analysis for each index test ignores correlations between the test results and requires an assumption that tests are missing completely at random. Our method provides an effective solution to this problem.

6 Discussion

This paper proposed a Bayesian hierarchical summary receiver operating characteristic model for network meta-analysis of diagnostic tests (HSROC-NMADT), which combines studies with different designs to simultaneously compare multiple diagnostic tests under a missing data framework. It accounts for correlations among multiple tests included in the same study and for between-study heterogeneity inherent in a meta-analysis. Disease prevalence is used to estimate PPV and NPV; its potential correlation with test accuracy indices (Li et al. 2007; Chu, Nie, Cole and Poole 2009; Leeflang et al. 2009; Liu et al. 2015) is taken into account. The HSROC-NMADT model estimates prevalence along with test accuracy indices and allows dependence of study-specific sensitivities and specificities on study-specific disease prevalence. The proposed method uses direct comparisons of diagnostic tests within a study and indirect comparisons combining studies. It also allows different studies to include different subsets of the index tests and does not require each study's reference test to be a gold standard. We illustrated the method using a real example and demonstrated its flexibility in the choice of summary statistics. Finally, simulation studies showed that our method fully uses the data and is more efficient than the HSROC meta-regression method.

The one-test HSROC framework (Rutter and Gatsonis 2001) assumes the accuracy and cutoff values are two independent intrinsic characteristics that together determine the sensitivity and specificity of a test in an individual study and also determine variation between studies in test performance. To compare multiple tests, we further assume correlations between test outcomes arise from correlations between accuracy values and between cutoff values, a natural extension of the one-test HSROC approach. Liu et al. (2015) made a similar assumption to estimate test accuracy indices from studies without a gold standard test. The meta-regression method for multiple test comparison is a special case of our approach, assuming tests are independent and have homogeneous variances. The proposed approach provides a way to model the covariance structure of multiple diagnostic tests; other methods, such as multivariate generalized linear mixed effects models, may also be feasible.

We have made several assumptions. First, we assume the latent variable Z_{ijk} is normally distributed. Liu et al. (2015) extended the HSROC framework by assuming the latent variable follows the location-scale distribution. We focused on the normal distribution but other location-scale families are easily implemented in our model, e.g., the case study considered a logistic distribution in place of the normal (Section 4.2). A second assumption is that tests omitted from studies are missing at random (MAR). The case study included a sensitivity analysis of this assumption in which a study's selection of index tests depended on the tests' sensitivities and specificities. A more general approach that allows the gold standard test to be MNAR needs further investigation. The third assumption is conditional independence, i.e., a study's test results are independent given true disease status and all level I parameters. This assumption fails if, e.g., two or more index tests applied to the same patient are conditionally dependent due to a factor other than disease status, such as a biological mechanism (Vacek 1985). A sensitivity analysis to the conditional independence assumption for the case study was given in Section 4.3. Several models have been proposed

that allow conditional dependence (Dendukuri and Joseph 2001; Dendukuri et al. 2009; Xu and Craig 2009; Dendukuri et al. 2012). This paper does not extend our model to incorporate conditional dependence for general cases though our model allows cutoff and accuracy values to be correlated between tests, which induces marginal correlations among sensitivity and specificity parameters, indirectly mitigating the effect of the conditional independence assumption. It is very challenging to model conditional dependence directly in any generality in meta-analysis of multiple diagnostic tests. More work is needed to address this problem.

We have extended network meta-analysis of randomized clinical trials (NMA-CT), which is widely used to simultaneously compare multiple treatments. The proposed model relies on a consistency assumption that is still being actively researched for NMA-CT. Inconsistency occurs when direct and indirect evidence conflict and can arise from various causes, such as non-comparability of studies. Some are skeptical about the validity of combining disparate sources of evidence, e.g., the Cochrane Collaboration (Higgins and Green 2011) has warned against pooling direct and indirect evidence. MNAR, discussed in Section 4.2, is one possible cause of inconsistency. We have not provided a general solution to this problem. Several methods have been proposed for detecting inconsistency in NMA-CT (Lu and Ades 2006; Lu and Ades 2009; Higgins et al. 2012; Piepho 2014; Zhao et al. 2016); these methods have not yet been applied to the HSROC-NMADT model.

Compared to a separate meta-analysis of each diagnostic test, approaches that simultaneously compare multiple tests have potential problems, particularly when the number of tests is large. In our model, K index tests have K^2+5K+2 unknown parameters, $(K+1)^2$ of which are related to the covariance matrices. One extra index test leads to $2K+6$ more parameters, so the computational burden increases polynomially in K . Estimating covariance matrices is known to be hard when the dimension is large and the number of studies is small. We used a separation strategy for the covariance matrices (Barnard et al. 2000; O'Malley and Zaslavsky 2008), which provides more flexibility than the common inverse Wishart prior. However, this strategy can be computationally challenging for large K . Alternative parameterizations could be helpful, e.g., Daniels and Kass (1999, 2001), Gelman (2006), or the covariance matrices could be simplified, as demonstrated in the case study.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

We thank the editor Professor Joseph G. Ibrahim, an associate editor and two anonymous reviewers for many constructive comments. Research reported in this publication was supported in part by NIDCR R03 DE024750 (H.C.), NLM R21 LM012197 (H.C.), NLM R21 LM012744 (H.C., J.H.), NIDDK U01 DK106786 (H.C.), AHRQ R03HS024743 (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

Barnard J McCulloch R Meng X Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage Statistica Sinica 2000 10 1281–1311

- Brooks, S, Gelman, A, Jones, G, Meng, X-L. Handbook of Markov Chain Monte Carlo. CRC press; 2011.
- Caldwell DM Ades AE Higgins JPT Simultaneous comparison of multiple treatments: combining direct and indirect evidence *BMJ : British Medical Journal* 2005 331 7521 897–900 [PubMed: 16223826]
- 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]
- 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]
- 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]
- Dendukuri N Hadgu A Wang L Modeling conditional dependence between diagnostic tests: A multiple latent variable model *Statistics in Medicine* 2009 28 3 441–461 [PubMed: 19067379]
- Dendukuri N Joseph L Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests *Biometrics* 2001 57 1 158–67 [PubMed: 11252592]
- 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]
- Dukic V Gatsonis C Meta-analysis of Diagnostic Test Accuracy Assessment Studies with Varying Number of Thresholds *Biometrics* 2003 59 4 936–946 [PubMed: 14969472]
- Gelman A Meng X-L Stern H Posterior predictive assessment of model fitness via realized discrepancies *Statistica Sinica* 1996 6 4 733–760
- Gelman A Rubin DB Inference from Iterative Simulation Using Multiple Sequences *Statistical Science* 1992 7 4 457–472
- 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]
- Higgins J, Green S. *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 [updated March 2011]. 2011
- Higgins JPT Jackson D Barrett JK Lu G Ades AE White IR Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies *Research Synthesis Methods* 2012 3 2 98–110 [PubMed: 26062084]
- Higgins JP White IR Wood AM Imputation methods for missing outcome data in meta-analysis of clinical trials *Clinical Trials* 2008 5 3 225–239 [PubMed: 18559412]
- Ibrahim JG Molenberghs G Missing data methods in longitudinal studies: a review *Test* 2009 18 1 1–43 [PubMed: 21218187]
- Irwig L Macaskill P Glasziou P Fahey M Meta-analytic methods for diagnostic test accuracy *Journal of Clinical Epidemiology* 1995 48 1 119–130 [PubMed: 7853038]
- Kyrle PA Eichinger S Deep vein thrombosis *The Lancet* 2005 365 9465 1163–1174
- Leeftang 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]
- Li J Fine JP Safdar N Prevalence-dependent diagnostic accuracy measures *Statistics in Medicine* 2007 26 17 3258–3273 [PubMed: 17212380]
- Little, RJA, Rubin, DB. 2nd Edition. John Wiley & Sons; New Jersey: 2002. *Statistical Analysis with Missing Data*.
- Liu Y Chen Y Chu H A unification of models for meta-analysis of diagnostic accuracy studies without a gold standard *Biometrics* 2015 71 2 538–547 [PubMed: 25358907]
- Lu G Ades AE Assessing Evidence Inconsistency in Mixed Treatment Comparisons *Journal of the American Statistical Association* 2006 101 474 447–459

- Lu G Ades AE Modeling between-trial variance structure in mixed treatment comparisons *Biostatistics* 2009 10 4 792–805 [PubMed: 19687150]
- Macaskill P, Gatsonis C, Deeks J, Harbord R, Takwoingi Y. “Chapter 10: Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis C (editors), *Cochrane. Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0*. The Cochrane Collaboration. 2010
- Mills EJ Ioannidis JP Thorlund K Schünemann HJ Puhan MA Guyatt GH How to use an article reporting a multiple treatment comparison meta-analysis *The Journal of the American Medical Association* 2012 308 12 1246–1253 [PubMed: 23011714]
- 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]
- O’Malley AJ Zaslavsky AM Domain-Level Covariance Analysis for Multilevel Survey Data with Structured Nonresponse *Journal of the American Statistical Association* 2008 103 484 1405–1418
- Perone N Bounameaux H Perrier A Comparison of four strategies for diagnosing deep vein thrombosis: a cost-effectiveness analysis *The American Journal of Medicine* 2001 110 1 33–40 [PubMed: 11152863]
- Piepho HP. Network-meta analysis made easy: detection of inconsistency using factorial analysis-of-variance models. *BMC Medical Research Methodology*. 2014; 14:61. [PubMed: 24885590]
- Plummer, M. Proceedings of the 3rd international workshop on distributed statistical computing. Vol. 124. Vienna, Austria: 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling; 125
- Plummer M. *rjags: Bayesian Graphical Models using MCMC*. 2016
- Reitsma JB Glas AS Rutjes AWS Scholten RJPM 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]
- Rutter CM Gatsonis CA A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations *Statistics in Medicine* 2001 20 19 2865–1884 [PubMed: 11568945]
- Sadatsafavi M Shahidi N Marra F FitzGerald MJ Elwood KR Guo N Marra CA 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]
- Scarvelis D Wells PS Diagnosis and treatment of deep-vein thrombosis *Canadian Medical Association Journal* 2006 175 9 1087–1092 [PubMed: 17060659]
- Spiegelhalter DJ Best NG Carlin BP Van Der Linde A Bayesian measures of model complexity and fit *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002 64 4 583–639
- Takwoingi Y Leeftang MM Deeks JJ Empirical evidence of the importance of comparative studies of diagnostic test accuracy *Annals of Internal Medicine* 2013 158 7 544–554 [PubMed: 23546566]
- Tovey C Wyatt S Diagnosis, investigation, and management of deep vein thrombosis *BMJ: British Medical Journal* 2003 326 7400 1180–1184 [PubMed: 12775619]
- Trikalinos TA, Hoaglin DC, Small KM, Terrin N, Schmid CH. Methods for the joint meta-analysis of multiple tests. *Research Synthesis Methods*. 2014; 5(4)
- Vacek P The effect of conditional dependence on the evaluation of diagnostic tests *Biometrics* 1985 41 4 959–968 [PubMed: 3830260]
- 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]
- Walter S Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data *Statistics in Medicine* 2002 21 9 1237–1256 [PubMed: 12111876]
- Walter SD Irwig L Glasziou PP Meta-Analysis of Diagnostic Tests With Imperfect Reference Standards *Journal of Clinical Epidemiology* 1999 52 10 943–951 [PubMed: 10513757]
- Wells PS Anderson DR Rodger M Forgie M Kearon C Dreyer J Kovacs G Mitchell M Lewandowski B Kovacs MJ Evaluation of D-Dimer in the Diagnosis of Suspected Deep-Vein Thrombosis *New England Journal of Medicine* 2003 349 1227–1235 [PubMed: 14507948]

- Xu H Craig BA A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests *Biometrics* 2009 65 4 1145–1155 [PubMed: 19210729]
- Zhang J Carlin B Neaton J Soon G Nie L Kane R Virnig B Chu H Network meta-analysis of randomized clinical trials: Reporting the proper summaries *Clinical Trials* 2014 11 2 246–262 [PubMed: 24096635]
- Zhao H Hodges JS Ma H Jiang Q Carlin BP Hierarchical Bayesian approaches for detecting inconsistency in network meta-analysis *Statistics in Medicine* 2016 35 20 3524–3536 [PubMed: 27037506]

Figure 1.

Forest Plots and Contour Plots for D-dimer and Ultrasound: Panels (a), (b) are forest plots of sensitivity and specificity of D-dimer, respectively. Panels (d), (e) are forest plots of sensitivity and specificity of ultrasound, respectively. Circles are study-specific posterior medians; solid and dashed lines denote 95% credible intervals when the test was included and not included in the study respectively (with the latter imputed by MCMC sampling). Panels (c), (f) show quantile contours of posterior false positive rate versus true positive rate at quantile levels 0.25, 0.5, 0.75, 0.90 and 0.95 for D-dimer and ultrasound, respectively.

Figure 2.

Estimated SROC curves, TP and FP rates, and pooled TP and FP rates; pooled TP = posterior median of sensitivity, pooled FP = 1 - posterior median of specificity.

Figure 3.

Simulation setups. Each circle represents a test, with its size reflecting the number of studies that included the test. Solid lines connecting the circles represent direct comparisons, with width proportional to the number of studies.

Table 1:
HSROC-NMADT model (Model 6): posterior medians and 95% credible intervals.

Prevalence		0.43 (0.37,0.49)	
	D-dimer	Ultrasonography	
Sensitivity	0.86 (0.67,0.99)	0.96	(0.80,1.00)
Specificity	0.88 (0.69,1.00)	0.82	(0.39,1.00)
PPV	0.84 (0.64,0.99)	0.80	(0.51, 1.00)
NPV	0.89 (0.76, 0.99)	0.96	(0.78, 1.00)

Table 2:Sensitivity to the prior distributions of P and S .

	Prior	$N(0, 0.2)$	$N(0, 0.8)$	$N(0, 2)$
Prevalence	Posterior median	0.43	0.43	0.43
	95% CI	(0.36, 0.51)	(0.37, 0.49)	(0.37, 0.49)
D-dimer				
Sensitivity	Posterior median	0.86	0.86	0.86
	95% CI	(0.67, 0.99)	(0.67, 0.99)	(0.57, 0.99)
Specificity	Posterior median	0.88	0.88	0.87
	95% CI	(0.68, 1.00)	(0.69, 1.00)	(0.59, 1.00)
Ultrasonography				
Sensitivity	Posterior median	0.95	0.96	0.96
	95% CI	(0.80, 1.00)	(0.80, 1.00)	(0.61, 1.00)
Specificity	Posterior median	0.82	0.82	0.83
	95% CI	(0.40, 1.00)	(0.39, 1.00)	(0.30, 1.00)

Table 3:
Meta-analysis of DVT tests: Estimates and credible intervals under different missingness assumptions. Missing Mechanism (MM) = “At Random” is equivalent to MAR; Missing Mechanism = “All” stands for the situation when the missingness is related to sensitivities and specificities of both tests.

Missing Mechanism	γ_{11}	γ_{12}	γ_{21}	γ_{22}	Prevalence	Se: D-dimer	Se: ultrasonography	Sp: D-dimer	Sp: ultrasonography
At Random	0	0	0	0	0.43 (0.37,0.49)	0.86 (0.67,0.99)	0.96 (0.80,1.00)	0.88 (0.69,1.00)	0.82 (0.39,1.00)
D-dimer	-0.5	-0.5	0	0	0.43 (0.37,0.49)	0.83 (0.46,0.99)	0.96 (0.84,1.00)	0.84 (0.45,0.99)	0.83 (0.50,1.00)
ultrasonography	0	0	-0.5	-0.5	0.43 (0.37,0.49)	0.88 (0.73,0.99)	0.92 (0.29,1.00)	0.89 (0.75,1.00)	0.63 (0.02,0.99)
Se	-0.5	0	-0.5	0	0.43 (0.37,0.49)	0.86 (0.59,0.99)	0.93 (0.16,1.00)	0.88 (0.73,0.99)	0.76 (0.22,1.00)
Sp	0	-0.5	0	-0.5	0.43 (0.37,0.50)	0.86 (0.70,0.99)	0.95 (0.76,1.00)	0.87 (0.57,1.00)	0.71 (0.06,1.00)
All	-0.5	-0.5	-0.5	-0.5	0.43 (0.37,0.50)	0.86 (0.64,0.98)	0.94 (0.43,1.00)	0.87 (0.63,0.99)	0.70 (0.04,1.00)
D-dimer	-1	-1	0	0	0.43 (0.37,0.50)	0.81 (0.43,0.99)	0.97 (0.84,1.00)	0.83 (0.40,0.99)	0.84 (0.53,1.00)
ultrasonography	0	0	-1	-1	0.43 (0.37,0.49)	0.88 (0.74,1.00)	0.91 (0.32,1.00)	0.90 (0.75,1.00)	0.59 (0.02,0.99)
Se	-1	0	-1	0	0.43 (0.37,0.49)	0.84 (0.53,0.99)	0.91 (0.13,1.00)	0.89 (0.74,0.99)	0.77 (0.24,1.00)
Sp	0	-1	0	-1	0.43 (0.37,0.50)	0.87 (0.71,0.99)	0.96 (0.78,1.00)	0.85 (0.52,0.99)	0.67 (0.09,0.99)
All	-1	-1	-1	-1	0.43 (0.37,0.50)	0.84 (0.51,0.99)	0.93 (0.36,1.00)	0.86 (0.50,1.00)	0.67 (0.03,1.00)
D-dimer	-2.5	-2.5	0	0	0.43 (0.37,0.50)	0.80 (0.39,0.99)	0.97 (0.85,1.00)	0.81 (0.39,0.99)	0.84 (0.55,1.00)
ultrasonography	0	0	-2.5	-2.5	0.43 (0.37,0.49)	0.88 (0.74,1.00)	0.90 (0.30,1.00)	0.90 (0.76,1.00)	0.58 (0.02,0.99)
Se	-2.5	0	-2.5	0	0.43 (0.37,0.50)	0.80 (0.46,0.99)	0.86 (0.10,1.00)	0.90 (0.75,1.00)	0.80 (0.31,1.00)
Sp	0	-2.5	0	-2.5	0.43 (0.37,0.49)	0.88 (0.71,0.99)	0.96 (0.82,1.00)	0.81 (0.43,0.99)	0.64 (0.12,0.99)
All	-2.5	-2.5	-2.5	-2.5	0.43 (0.37,0.50)	0.83 (0.37,1.00)	0.92 (0.30,1.00)	0.85 (0.30,1.00)	0.63 (0.01,1.00)

Table 4:
Sensitivity analysis of the conditional independence assumption

Prevalence		0.43 (0.37,0.50)
	D-dimer	Ultrasonography
Sensitivity	0.85 (0.68,0.99)	0.96 (0.79,1.00)
Specilicity	0.87 (0.68,0.99)	0.79 (0.42,1.00)

Table 5:
Simulation results: absolute biases, MSEs and CPs of sensitivities (Se_k), specilcities (Sp_k), and overall prevalence (Π). Estimates from HSROC-NMADT model using all the data are compared to those from HSROC meta-regression model using partial data.

	HSROC-NMADT			HSROC meta-regression			RE ^a
Parameter (true)	Bias	MSE	CP	Bias	MSE	CP	
Scenario I							
Covariance assumption (1)							
Π (0.55)	0.003	0.004	0.940	-	-	-	-
Se_1 (0.9)	-0.004	0.001	0.942	0.005	0.002	0.967	1.360
Se_2 (0.8)	0.000	0.002	0.953	0.008	0.003	0.960	1.734
Se_3 (0.7)	0.003	0.005	0.944	0.003	0.016	0.991	3.413
Sp_1 (0.85)	-0.003	0.004	0.949	0.002	0.006	0.973	1.704
Sp_2 (0.7)	-0.002	0.004	0.955	0.005	0.008	0.974	1.855
Sp_3 (0.75)	0.001	0.007	0.945	-0.004	0.024	0.987	3.642
Covariance assumption (2)							
Π (0.55)	0.002	0.003	0.949	-	-	-	-
Se_1 (0.9)	-0.001	0.001	0.959	0.004	0.002	0.965	1.879
Se_2 (0.8)	0.001	0.002	0.950	0.009	0.003	0.959	1.879
Se_3 (0.7)	0.008	0.004	0.952	0.005	0.016	0.989	4.028
Sp_1 (0.85)	-0.000	0.003	0.953	0.004	0.008	0.967	1.910
Sp_2 (0.7)	-0.000	0.004	0.954	0.004	0.008	0.967	1.957
Sp_3 (0.75)	0.008	0.006	0.938	-0.005	0.022	0.993	3.745
Covariance assumption (3)							
Π (0.55)	0.001	0.002	0.956	-	-	-	-
Se_1 (0.9)	-0.002	0.001	0.942	0.005	0.002	0.965	1.644
Se_2 (0.8)	-0.004	0.003	0.950	0.011	0.005	0.967	1.734
Se_3 (0.7)	0.004	0.009	0.938	0.008	0.030	0.989	3.284
Sp_1 (0.85)	-0.008	0.004	0.952	0.005	0.006	0.971	1.754
Sp_2 (0.7)	-0.000	0.007	0.945	0.006	0.015	0.965	1.962
Sp_3 (0.75)	0.004	0.013	0.940	-0.024	0.051	0.989	3.959
Scenario II							
Π (0.55)	0.011	0.005	0.946	-	-	-	-
Se_1 (0.9)	-0.007	0.004	0.946	0.002	0.007	0.976	1.705
Se_2 (0.8)	-0.002	0.002	0.946	0.006	0.005	0.976	2.150
Se_3 (0.7)	0.013	0.004	0.939	-	-	-	-
Sp_1 (0.85)	-0.003	0.001	0.957	0.005	0.002	0.970	1.624
Sp_2 (0.7)	0.002	0.005	0.954	0.002	0.013	0.976	2.486
Sp_3 (0.75)	0.042	0.022	0.928	-	-	-	-

^a“RE” denotes the relative efficiency of the estimate from HSROC-NMADT to that from HSROC meta-regression.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript