Statistical Methods for Multivariate Meta-Analysis of Diagnostic Tests

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Xiaoye Ma

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Advised by Prof. Haitao Chu

May, 2015

Acknowledgements

I would like to express the deepest appreciation to my advisor, Dr. Haitao Chu, for his excellent guidance and persistent support throughout my master and doctoral study, for his patience, enthusiasm, and immense knowledge. I am extremely grateful that Dr. Haitao Chu agreed to be my advisor since my Master's study, when I was enlightened at the first glance of research. My journey of Ph.D. study is challenging, but with his persistent support and sincere advice, it turned out to be extremely rewarding and memorable. I am truly thankful to have Dr.Bradley P. Carlin be my committee chair. His intelligence and expertise has invaluable input in my dissertaion. I am also grateful that Dr.Bradley P. Carlin referred me to Dr.David Ohlssen at Novartis for a summer internship position. Without this experience I would not be able to find a full-time job as smoothly. Many thanks also go to my committee members Dr.Wei Pan and Dr.Richard MacLehose, for their friendly guidance, thought-provoking suggestions, and invaluable input. Thanks go to Dr.Muhammad Fareed K. Suri, who got me involved in his interesting research project. Thanks to all professors who taught me classes in my graduate study and all my classmates and officemates who shared ideas, discussions and help with each other's difficulties.

I also want to send my thanks to my family overseas. Thank you mom and dad for sending me abroad to study, trusting me to pursue my dream, which opened a gate to an incredibly amazing life. Thank you grandpas and grandmas, in China and in heaven, for your endless care, and your lovely effort to try to remember name of my school. At last, special thanks to my husband, Chuan Shi. Thank you for always staying by my side, facing all difficulties together and helping me bulid my confidence. I can always feel your love and support no matter how far apart we are.

Dedication

To my husband, Chuan Shi, and my parents Lin Ma and Xun Sun, for their infinite love, trust and support.

Abstract

Accurate diagnosis is often the first step towards the treatment and prevention of disease. Many quantitative comparisons of diagnostic tests have relied on meta-analyses, which are statistical methods to synthesize all available information in various clinical studies. In addition, in order to effectively compare the growing number of diagnostic tests for a specific disease, innovative and efficient statistical methods to simultaneously compare multiple diagnostic tests are urgently needed for physicians and patients to make better decisions.

In the literature of meta-analysis of diagnostic tests (MA-DT), discussions have been focused on statistical models under two scenarios: (1) when the reference test can be considered a gold standard, and (2) when the reference test cannot be considered a gold standard. We present an overview of statistical methods for MA-DT in both scenarios. This dissertation covers both conventional and advanced multivariate approaches for the first scenario, and a latent class random effects model when the reference test itself is imperfect.

As study design and populations vary, the definition of disease status or severity could differ across studies. A trivariate generalized linear mixed model (TGLMM) has been proposed to account for this situation; however, its application is limited to cohort studies. In practice, meta-analytic data is often a mixture of cohort and case-control studies. In addition, some diagnostic accuracy studies only select a subset of samples to be verified by the reference test, which is known as potential source of partial verification bias in single studies. The impact of this bias on a meta-analysis has not been investigated. We propose a novel hybrid Bayesian hierarchical model to combine cohort and case-control studies, and correct partial verification bias at the same time.

A recent paper proposed an intent-to-diagnose approach to handle non-evaluable index test results, and discussed several alternative approaches. However, no simulation studies have been conducted to test the performance of the methods. We propose an extended TGLMM to handle non-evaluable index test results, and examine the performance of the intent-to-diagnose approach, the alternative approaches, and the proposed approach by extensive simulation studies.

To compare the accuracy of multiple tests in a single study, three designs are commonly used: 1) the multiple test comparison design; 2) the randomized design; and 3) the non-comparative design. Existing MA-DT methods 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 an efficient and flexible meta-analysis framework to combine all designs for simultaneous inference. We develop a missing data framework and a Bayesian hierarchical model for network meta-analysis of diagnostic tests (NMA-DT), and offer key advantages over traditional MA-DT methods.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Accurate diagnosis of a disease is often the first step toward its treatment and prevention. The performance of a binary test of interest (candidate or index test) is commonly compared to a reference test (preferably a "gold standard" test), then measured by a pair of indices such as sensitivity (Se) and specificity (Sp). Sensitivity is defined as the probability of testing positive given a person is diseased, and specificity is defined as the probability of testing negative given a person is disease-free[1]. For a toy example, ultrasound is the candidate or index test in diagnosing rotator cuff tears, and the gold standard test is arthroscopic surgery. In one study, to estimate the Se and Sp of ultrasound, a group of participants are tested by both ultrasound and the gold standard, and test outcomes are compared in a cross-tabulated 2 × 2 table (Table 1.1). In this study, Se is estimated as $Pr(\text{Ultrasound} = +|\text{Diseased}) = 80/100 = 0.8$ and Sp is estimated as $Pr(\text{Ultrasound} = -|\text{Non-diseased}) = 180/200 = 0.9$. Other frequently used indices include positive and negative predictive values (PPV and NPV), and positive and negative diagnostic likelihood ratios (LR+ and LR$-$). PPV is defined as the probability of being diseased given a positive index test result, and NPV is defined as the probability of being disease-free given a negative index test result. In this example, PPV is estimated as $Pr(\text{Diseased}|\text{Ultrasound} = +) = 80/100 = 0.8$ and NPV is estimated as $Pr(\text{Non-diseased}|\text{Ultrasound} = -) = 180/200 = 0.9$.

The growing number of assessment instruments, as well as a rapid escalation in trial costs, has generated an increasing need for scientifically rigorous comparisons of the diagnostic tests in clinical practice. *Meta-analysis of diagnostic test (MA-DT)* is a useful

Table 1.1: 2 × 2 table for a toy example

Ultrasound	Arthroscopic + (Diseased)	 − (Non-diseased)	Total
+	80	20	100
−	20	180	200
Total	100	200	300

tool to combine evidence on diagnostic accuracies from multiple studies. Compared to conventional meta-analyses of controlled clinical trials, it has several additional statistical challenges that have been extensively studied in the literature, such as correlation between test accuracy indices and heterogeneity of test performance across studies. Other important topics in MA-DT, such as partial verification and mixture of study designs remain challenging.

The increasing number of available diagnostic instruments for a disease condition has generated a need to develop an efficient and flexible meta-analysis framework for simultaneous inference. As a result, in order to effectively compare multiple diagnostic tests, extending MA-DT from studying the performance of a single test, to enabling simultaneous comparison of multiple test performance by a framework of *network meta-analysis of diagnostic tests (NMA-DT)*, is urgently needed for physicians and patients to make better decisions in selecting tests.

1.1 Current development and challenges in meta-analysis of diagnostic tests (MA-DT)

1.1.1 Literature review in MA-DT

In MA-DT, there is a great potential for heterogeneity due to differences in such things as disease prevalence, study population characteristics, laboratory methods, and study designs. While some study level covariates such as the mean age may explain some of the variability, random effects models are commonly recommended to account for other unobserved sources of variation. When a reference test can be considered as a gold standard, several meta-analysis methods are available to account for this heterogeneity[2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. Specifically, random effects models, including the hierarchical summary receiver operating characteristic model [2] and bivariate random effects meta-analysis on sensitivities and specificities[4, 10, 11], which are identical in some situations, have been recommended [5, 8, 12]. Indeed, extensive examples and simulations demonstrated that bivariate random-effects meta-analysis offers numerous advantages over separate univariate meta-analysis [13, 14]. In general, generalized linear mixed models (GLMM), which use the exact binomial likelihood, often perform better than the linear mixed models which use a normal approximation[10, 15]. In addition, a trivariate GLMM (TGLMM) was also proposed to jointly model the disease prevalence, sensitivities and specificities[16].

In practice, disease status is often measured by a reference test that is subject to nontrivial measurement error. This leads to a setting "without a gold standard". When the reference test is subject to measurement error, the evaluation of diagnostic tests in a meta-analysis setting becomes more challenging. Only a few articles have described meta-analysis models for diagnostic tests in the absence of a gold standard. Walter et al.[17] discussed a latent class model for a meta-analysis of two diagnostic tests assuming varying prevalence, but constant sensitivity and specificity across studies. A more general latent class random effects model by Chu et al.[18] assumes sensitivity and specificity of both tests as well as prevalence to be random effects. Sadatsafavi et al.[19] presented a model where conditional dependence between tests is allowed but other than prevalence, only one of the sensitivity or specificity can be implemented as random effects. Dendukuri et al.[20] presented a Bayesian meta-analysis for the accuracy of a test for tuberculous pleuritis in the absence of a gold standard. In this thesis, we will perform a systematic review and comparison of the above mentioned methods.

1.1.2 Mixture of case-control and cohort studies

As introduced in the last section, the paired indices measuring diagnostic test performance are potentially correlated and heterogeneous across studies, such that bivariate random effects models on sensitivities and specificities have been recommended to account for such correlation and heterogeneity [5, 10, 4]. In addition, because the classification of disease status is typically based on a continuum of measurable traits, and such continuous traits not only determine disease prevalence, but also misclassification rates (subjects with true levels close to the cut-point are more likely to be misclassified), sensitivities and specificities can be correlated with study prevalences [21]. TGLMMs on prevalence, sensitivities and specificities were proposed to account for such correlations [22]. However, many meta-analyses of diagnostic tests in practice contain both cohort and case-control study designs [23]. Using cohort design, a study first tests participants with the index test, and then confirms disease status with the gold standard[24]. In case-control design studies, groups of patients with and without disease are identified before performing the index test [25]. Thus, case-control studies cannot be used to estimate disease prevalence, and direct application of the trivariate random effects models has been restricted to a meta-analysis with cohort studies only. In such situations, ignoring the information on prevalence to fit the bivariate random effects model [10, 4, 8, 11] on Se and Sp, or excluding case-control studies to fit the trivariate random effects model [22] on prevalence, can potentially lead to a substantial loss of information contained in the data. For example, the former approach ignores disease prevalence information and the correlations between disease prevalence Se and Sp, which can lead to incorrect estimation of PPV and NPV.

1.1.3 Partial verification bias

Partial verification is a common and important potential source of bias that usually arises when the selection of samples to be tested by a reference standard test is affected by the results of a diagnostic test [26, 27]. As stated in the quality assessment tool for diagnostic accuracy studies (QUADAS), partial verification bias occurs when not all of the study group receive confirmation of the diagnosis by the reference standard [28]. As an illustration, let us use the previous example in Table 1.1 and assume that true Se and Sp of ultrasound are 0.8 and 0.9, respectively and no sampling variation. However, 80% of the subjects with ultrasound positive outcomes are verified, while only 20% of the subjects with ultrasound negative outcomes are verified by the gold standard. Let n_{td} denote the number of subjects with test results $T = t$ and disease status $Dis = d$ ($t, d = 0, 1, m$ indicating negative, positive and missing results, respectively). We will have $n_{11} = 64$, $n_{01} = 4$, $n_{00} = 36$, $n_{10} = 16$, $n_{1m} = 20$ and $n_{0m} = 160$. Now, if we only use verified samples, we overestimate Se as $\widehat{Se} = n_{11}/(n_{11} + n_{01}) = 0.94$ and underestimate Sp as $\widehat{Sp} = n_{00}/(n_{00} + n_{10}) = 0.69$. Moreover, the direction and magnitude of such bias depends on selection probabilities [29]. To avoid such bias, ideally, all subjects should be verified. However, due to some practical issues such as ethical and economic considerations, partial verification is ubiquitous. In a systematic review of bias and variation in meta-analysis of diagnostic accuracy studies, 15 out of 31 (48%) meta-analyses contain at least one study with partial verification [29]. Thus, it is important to adjust for partial verification bias in meta-analysis of diagnostic tests [29, 30].

Methods to adjust for verification bias in a single study are widely published. Most of the methods are built upon the missing at random (MAR) assumption, when the decision to ascertain disease status only depends on the observed index test result, T. Violations of this condition can happen when, for example, subjects with family disease history are more likely to verify their desease status[1]. Begg and Greenes [31] proposed a simple method based on Bayes theorem. Other methods such as multiple imputation, direct maximum likelihood, or Bayesian approaches have been proposed [32, 33, 34, 35, 36, 27]. These methods give unbiased estimates of Se and Sp for individual studies, instead of recovering missing counts of subjects. Thus we would not be able to apply the exact binomial likelihood assumption for a GLMM approach under meta-analysis settings. Few sensitivity analysis methods are available under the assumption of Missing Not At Random (MNAR), i.e., the probability of being verified by a reference standard depends on the unobserved data[37, 38].

On the other hand, only limited papers are available on methods to adjust verification bias in a meta-analysis setting. De Groot et al. [39] extended the Bayes theorem method to adjusting for this bias in meta-analysis of diagnostic tests with nominal outcomes. A two-stage Bayesian approach was described, where in the 1st stage the probability distribution of the index test was calculated and in the 2nd stage PPV and NPV are calculated using observed data based on their unbiasedness property under the MAR assumption [1]. Bayes theorem is then applied to achieve pooled sensitivity and specificity estimates. A few papers have discussed the missing data problem caused by imperfect reference standards, but these papers are not aimed at partial verification problems specifically. Previously introduced papers by Chu et al. [18] and Sdatsafavi et al. [19] disscuss models for such a scenario.

1.1.4 Non-evaluable subjects

Most papers in the literature have discussed missing reference test outcomes (missing disease status) and how to correct such bias, known as partial verification bias[31, 39, 34, 26]. However, index test outcomes can be non-evaluable as well, especially for tests yielding dichotomous results, and different situations were discussed where index test result can be non-evaluable: uninterpretable, intermediate and indeterminate [40, 41].

For a single study, there are many discussions about how to deal with non-evaluable index test outcomes, such as excluding them [42], grouping them with positive or negative outcomes[42, 40], or using a 3 × 2 table to report them as an extension of the standard 2 × 2 table[42]. On the other hand, in meta-analysis, there is little discussion of how to deal with missing index test outcomes[41]. The "classic" 2 × 2 table models such as the bivariate linear mixed models[2, 11, 5, 43, 4, 8], bivariate GLMM[10, 44, 45] and TGLMM [22] ignore missing index test outcomes. Recently, a paper by Schuetz et al.[41] discussed this issue by studying different approaches dealing with index test non-evaluable subjects. The paper conducted a meta-analysis of coronary CT angiography studies and presented an intent-to-diagnose approach together with three commonly applied alternative approaches. The intent-to-diagnose approach takes non-evaluable diseased subjects as false positives and non-diseased subjects as false negatives such that sensitivity and specificity won't be overestimated. The other three alternative approaches in Schuetz et al.[41] are described in Chapter 4. The authors concluded that excluding the index test non-evaluable subjects leads to over-estimation of sensitivity and specificity and recommended the conservative intent-to-diagnose approach by treating non-evaluable diseased subjects as false negatives and non-evaluable non-diseased subjects as false positives. However, no simulation studies have been conducted to evaluate the performance of these approaches.

We can treat index test non-evaluable subjects as missing data. Schuetz et al.[41] concluded that sensitivity and specificity could be over-estimated by excluding non-evaluable subjects. In fact, under a reasonable general assumption, MAR, excluding non-evaluable subjects can provide unbiased estimates of them. A special case of MAR is missing completely at random (MCAR), where missingness is independent of both observed and unobserved variables [46]. Under MAR, T and M are independent given disease status Dis, where $M = 1, 0$ indicates missingness of index test outcome. Hence, excluding non-evaluable subjects will have unbiased estimates of Se and Sp: $\widehat{Se} = Pr(T = 1|Dis = 1, M = 0) = Pr(T = 1|Dis = 1)$ and $\widehat{Sp} = Pr(T = 0|Dis = 0, M = 0) = Pr(T = 0|Dis = 0)$. Similarly, positive and negative likelihood ratios (LR+ and LR−) and area under the curve (AUC) estimates are unbiased too. Under MCAR, $Pr(M = 1|Dis = 1) = Pr(M = 1|Dis = 0)$, and hence disease prevalence (π) estimate is also unbiased if non-evaluable subjects are excluded. However, when missing probabilities are not equal between diseased and non-diseased participants, disease prevalence estimate can be biased if non-evaluable subjects are excluded, leading overall estimates of PPV and NPV to be biased. PPV and NPV are generally preferred by clinicians as measurements of how well a test predicts true disease status because their interpretations are more intuitive: PPV is the probability that a subject with positive intex test result is truly diseased and NPV is the probability that a subject with negative intex test result is truly non-diseased[1]. However, none of the approaches discussed in Schuetz et al. [41] can correct the bias in their estimates.

1.2 Network meta-analysis of diagnostic tests (NMA-DT)

As discussed, 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. However, in practice, it is becoming common to compare multiple diagnostic tests in a meta-analysis, where studies may compare different candidate tests and some studies may not include a gold standard [47, 48, 19, 49, 50, 51]. As a consequence, existing meta-analysis methods reviewed previously are not able to effectively analyze such data.

To compare the accuracy of multiple tests in a single study, three designs are commonly used [52]: 1) the multiple test comparison design where all subjects are diagnosed by all candidate tests and verified by a gold standard; 2) the randomized design where subjects are randomly assigned to one of the candidate tests, and all subjects are verified by a gold standard; and 3) 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. In the first two types of designs, confounding can be avoided because the comparisons are made on the same population or randomly assigned sub-populations. However, in practice, many studies adopt the non-comparative design. 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 [2, 11]. The growing number of assessment instruments, as well as the rapid escalation in their cost, have generated an increasing need for scientifically rigorous comparisons of multiple diagnostic tests in clinical practice. Thus, a flexible meta-analysis framework is needed to combine information from all three designs for effectively ranking all candidate tests.

Very few papers have discussed how to simultaneously compare multiple candidate tests in meta-analysis [19, 51]. A naive procedure is to conduct separate MA-DT of each candidate test then compare their summary estimates [53]. However, there are some important drawbacks of this procedure. First, for studies that compared multiple tests, the accuracy estimates of each candidate tests from separate MA-DT are typically correlated. Ignoring such correlations can potentially lead to invalid inference. Secondly, when a candidate test is compared to a non-gold standard reference test in some studies, at least a second study comparing the same set of tests is needed to solve the non-identifiability problem [18]. 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 [15, 44]. In addition, as the test performance is summarized using a different study population, 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.

The remainder of this thesis is structured as follows. First, Chapter 2 provides a comprehensive review of the pros and cons of existing statistical methods for MA-DT, including models for settings with and without a gold-standard test. We go through both traditional and advanced methods in detail, and make recommendations for their application. Chapter 3 then proposes a hybrid GLMM to combine information from cohort and case-control studies, and to correct partial verification bias in meta-analyses of diagnostic tests simultaneously. We build this model under the assumption of a gold standard reference test. Model properties are investigated via simulation studies and model application is demonstrated by case studies. In Chapter 4, we discuss the situation with non-evaluable subjects. We extend the TGLMM approach[22] by treating non-evaluable subjects as missing data to adjust for potential bias. By extending the TGLMM to account for missing data, potential bias in disease prevalence estimate can be adjusted, and thus, bias in PPV and NPV estimates can be avoided. We add simulation studies to investigate and compare the extended TGLMM and alternative methods discussed in Schuetz et al.[41]. Next, we extend our topic from MA-DT to NMA-DT in Chapter 5, where we develop a NMA-DT framework from the perspective of missing data analysis. By simultaneously comparing all candidate tests and the 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. Finally, Chapter 6 summarizes our findings, limitations, and discuss areas for potential future work.

Chapter 2

Statistical methods for multivariate MA-DT

In this chapter, we provide a comprehensive review of existing statistical methods for MA-DT, including models for settings with and without a gold-standard test. Both conventional and advanced models are illustrated and compared for their advantages and disadvantages. In section 2.1, we summarize and compare different models when the referent test can be considered as a gold standard. In section 2.2, we introduce models in the absence of a gold standard. In section 2.3 we draw summaries of all methods and give recommendations.

2.1 Statistical methods when the reference test is a gold standard

When the reference test can be considered as a gold standard, let n_{itd} denote the number of subjects with index test results $T = t$ and disease status $Dis = d$ for study i ($i = 1, 2, \ldots, N$), where t and d are defined in Chapter 1. Thus, n_{i11}, n_{i00}, n_{i01}, and n_{i10} are the number of true positives, true negatives, false positives and false negatives for the ith study, respectively. Let $n_{i1+} = n_{i11} + n_{i10}$ and $n_{i0+} = n_{i01} + n_{i00}$ be the study-specific numbers of diseased and disease-free subjects. Then the study-specific sensitivity and specificity can be estimated as $\widehat{Se_i} = n_{i11}/n_{i1+}$, and $\widehat{Sp_i} = n_{i00}/n_{i0+}$. See Table 2.1 for

Table 2.1: 2 by 2 table for ith study

Index Test	Reference test		Total
	Positive (+)	Negative (−)	
Positive(+)	n_{i11}	n_{ni01}	
Negative(−)	n_{i10}	n_{i00}	
Total	n_{i1+}	n_{i0+}	n_{i++}

a typical 2 by 2 table for study i.

In this section, we will first discuss the conventional summary receiver operating characteristic (ROC) approach and a bivariate approach using linear mixed models (LMM). Both methods require direct calculations of study-specific sensitivities and specificities, and an ad hoc continuity correction when there are zero events in either arm of a study. Second, we will discuss the hierarchical summary ROC approach for jointly modeling positivity criteria and accuracy parameters, and a bivariate approach using GLMM for jointly modeling sensitivities and specificities. At last, we will discuss a trivariate approach using GLMM for jointly modeling prevalence, sensitivities and specificities to assess the correlations among the three parameters. The hierarchical summary ROC approach, and the bivariate and trivariate approaches are based on the exact binomial distribution and thus do not require any ad hoc continuity correction. [43]

2.1.1 The summary ROC method

The summary ROC curve method was first proposed by Moses et al. [54]. Reflecting the trade-off between sensitivity and specificity caused by implicit thresholds, this method has been widely used in diagnostic tests studies. The observed Se and Sp estimates form a concave ROC curve shape as the threshold varies. Such curve can be fitted by back-transforming the linear relationship between logit transformations of Se and Sp to the ROC space: First, if some studies have $n_{i11} = 0$ or $n_{i00} = 0$, an ad hoc continuity correction is applied by adding 0.5 to each of the 4 cells of such studies. After the correction, sensitivity is estimated to be $\widehat{Se_i} = (n_{i11} + 0.5)/(n_{i1+} + 1)$ and specificity is estimated to be $\widehat{Sp_i} = (n_{i00} + 0.5)/(n_{i0+} + 1)$ for the ith study. Second, define variables

S and D as the sum and the difference of logit transformed sensitivity and specificity, such that $S_i = \text{logit}(\widehat{Se_i}) + \text{logit}(\widehat{Sp_i})$ and $D_i = \text{logit}(\widehat{Se_i}) - \text{logit}(\widehat{Sp_i})$, where $\text{logit}(p) = \log\{p/(1-p)\}$, $0 < p < 1$. (This notation is slightly different than Moses et al.[54] because the original transformation is on Se and false positive rate (FPR, equivalent to 1-Sp)). One can see that $S_i = \log(\widehat{OR_i})$, where $\widehat{OR_i} = \frac{n_{i11}}{n_{i10}} \big/ \frac{n_{i01}}{n_{i00}}$ is the diagnostic odds ratio for the ith study. Third, for N studies, fit a linear regression line $S = a + bD$ either by an ordinary least squares or by a weighted least squares method weighing by the inverse of within-study variances $\text{var}(\log(OR_i))^{-1}$, where $\text{var}(\log(OR_i)) = 1/n_{i11} + 1/n_{i10} + 1/n_{i01} + 1/n_{i00}$ [4]. After fitting the regression line using either method, one can plot the summary ROC curve by the two estimated coefficients (i.e., intercept $\hat{a}$ and slope $\hat{b}$),

$$Se = \{1 + e^{-\hat{a}/(1-\hat{b})} \times \big(Sp/(1-Sp)\big)^{(1+\hat{b})/(1-\hat{b})}\}^{-1}, \tag{2.1}$$

with Se on the y-axis and $1 - Sp$ on the x-axis. To adjust for study-level covariates $\boldsymbol{Z}$ (e.g., different sites the diagnostic tests were taken), one can fit a model with $S_i = a + bD_i + cZ_i$. We can then have $S_i = \hat{a} + \hat{b}D_i + \hat{c}Z_i = (\hat{a} + \hat{c}Z_i) + \hat{b}D_i = \hat{a}' + \hat{b}'D_i$. The summary ROC curve can be plotted according to new estimates $\hat{a}'$ and $\hat{b}'$ given $\boldsymbol{Z}$.

The summary ROC method is easy to perform but suffers some shortcomings. On the one hand, its interpretations are known to be problematic. Walter discussed the interpretation of area under the curve (AUC) [17]. A summary ROC curve located closer to the left upper corner of the ROC space will have a larger AUC, indicating better predictive accuracy of a test [17]. However, the conclusion becomes unreliable when comparing tests whose summary ROC curves may cross each other. Alternative statistics, such as the partial AUC [55] and the Q point [56] also have limited applications. On the other hand, the model setting has some drawbacks. First, because $S_i = \log(\widehat{OR_i})$, the data are reduced to one outcome measure per study: diagnostic odds ratio. Independent summaries of sensitivity and specificity are not available, which could be important in test evaluating. Second, the model is restricted in that the between-study heterogeneity can only be adjusted by study level covariates, such that some components of the variance might not be explained. This is the reason why both Moses et al.[54] and Irwig et al. [57] recommended the unweighted least squares rather than the weighted as in a fixed effect model as a few large studies may dominate the result if the between-study variation is present. Third, in practice, study characteristics besides the cut-point effect contribute to the trade-off between sensitivity and specificity within a study[54, 58], which are not incorporated in the summary ROC curves. Finally, an arbitrary continuity correction is needed to handle zero cells. Moses showed that it can push the summary ROC curve far from the ideal upper left corner of the ROC space, giving biased results [55].

2.1.2 A bivariate approach based on LMM

To improve over the summary ROC method, Reitsma et al. proposed a bivariate LMM [11]. The model proceeds as follows. First, a logit transform of the sensitivity and specificity is applied to each study. Different from the summary ROC method, they are considered as random by allowing variation according to normal distributions, that is $\text{logit}(Se_i) \sim N(\alpha, \sigma_\mu^2)$ and $\text{logit}(Sp_i) \sim N(\beta, \sigma_\nu^2)$. A bivariate normal distribution can include possible correlation between sensitivity and specificity within study: $\begin{pmatrix} \text{logit}(Se_i) \\ \text{logit}(Sp_i) \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \mathbf{\Sigma}\right)$, where $\mathbf{\Sigma} = \begin{pmatrix} \sigma_\mu^2 & \sigma_{\mu\nu} \\ \sigma_{\mu\nu} & \sigma_\nu^2 \end{pmatrix}$ and $\sigma_{\mu\nu}$ denotes the covariance between logit sensitivity and specificity.

Second, to account for the sampling variation, the estimated logit sensitivity and specificity are assumed to be normally distributed as $\begin{pmatrix} \text{logit}(\widehat{Se_i}) \\ \text{logit}(\widehat{Sp_i}) \end{pmatrix} \sim N\left(\begin{pmatrix} \text{logit}(Se_i) \\ \text{logit}(Sp_i) \end{pmatrix}, C_i\right)$ for study i, with $C_i = \begin{pmatrix} \text{var}(\text{logit}(\widehat{Se_i})) & 0 \\ 0 & \text{var}(\text{logit}(\widehat{Sp_i})) \end{pmatrix}$, $\text{var}(\text{logit}(\widehat{Se_i})) = 1/n_{i11} + 1/n_{i10}$ and $\text{var}(\text{logit}(\widehat{Sp_i})) = 1/n_{i10} + 1/n_{i00}$. Note that, the general rule that $n_{i1+}\widehat{Se_i}$, $n_{i1+}(1-\widehat{Se_i})$, $n_{i0+}\widehat{Sp_i}$ and $n_{i0+}(1-\widehat{Sp_i})$ are at least five need to hold for normal approximation to be valid. Consequently, $\text{logit}(\widehat{Se_i})$ and $\text{logit}(\widehat{Sp_i})$ are assumed to have the following bivariate normal distribution:

$$\begin{pmatrix} \text{logit}(\widehat{Se_i}) \\ \text{logit}(\widehat{Sp_i}) \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \mathbf{\Sigma} + C_i\right) \tag{2.2}$$

Because the distributions of sensitivity and specificity are often skewed, one may prefer inference based on the medians rather than means as the overall diagnostic test performance summary. Based on parameter estimates, the median sensitivity and specificity can be back-transformed as $\widehat{Se_M} = \text{logit}^{-1}(\widehat{\alpha})$ and $\widehat{Sp_M} = \text{logit}^{-1}(\widehat{\beta})$, where $\text{logit}^{-1}(\cdot)$ is the inverse logit function such that $\text{logit}^{-1}(x) = 1/(1+\exp(-x))$. Similarly, confidence intervals for $\widehat{Se_M}$ and $\widehat{Sp_M}$ can be transformed from the confidence intervals of $\hat{\alpha}$ and $\widehat{\beta}$. The correlation between sensitivity and specificity can be estimated as $\frac{\hat{\sigma}_{\mu\nu}}{\hat{\sigma}_{\mu}\times\hat{\sigma}_{\nu}}$. The standard errors are $SE(\widehat{Se_M}) = \frac{SE(\alpha)}{1/\widehat{Se_M}+1/(1-\widehat{Se_M})}$ and $SE(\widehat{Sp_M}) = \frac{SE(\beta)}{1/\widehat{Sp_M}+1/(1-\widehat{Sp_M})}$ based on the Delta method. A summary ROC curve can be constructed by the regression of logit sensitivity over given specificity as

$$\text{logit}(Se) = \hat{\alpha} + \frac{\hat{\sigma}_{\mu\nu}}{\hat{\sigma}^2_{\nu}}\left(\text{logit}(Sp) - \widehat{\beta}\right). \tag{2.3}$$

In general, this approach is superior to the summary ROC model by analyzing sensitivity and specificity jointly in a bivariate LMM (BLMM). However, the bivariate approach estimates the degree of correlation between sensitivity and specificity, as well as both within- and between-study variation in the two indicators separately. A drawback of this approach is that an ad hoc continuity correction is required in the presence of zero cells, as with the summary ROC approach. In addition, the general rule of normal approximation is sometimes violated in practice[10]. The bivariate model can adjust for covariates by regression model in the mean vector of the bivariate normal distribution: $\begin{pmatrix} \text{logit}(Se_i) \\ \text{logit}(Sp_i) \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha+\gamma Z_i \\ \beta+\lambda Z_i \end{pmatrix}, \mathbf{\Sigma}\right)$, where Z_i is the study-level covariate and γ, λ are the corresponding coefficient parameters [4].

2.1.3 The hierarchical summary ROC approach

Rutter and Gatsonis proposed a hierarchical summary ROC approach[2], which is a simplification of the ordinal regression model by Tosteson and Begg: $g(\gamma_j(\boldsymbol{x})) = (\theta_j - \boldsymbol{\alpha}'\boldsymbol{x})e^{\boldsymbol{\beta}'\boldsymbol{x}}$, where $g(\cdot)$ is a link function, $\gamma_j(\boldsymbol{x})$ is the probability of a response being in one of the ordered categories given covariates $\boldsymbol{x}$, θ_j is the cutoff values of each category, $\boldsymbol{\alpha}$ is the location parameters and $\boldsymbol{\beta}$ is the scale parameter[59]. The hierarchical summary ROC approach reduces the ordinal regression model to two categories ($j = 1, 2$), $\boldsymbol{x}$ indicates true disease status (coded as 0.5 for $Dis = 1$ and -0.5 for $Dis = 0$) and $\gamma_j(\boldsymbol{x})$ correspond to positive test rates: Se_i and $1 - Sp_i$ (FPR)[2].

The first stage assumes binomial distributions of the number of positive outcomes in the ith study, i.e., $n_{i11} \sim \text{Bin}(n_{i1+}, Se_i)$ and $n_{i01} \sim \text{Bin}(n_{i0+}, 1 - Sp_i)$. Choose $g(\cdot)$

to be logit link, the model is written as

$$\text{logit}(Se_i) = (\theta_i + 0.5\alpha_i)e^{-0.5\beta},\ \text{logit}(1 - Sp_i) = (\theta_i - 0.5\alpha_i)e^{0.5\beta}, \tag{2.4}$$

where the latter is the same as $\text{logit}(Sp_i) = -(\theta_i - 0.5\alpha_i)e^{0.5\beta}$. The positivity criteria θ_i model the tradeoff between sensitivity and specificity in each study. Direct interpretations of the accuracy parameters α_i are that when $\beta = 0$, $\alpha_i = \text{logit}(Se_i) + \text{logit}(Sp_i) = log(OR_i)$, which is independent of θ_i. In the second stage, Rutter and Gatsonis allow θ_i and α_i to vary across studies[2]. Thus, θ_i and α_i are assumed independently normally distributed as: $\begin{pmatrix} \theta_i \\ \alpha_i \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_0 \\ \alpha_0 \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & 0 \\ 0 & \sigma_\alpha^2 \end{pmatrix}\right)$.

A summary ROC curve can be derived based on solving functions in (2.4) as

$$\text{logit}(Se_i) = \alpha_i e^{-\beta/2} + e^{-\beta}\text{logit}(1 - Sp_i)$$

Another possible construction of summary ROC curve pointed out by Chu et al.[12] is based on the bivariate normal distribution of θ_i and α_i and the Delta method as

$$\text{logit}(Se) = e^{-0.5\hat{\beta}}(0.5\hat{\alpha_0} + \hat{\theta_0}) + \frac{0.25\hat{\sigma}_\alpha^2 - \hat{\sigma}_\theta^2}{0.25\hat{\sigma}_\alpha^2 + \hat{\sigma}_\theta^2} \times e^{-\hat{\beta}}\{\text{logit}(Sp) - e^{-0.5\hat{\beta}}(0.5\hat{\alpha}_0 - \hat{\theta}_0)\}. \tag{2.5}$$

In addition, Arends et al. discussed several choices of SROC curves[9]. Median sensitivity and specificity estimates are $\widehat{Se_M} = \left\{1 + exp\{-(\hat{\theta}_0 + 0.5\hat{\alpha})e^{-0.5\hat{\beta}}\}\right\}^{-1}$ and $\widehat{Sp_M} = \left\{1 + exp\{(\hat{\theta}_0 - 0.5\hat{\alpha})e^{0.5\hat{\beta}}\}\right\}^{-1}$. Also, similar as the previous model, the hierarchical summary ROC approach can incorporate study level covariates by $\begin{pmatrix} \theta_i \\ \alpha_i \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_0 + \gamma Z_i \\ \alpha_0 + \lambda Z_i \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & 0 \\ 0 & \sigma_\alpha^2 \end{pmatrix}\right)$.

The hierarchical summary ROC approach incorporates both within- and between-study variability and the correlation between the summary statistics by random effects θ_i and α_i. Because sparse data is common in meta-analysis of diagnostic tests, especially under low event rates, an important advantage over the previous models is that the hierarchical summary ROC approach avoids the continuity correction by assuming binomial distributions[2]. A practical limitation of this model is that originally it is fitted via Bayesian Markov Chain Monte Carlo approach using BUGS, which requires some level of programming skills. This approach is found to be the same as the following bivariate GLMM with alternative parameterizations in some situations.

2.1.4 The bivariate generalized linear mixed model (GLMM)

Chu and Cole presented a bivariate GLMM to jointly analyze sensitivity and specificity using logit link[10]. Later, the bivariate GLMM was adjusted to a general link function [45]. The model starts with binomial distribution assumptions and applies link functions on the probability parameters:

$$n_{i11} \sim \text{Bin}(n_{i1+}, Se_i),\ n_{i00} \sim \text{Bin}(n_{i0+}, Sp_i),\ g(Se_i) = \alpha + \mu_i,\ g(Sp_i) = \beta + \nu_i, \quad (2.6)$$

where μ_i and ν_i are random effects follow bivariate normal distribution $\begin{pmatrix} \mu_i \\ \nu_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\mu^2 & \rho_{\mu\nu}\sigma_\mu\sigma_\nu \\ \rho_{\mu\nu}\sigma_\mu\sigma_\nu & \sigma_\nu^2 \end{pmatrix}\right)$, and $g(\cdot)$ is a link function such as the logit, probit, and complimentary log-log link. Different link functions can be applied to sensitivity and specificity separately. Though logit link is widely used in meta-analysis to date, Chu et al. argued that, for some meta-analyses, the choice of the link may affect model fit and inference[45]. The parameters σ_μ^2 and σ_ν^2 estimate the between-study variances and $\rho_{\mu\nu}$ explain possible correlations.

The model gives median estimates as $\widehat{Se_M} = \text{logit}^{-1}(\hat{\alpha})$ and $\widehat{Sp_M} = \text{logit}^{-1}(\hat{\beta})$. Similarly, confidence intervals for $\widehat{Se_M}$ and $\widehat{Sp_M}$ can be transformed from the confidence intervals of $\hat{\alpha}$ and $\hat{\beta}$. Study-level covariate $\boldsymbol{Z}$ can be adjusted by $g(Se_i) = \alpha + \mu_i + \gamma Z_i$ and $g(Sp_i) = \beta + \nu_i + \lambda Z_i$, where γ, λ are corresponding coefficient parameters. Different covariates could be adjusted for sensitivity and specificity. A regression line of $g(Se)$ on $g(Sp)$, $g(Se) = \hat{\alpha} + \hat{\rho}_{\mu\nu}\frac{\hat{\sigma}_\mu}{\hat{\sigma}_\nu}[g(Sp) - \hat{\beta}]$, gives the summary ROC curve by transforming to the ROC space. Also, alternative choices of the regression lines can construct different summary ROC curves with corresponding interpretations[9].

In addition to estimating the heterogeneity and correlation parameters, both hierarchical summary ROC and bivariate GLMM approaches have advantages over the bivariate LMM. First, the bivariate GLMM does not require the general rule of normal approximation to estimate var(logit($\widehat{Se_i}$)) and var(logit($\widehat{Sp_i}$)). Second, neither the two approaches require continuity correction because direct calculation of study-specific sensitivities and specificities is not involved. In the absence of study-level covariates, the two approaches are the same model with alternative parameterizations[5].

Both hierarchical summary ROC and bivariate GLMM can be fitted using the maximum likelihood approach. Several numerical methods might be used, for instance, the dual quasi-Newton optimization techniques, as implemented in SAS NLMIXED procedure. The standard errors and confidence intervals for interested parameters are estimated by the Delta method and are reported if specified in the ESTIMATE statement. To restrict the correlation coefficient $\rho_{\mu\nu}$ in the range [-1, 1] in the bivariate GLMM, one can use the Fisher's z transformation of $\rho_{\mu\nu}$ in programming. AUC for both hierarchical summary ROC and bivariate GLMM can be computed by numerical integration implemented in SAS macro.

2.1.5 The trivariate GLMM

The above approaches involving only sensitivities and specificities work best if all or the majority of the studies use case-control designs with non-identifiable prevalence. When disease prevalence estimation is allowed in cohort study designs, we can derive other clinically interested indices such as positive and negative predictive values (PPV and NPV) by estimates of sensitivity, specificity and prevalence. In this case, a challenge is the potential dependence of test performance on prevalence, which can be termed spectrum bias[26]. Typically, such dependence is mostly concerned when the bivariate diagnostic outcome is based on the cut point on continuous traits, thus misclassification rates could be higher among subjects with true value around the cut point[60]. To account for this potential dependence, Chu et al. extended the bivariate GLMM to a trivariate GLMM jointly modeling the disease prevalence, sensitivity and specificity[16]. Recently, Li and Fine proposed a Pearson type correlation coefficient to assess this dependence by an estimating equation-based regression framework[61].

Here, we only consider a trivariate GLMM based on the parameterization of π_i, Se_i and Sp_i, where π_i is the disease prevalence in ith study. The first level of this model assumes binomial distributions:

$$n_{i1+} \sim \mathrm{Bin}(n_{i++}, \pi_i),\ n_{i11} \sim \mathrm{Bin}(n_{i1+}, Se_i),\ n_{i00} \sim \mathrm{Bin}(n_{i0+}, Sp_i). \tag{2.7}$$

The parameters are modeled via link functions: $g(\pi_i) = \eta + \varepsilon_i$, $g(Se_i) = \alpha + \mu_i$ and $g(Sp_i) = \beta + \nu_i$. See Table 2.2 a two by two table accounting for disease prevalence.

To consider heterogeneity and potential correlations of the three parameters, ε_i, μ_i

Table 2.2: 2 by 2 table for ith study accounting for prevalence

Index Test	Reference test		Total
	Positive (+)	Negative (−)	
Positive(+)	n_{i11} $\pi_i Se_i$	n_{ni01} $(1-\pi_i)(1-Sp_i)$	
Negative(−)	n_{i10} $\pi_i(1-Se_i)$	n_{i00} $(1-\pi_i)Sp_i$	
Total	n_{i1+} π_i	n_{i0+} $1-\pi_i$	n_{i++} 1

and ν_i are assumed to be random effects with trivariate normal distribution:

$$\begin{pmatrix} \varepsilon_i \\ \mu_i \\ \nu_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma\right), \text{ where } \Sigma = \begin{pmatrix} \sigma_\varepsilon^2 & \rho_{\varepsilon\mu}\sigma_\mu\sigma_\varepsilon & \rho_{\varepsilon\nu}\sigma_\nu\sigma_\varepsilon \\ & \sigma_\mu^2 & \rho_{\mu\nu}\sigma_\mu\sigma_\nu \\ & & \sigma_\nu^2 \end{pmatrix}$$

The parameters σ_ε^2, σ_μ^2 and σ_ν^2 capture the between-study variance of the disease prevalence, sensitivity and specificity while $\rho_{\varepsilon\mu}$, $\rho_{\varepsilon\nu}$ and $\rho_{\mu\nu}$ represent correlations.

Standard software such as SAS NLMIXED can maximize the likelihood. To avoid including unnecessary parameters, model selection criteria such as Akaike information criterion (AIC) can be used. By parameter estimates, the medians are derived as $\widehat{\pi_M} = g^{-1}(\hat{\eta})$, $\widehat{Se_M} = g^{-1}(\hat{\alpha})$ and $\widehat{Sp_M} = g^{-1}(\hat{\beta})$. In this model, covariates can be incorporated in sensitivities, specificities and disease prevalence as was done for the bivariate GLMM.

2.2 Statistical methods when the reference test is not a gold standard

Limited meta-analysis tools are available when the reference test is imperfect. Walter et al. discussed the latent class model for a meta-analysis of two diagnostic tests[17]. Sadatsafavi et al. presented a latent class random effects model[19]. However, other than prevalence, only one of the sensitivity and specificity can be implemented as a random effect. Dendukuri et al. presented a Bayesian approach, which is an extension of the hierarchical summary ROC model, to adjust for different reference standards[20].

We hereby introduce the latent class random effects model by Chu et al. using random effects to allow variation and correlation in sensitivity, specificity and prevalence between studies[18].

Let (Se_{Bi}, Sp_{Bi}) be the pair of diagnostic accuracy parameters for the reference test while (Se_{Ai}, Sp_{Ai}) be the pair for the index test. To construct the 2 by 2 table (Table 2.3) for such studies, both the above pairs of statistics and the disease prevalence are needed.

Table 2.3: 2 by 2 table when the reference test is not a gold standard

Index Test	Reference test		Total
	Positive (+)	Negative (−)	
Positive(+)	n_{i11} $p_{i11} = \pi_i Se_{Ai} Se_{Bi} + (1-\pi_i)(1-Sp_{Ai})(1-Sp_{Bi})$	n_{i01} $p_{i01} = \pi_i Se_{Ai}(1-Se_{Bi}) + (1-\pi_i)(1-Sp_{Ai})Sp_{Bi}$	
Negative(−)	n_{i10} $p_{i10} = \pi_i(1-Se_{Ai})Se_{Bi} + (1-\pi_i)Sp_{Ai}(1-Sp_{Bi})$	n_{i00} $p_{i00} = \pi_i(1-Se_{Ai})(1-Se_{Bi}) + (1-\pi_i)Sp_{Ai}Sp_{Bi}$	
Total	n_{i1+} $p_{i1+} = \pi_i Se_{Bi} + (1-\pi_i)(1-Sp_{Bi})$	n_{i0+} $p_{i0+} = \pi_i(1-Se_{Bi}) + (1-\pi_i)Sp_{Bi}$	n_{i++} 1

The four counts in Table 2.3 follow a multinomial distribution, with log-likelihood being:

$$\log L = \sum_i \{n_{i11}\log(p_{i11}) + n_{i10}\log(p_{i10}) + n_{i01}\log(p_{i01}) + n_{i00}\log(p_{i00})\}. \tag{2.8}$$

Chu et al. used random effects to model between and with-in study heterogeneity and potential correlations[18]. We write this model in a form suit for a general link function $g()$:

$$g(\pi_i) = \eta + \varepsilon_i;\ g(Se_{Ai}) = \alpha_A + \mu_{Ai};\ g(Sp_{Ai}) = \beta_A + \nu_{Ai};$$

$$g(Se_{Bi}) = \alpha_B + \mu_{Bi};\ g(Sp_{Bi}) = \beta_B + \nu_{Bi}$$

where random effects follow a multivariate normal distribution: $(\varepsilon_i, \mu_{Ai}, \nu_{Ai}, \mu_{Bi}, \nu_{Bi})' \sim N(\mathbf{0}, \mathbf{\Sigma})$ with variance-covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_\varepsilon^2 & \rho_{\varepsilon\mu_A}\sigma_\varepsilon\sigma_{\mu_A} & \rho_{\varepsilon\nu_A}\sigma_\varepsilon\sigma_{\nu_A} & \rho_{\varepsilon\mu_B}\sigma_\varepsilon\sigma_{\mu_B} & \rho_{\varepsilon\nu_B}\sigma_\varepsilon\sigma_{\nu_B} \\ & \sigma_{\mu_A}^2 & \rho_{\mu_A\nu_A}\sigma_{\mu_A}\sigma_{\nu_A} & \rho_{\mu_A\mu_B}\sigma_{\mu_A}\sigma_{\mu_B} & \rho_{\mu_A\nu_B}\sigma_{\mu_A}\sigma_{\nu_B} \\ & & \sigma_{\nu_A}^2 & \rho_{\nu_A\mu_B}\sigma_{\nu_A}\sigma_{\mu_B} & \rho_{\nu_A\nu_B}\sigma_{\nu_A}\sigma_{\nu_B} \\ & & & \sigma_{\mu_B}^2 & \rho_{\mu_B\nu_B}\sigma_{\mu_B}\sigma_{\nu_B} \\ & & & & \sigma_{\nu_B}^2 \end{pmatrix}$$

Median estimates of prevalence, sensitivities and specificities can be achieved by $\widehat{\pi_M} = g^{-1}(\widehat{\eta})$, $\widehat{Se_{AM}} = g^{-1}(\widehat{\alpha_A})$, $\widehat{Sp_{AM}} = g^{-1}(\widehat{\beta_A})$, $\widehat{Se_{BM}} = g^{-1}(\widehat{\alpha_B})$ and $\widehat{Sp_{BM}} = g^{-1}(\widehat{\beta_B})$. Variance and correlation parameter estimates can be derived from $\widehat{\mathbf{\Sigma}}$. Covariates $\boldsymbol{Z}$ can be adjusted by linear regressions in the mean vectors, for instance $g(\pi_i) = \eta + \varepsilon_i + \gamma Z_i$.

This latent class random effects model fills in the gap of models for meta-analysis of diagnostic test under imperfect reference test condition. It can evaluate the performance of both the diagnostic test of interest and the reference test while keep all the advantages of the GLMMs. A limitation applies when fitting this model by SAS NLMIXED. One could encounter convergence problem because of limited number of studies and relatively large number of parameters. Possible simplifying assumption could be independence of disease prevalence against the other parameters. Also, to avoid including redundant random effects whose variance approximates zero, one can apply a forward selection based on AIC.

2.3 Discussion

In this chapter, we discussed the methods for evaluating the performance of diagnostic tests for situations when the reference test can be considered a gold standard, as well as situations when it is error-prone. Under the scenario with gold standard, we reviewed the traditional summary ROC method, bivariate LMM and the hierarchical summary ROC model. Then we focused on the random effect GLMM, because it has several advantages over simpler methods. In this section, we showed how the bivariate GLMM can be fitted using different link functions other than logit link, and extended the approach to a trivariate GLMM to model prevalence as well as sensitivity and specificity. Under the situation with no gold standard, we built upon the latent class model proposed by Walter et al.[17] by adding random effects to quantify possible correlation and variation following the method by Chu et al.[18].

Among the models presented, the summary ROC approach is simple and widely used, while receives a number of critical comments on problems related to the interpretation of summary ROC curves, the fixed effects model and the continuity correction. The bivariate LMM improves over the summary ROC approach by assuming random effects to explain both within- and between study variations and possible correlation. The bivariate LMM can give inferences both in terms of summary ROC curves and summary statistics of overall test performance. However, it has some limitations due to the continuity correction and the validity of normal approximation. The GLMMs do not have the limitations of the above models by assuming exact binomial distributions. The bivariate GLMM, which is essentially the same as the hierarchical summary ROC model in some situations, can be used when the majority of the studies use case-control designs and the trivariate GLMM can be used when most of the studies are cohort studies.

A limitation related to the GLMMs is that the meta-analysis reported often includes a mixture of case-control and cohort studies designs. Thus using either the bivariate or the trivariate GLMM for all the studies can lead to some issues. Another problem arises when fitting the trivariate GLMM and the latent class random effects models in SAS procedure NLMIXED. The more random effects included, the longer it took to converge. Under such situations, one can first get raw estimates of the desired parameters by fitting the data in models with less random effects. The raw estimates can then be used to adjust the initial values to improve convergence. For the latent class random effects model, one may need to apply simpler assumptions for ease of fitting.

Chapter 3

A hybrid Bayesian hierarchical model combining cohort and case-control studies for meta-analysis of diagnostic tests: Accounting for partial verification bias

As discussed in the previous Chapter, many meta-analyses of diagnostic tests in practice contain both cohort and case-control study designs [23]. However, the trivariate GLMM [16] can only include cohort studies with information estimating study-specific disease prevalence and the bivaraite GLMM cannot estimate prevalence and account for the correlation between prevalence and accuracy parameters. On the other hand, some diagnostic accuracy studies only select a subset of samples to be verified by the reference test. It is known that ignoring unverified subjects may lead to partial verification bias in the estimation of prevalence, sensitivities, and specificities in a single study [31]. However, the impact of this bias on a meta-analysis has not been investigated. In this chapter, we propose a novel hybrid Bayesian hierarchical model combining cohort and casecontrol studies and correcting partial verification bias at the same time. We first describe the proposed method in Section 3.1. We investigate the performance of the proposed methods through a set of simulation studies in Section 3.2. Section 3.3 provides two motivating case studies: assessing the diagnostic accuracy of gadolinium-enhanced magnetic resonance imaging (MRI) in detecting lymph node metastases and of adrenal fluorine-18 fluorodeoxyglucose positron emission tomography (PET) in characterizing adrenal masses. This chapter ends with a discussion in Section 3.4. The data sets for the two case studies are given in Appendix A.

3.1 Bayesian hierarchical model

3.1.1 Notations

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

Table 3.1 presents the data structure and notation for the ith study when it is a cohort study or a case-control study. In each cell, the number of cell counts and the corresponding probabilities are presented. The left panel is for a cohort studies, which extends a standard 2×2 table (Table 2.2) to allow for partial verification. The sum of all cell probabilities is one. The right half is for a case-control studies with a typical

2×2 table (Table 2.1). The cell probabilities sum up to one for diseased and non-diseased subjects separately. Derivations of the cell probabilities for cohort studies are also provided at the footnote of Table 3.1 .

Table 3.1: Data display for the ith study when it is a cohort study and when it is a case-control study.

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

In each cell, the number of cell counts are presented and the probabilities corresponding to the cell counts are presented below the cell counts.

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

$p_{i11} = P(V=1|Dis=1, T=1)P(Dis=1)P(T=1|Dis=1) = (1-\omega_{i1m}-\omega_{im1})\pi_i Se_i$,

$p_{i10} = P(V=1|Dis=0, T=1)P(Dis=0)P(T=1|Dis=0) = (1-\omega_{i1m}-\omega_{im1})(1-\pi_i)(1-Sp_i)$,

$p_{i01} = P(V=1|Dis=1, T=0)P(Dis=1)P(T=0|Dis=1) = (1-\omega_{i0m}-\omega_{im1})\pi_i(1-Se_i)$,

$p_{i00} = P(T=0|Dis=0)P(V=1|Dis=0, T=0)P(Dis=0) = (1-\omega_{i0m}-\omega_{im0})(1-\pi_i)Sp_i$.

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

$p_{i1m} = P(V=0|T=1)P(T=1) = P(V=0|T=1)\{P(T=1, Dis=1) + P(T=1, Dis=0)\} = P(V=0|T=1)\{P(Dis=1)P(T=1|Dis=1) + P(Dis=0)P(T=1|Dis=0)\} = \omega_{i1m}\{\pi_i Se_i + (1-\pi_i)(1-Sp_i)\}$,

$p_{i0m} = P(V=0|T=0)P(T=0) = P(V=0|T=0)\{P(T=0, Dis=1) + P(T=0, Dis=0)\} = P(V=0|T=0)\{P(Dis=1)P(T=0|Dis=1) + P(Dis=0)P(T=0|Dis=0)\} = \omega_{i0m}\{\pi_i(1-Se_i) + (1-\pi_i)Sp_i\}$,

$p_{im1} = P(V=0|Dis=1)P(Dis=1) = \omega_{im1}\pi_i$, and $p_{im0} = P(V=0|Dis=0)P(Dis=0) = \omega_{im0}(1-\pi_i)$.

3.1.2 The likelihood with random effects accounting for heterogeneity

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

$$L(\boldsymbol{\omega}|\text{Data}) \propto \prod_{i=1}^{N_1} \left\{ \omega_{i1m}^{n_{i1m}} \omega_{im1}^{n_{im1}} \omega_{im0}^{n_{im0}} \omega_{i0m}^{n_{i0m}} \prod_{j,k=0,1} (1 - \omega_{ijm} - \omega_{imk})^{n_{ijk}} \right\} \tag{3.1}$$

and

$$L(\boldsymbol{\theta}|\text{Data}) \propto \prod_{i=1}^{N} Se_i^{n_{i11}} (1 - Sp_i)^{n_{i10}} (1 - Se_i)^{n_{i01}} Sp_i^{n_{i00}} \prod_{i=1}^{N_1} \pi_i^{\sum_j n_{ij1}} (1 - \pi_i)^{\sum_j n_{ij0}} h_{i1}^{n_{i1m}} h_{i0}^{n_{i0m}}, \tag{3.2}$$

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

To account for potential between-study heterogeneity, we consider a GLMM:

$$g(\pi_i) = \eta + \varepsilon_i; \quad g(Se_i) = \alpha + \mu_i; \quad g(Sp_i) = \beta + \nu_i, \tag{3.3}$$

where $g()$ is a link function, and $(\varepsilon_i, \mu_i, \nu_i)^T$ is a random effect vector. To account for potential correlation among π_i, Se_i and Sp_i, $(\varepsilon_i, \mu_i, \nu_i)^T$ is assumed to be multivariate normally distributed as $(\varepsilon_i, \mu_i, \nu_i)^T \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where

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

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

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

The marginal likelihood integrated over random effects is:

$$L(\boldsymbol{\theta}, \boldsymbol{\omega}) = \iiint L(\boldsymbol{\theta}, \boldsymbol{\omega} | \text{Data}) \times p(\mu_i, \nu_i, \varepsilon_i | \boldsymbol{\Sigma}) \, d\varepsilon_i \, d\mu_i \, d\nu_i \tag{3.4}$$

Frequentist methods (such as the maximum likelihood estimate) may converge slowly or have convergence problems due to the need to maximize the marginal likelihood with trivariate integrations, and the corresponding asymptotic approximations for standard errors of functions of parameters may not be sufficiently accurate [44].

3.1.3 Bayesian posterior sampling approach

In this chapter, we consider fully Bayesian approaches using Markov chain Monte Carlo (MCMC) methods for parameter estimation. In most instances, inferences obtained by Bayesian and classical frequentist methods are similar when the former uses non-informative or weakly informative prior distributions for all model parameters [62]. Compared to the frequentist methods, MCMC algorithms permit full posterior inference (e.g., credible intervals) even when the normality approximation based on large sample theory is insufficient, which is valuable here because the sampling distributions of π, Se, Sp, PPV, NPV, LR+ and LR− are often skewed and the number of studies in the meta-analysis is relatively small or moderate (e.g., $N < 30$). Specifically, we will draw posterior inference using Gibbs and Metropolis-Hastings sampling algorithms [63, 64, 65, 66] with convergence assessed using trace plots, sample autocorrelations, and statistical convergence diagnostic tests [67, 68].

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

$$p(\eta) \sim N(0, 10^2); \quad p(\alpha) \sim N(0, 10^2); \quad p(\beta) \sim N(0, 10^2); \quad p(\mathbf{\Sigma}^{-1}) \sim W(\mathbf{R}, v), \tag{3.5}$$

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

$$p(\eta, \alpha, \beta, \mathbf{\Sigma}|\text{Data}) \propto L(\boldsymbol{\theta}|\text{Data}) p(\eta) p(\alpha) p(\beta) p(\mathbf{\Sigma}) \prod_{i=1}^{N} p(\varepsilon_i, \mu_i, \nu_i|\mathbf{\Sigma}) \tag{3.6}$$

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

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

$$PPV = \frac{g^{-1}(\eta) g^{-1}(\alpha)}{g^{-1}(\eta) g^{-1}(\alpha) + \{1 - g^{-1}(\eta)\}\{1 - g^{-1}(\beta)\}}$$

$$NPV = \frac{\{1 - g^{-1}(\eta)\} g^{-1}(\beta)}{\{1 - g^{-1}(\eta)\} g^{-1}(\beta) + g^{-1}(\eta)\{1 - g^{-1}(\alpha)\}}$$

$$LR{+} = g^{-1}(\alpha)/\{1 - g^{-1}(\beta)\}, \quad LR{-} = \{1 - g^{-1}(\alpha)\}/g^{-1}(\beta)$$

3.2 Simulation

3.2.1 Simulation design

We conduct 12 sets of simulations to compare the proposed Bayesian hybrid GLMM (3.1) to two alternative approaches which researchers are likely to apply in practice: 1) a complete case analysis approach in which subjects not verified are ignored (model 3.2); and 2) a trivariate GLMM approach in which case-control studies are excluded from the analysis (model 3.3). To fit model 3.2, case-control and cohort studies are combined as in the hybrid GLMM, while the missing counts are excluded. To fit model 3.3, the missing counts are accounted for as in the hybrid GLMM, while all case-control studies are excluded from the data. To investigate the performance of the proposed hybrid GLMM, for each generated dataset, we fit the hybrid GLMM, model 3.2 and model 3.3 separately using R package BRugs [69]. Each dataset contains equal numbers of case-control and cohort studies, where cohort studies are subject to partial verification. The probabilities of missing a reference test are 0.2 and 0.8, given diagnostic test results being positive and negative, respectively. The median prevalence is set to be 0.2 with the variances as $\sigma_{\varepsilon}^2 = \sigma_{\mu}^2 = \sigma_{\nu}^2 = 1$, and the number of subjects per study is chosen to be similar to the case studies in Section 3.3. Specifically, we consider 12 settings with small (10) or moderate (30) number of studies in a meta-analysis and high sensitivity (specificity) as 0.9 (0.95), or low sensitivity (specificity) as 0.7 (0.8), respectively. To evaluate the impact of the correlation structure, the correlation parameters $(\rho_{\varepsilon\mu}, \rho_{\varepsilon\nu}, \rho_{\mu\nu})$ are chosen as (0, 0, 0), (0.5, -0.5, -0.5) or (0.8, -0.8, -0.8) to correspond to no correlation, moderate or strong correlations among disease prevalence and test sensitivity and specificity (in logit scale). We assume a positive correlation between π_i and Se_i as it is likely to happen when population with higher prevalence may have more patients with clear-cut disease condition, leading to a higher sensitivity. However, a negative correlation was also observed in some studies[21]. For each setting, 2000 replicates are generated using the trivariate logit-normal random effects model. The posterior statistics (median and 95% equal tailed credible interval) are summarized from 10000 posterior samples with 5000 burn-in iterations. Model performance is evaluated by comparing bias, relative efficiency (RE) and 95% equal tailed credible interval coverage probability (CP) of the three models. The REs are calculated as the ratio of the variances of estimates from the hybrid model and the variances of the estimates from an alternative model. The larger RE, the more efficient the estimate.

3.2.2 Simulation results

We summarized in Table 3.2 the bias, RE and CP of estimated overall Se, Sp, π, NPV and PPV for settings with 30 studies and median Se/Sp as 0.7/0.8. Simulation results under other simulation settings are attached in Appendix B. Under all settings, the hybrid GLMM gives nearly unbiased estimates and satisfactory CP of Se, Sp, π, PPV and NPV that are close to the nominal level of 95%.

Table 3.2: Summary of 2000 simulations with data generated from settings with 30 studies, true Se (Sp)=0.7 (0.8) and different correlation assumptions.

		Sp			Se			π			PPV			NPV		
Corr[a]	Model[b]	Bias	RE	CP	Bias	RE	CP	Bias	RE	CP	Bias	RE	CP	Bias	RE	CP
0	3.1	0	1	0.93	0	1	0.94	0.01	1	0.92	0.01	1	0.93	−0.01	1	0.93
0	3.2	−0.13	NA	0.29	0.1	NA	0.42	0.09	NA	0.71	0.02	NA	0.92	−0.02	NA	0.88
0	3.3	−0.01	0.47	0.93	0	0.29	0.93	0.01	1.07	0.93	0	0.83	0.94	−0.01	0.62	0.93
0.5	3.1	0	1	0.94	0	1	0.95	0.01	1	0.93	0	1	0.94	0	1	0.94
0.5	3.2	−0.13	NA	0.26	0.11	NA	0.34	0.06	NA	0.84	−0.01	NA	0.93	−0.01	NA	0.94
0.5	3.3	−0.01	0.5	0.94	0.02	0.32	0.94	0.01	0.95	0.93	0	0.89	0.95	0	0.68	0.95
0.8	3.1	0	1	0.93	0	1	0.94	0	1	0.95	0	1	0.95	0	1	0.96
0.8	3.2	−0.13	NA	0.29	0.11	NA	0.3	0.04	NA	0.88	−0.03	NA	0.94	0.01	NA	0.96
0.8	3.3	0	0.48	0.93	0.03	0.33	0.94	0	0.75	0.94	0.01	0.89	0.97	0.01	0.56	0.96

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

[b] Model = 3.1: Hybrid GLMM; Model = 3.2: a complete case analysis approach in which subjects not verified are ignored; Model = 3.3: a trivariate GLMM approach in which case-control studies are excluded from the analysis.

As expected, when the partial verification is ignored as in model 3.2, some of the posterior estimates were considerably biased with grossly small CP. Under our simulation assumptions, specificities are under-estimated, and prevalences and sensitivities are over-estimated, which agrees with the illustrative example described in the introduction. An intuitive explanation is that if we assume $\omega_{i1m} = 0$ and $\omega_{i0m} > 0$ such that partial verification would decrease n_{i10} and n_{i00} but n_{i11} and n_{i10} remain the same, leading to increased Se and decreased Sp estimates. From the simulations we also observe that the bias in π is larger when true Se (Sp) was 0.9 (0.95) (ranges from 0.13 to 0.2) than when true Se (Sp) was 0.7 (0.8) (ranges from 0.04 to 0.11), respectively. On the contrary, Sp and Se estimates are more biased when true Se (Sp) is 0.7 (0.8) (ranges from 0.04 to 0.14 and from 0.09 to 0.11 respectively) than when true Se (Sp) is 0.9 (0.95) (ranges from 0.01 to 0.04 and from 0.03 to 0.04 respectively). Because the estimates are biased, we do not calculate the RE of these estimates. Estimates of PPV and NPV from model 3.2 are nearly unbiased. Under the MAR assumption, we have $P(V = 1|Dis = 1, T = 1) = P(V = 1|T = 1)$, where $V = 1$ indicates verification of disease status, which would imply that $P(Dis = 1|V = 1, Y = 1) = P(Dis = 1|Y = 1)$[1].

When only cohort studies are included as in model 3.3, the estimates are nearly unbiased and the CPs remain close to the nominal level. Specifically, for estimation of prevalence, when there is no correlation, model 3 performs as well as the hybrid GLMM because only the cohort studies have information of π. However, as the correlation becomes larger, the RE of model 3.3 becomes smaller indicating the hybrid GLMM is gaining efficiency. This is because information of estimating prevalence is borrowed from Se and Sp estimates from case-control studies. For estimations of Se and Sp, substantial loss of efficiency can be observed using model 3.3 with REs around 0.3 and 0.5. The reason is that half of the whole study set (the case-control studies) are discarded in model 3.3, which contains important information to estimate Se and Sp. For estimations of PPV and NPV, loss of efficiency can also be observed with REs ranging from 0.76 to 0.92, and from 0.44 to 0.69, respectively. Generally, the relative efficiencies indicate that estimates from hybrid model are preferable. In summary, the hybrid model performs well in correcting partial verification bias and gaining efficiency by combining the information from cohort and case-control studies.

3.3 Case study

3.3.1 Meta-analysis of gadolinium-enhanced MRI in detecting lymph node metastases

We reanalyze the meta-analysis conducted by Klerkx et al. [23] using the proposed approach. Thirty-two studies were reported assessing diagnostic accuracy of gadolinium-enhanced MRI in detecting lymph node metastases, with histopathology test as the reference gold standard test. A bivariate random effects model [11] was applied by Klerkx et al. [23]. Overall sensitivity and specificity were estimated as 0.72 with 95% confidence interval (0.66, 0.79) and 0.87 with 95% confidence interval (0.82, 0.91), respectively. Data for each study is reported in the systematic review, as well as the QUADAS [28] quality assessment checklist.

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

Model fitting via bayesian approach

We fit the data using the hybrid GLMM with logit link function using WinBUGS [70] to draw posterior samples. Model 3.2 and model 3.3 are also fitted for comparison. Non-informative normal priors $N(0, 10^2)$ are given to η, α and β and a Wishart prior

$W(\mathbf{R}, \mathbf{v})$ is given to the precision matrix $\mathbf{\Sigma}^{-1}$ as in (3.5). The degree of freedom v in the Wishart prior is set as $v = 4$, as pointed out by Tokuda et al. that when $v = k + 1$, where k is the dimension of $\mathbf{\Sigma}$, the correlation coefficient parameters in $\mathbf{\Sigma}$ will have an approximately Uniform $(-1, 1)$ vague prior [71, 72]. A scaled Wishart prior method is applied by setting $v = 4$ and $\mathbf{R}$ as a 3 by 3 identity matrix. Wishart prior is known as a conjugate prior for the precision matrix in a multivariate normal distribution. However, it is restricted in that it implies the same prior assumption on all of the variance components. The scaled Wishart prior method allows the flexibility of having separate priors on each of the precision parameter, while keeping the conjugacy property [73]. The same priors are applied to model 3.2 and model 3.3. After 100,000 burn-in samples, 1,000,000 posterior samples are collected. The median estimates and 95% credible interval (CI) of interested parameters are presented in Table 3.3, where the estimates from hybrid GLMM are in bold.

Table 3.3: Median estimates and 95% CI for meta-analysis of MRI: comparing two prior families (the scaled and unscaled inverse Wishart prior) and comparing different choices of the Wishart prior parameter $\boldsymbol{R}$.

	Scaled Method			Unscaled Method	
	Hybrid GLMM	Model 3.2	Model 3.3	Hybrid GLMM	Hybrid GLMM
Parameters	**R=I**	**R=I**	**R=I**	**R=I**	diag(**R**)=**(9.8,3.3,2.2)**
π	**0.39 (0.35,0.44)**	0.37 (0.32,0.42)	0.39 (0.35,0.44)	0.39 (0.34, 0.45)	0.39 (0.34,0.45)
σ_ε	**0.32 (0.08,0.57)**	0.45 (0.22,0.73)	0.32 (0.10,0.57)	0.46 (0.31,0.69)	0.47 (0.33,0.69)
Se	**0.76 (0.70,0.82)**	0.73 (0.66,0.78)	0.77 (0.71,0.83)	0.77 (0.70,0.82)	0.78 (0.69,0.85)
σ_μ	**0.55 (0.21,0.99)**	0.47 (0.17,0.92)	0.47 (0.10,0.99)	0.64 (0.41,1.00)	1.03 (0.76,1.45)
Sp	**0.84 (0.79,0.89)**	0.87 (0.82,0.91)	0.85 (0.79,0.90)	0.84 (0.79,0.89)	0.85 (0.79,0.90)
σ_ν	**0.92 (0.62,1.33)**	0.89 (0.62,1.31)	0.74 (0.41,1.22)	0.88 (0.61,1.27)	0.97 (0.69,1.37)
$\rho_{\mu\nu}$	**−0.47 (−0.92,0.15)**	−0.56 (−0.96,0.11)	−0.60 (−0.97,0.31)	−0.39 (−0.76,0.17)	−0.49 (−0.83,0.12)
$\rho_{\varepsilon\mu}$	**0.08 (−0.74,0.85)**	0.37 (−0.50,0.94)	0.16 (−0.71,0.89)	−0.01 (−0.55,0.56)	0.04 (−0.56,0.61)
$\rho_{\varepsilon\nu}$	**−0.42 (−0.92,0.40)**	−0.57 (−0.93,0.09)	−0.41 (−0.91,0.36)	−0.30 (−0.72,0.31)	−0.34 (−0.76,0.31)
NPV	**0.85 (0.80,0.88)**	0.85 (0.80,0.88)	0.85 (0.81,0.89)	0.85 (0.80,0.89)	0.86 (0.80,0.91)
PPV	**0.76 (0.69,0.83)**	0.76 (0.69,0.82)	0.76 (0.69,0.83)	0.76 (0.69,0.83)	0.77 (0.69,0.84)
LR+	**3.22 (2.37,4.52)**	2.65 (1.98,3.61)	3.34 (2.45,4.82)	3.25 (2.35,4.63)	3.53 (2.25,5.76)
LR−	**0.31 (0.22,0.42)**	0.38 (0.28,0.51)	0.30 (0.21,0.41)	0.31 (0.22,0.43)	0.28 (0.17,0.44)

Model 3.2 stands for a complete-case analysis where case-control and cohort studies are combined while partial verification are ignored and model 3.3 stands for a trivariate GLMM where partial verificaiton bias is adjusted while case-control studies are excluded.

The hybrid GLMM gives posterior median estimates of overall sensitivity as 0.76, which is 0.04 higher than the estimate reported by Klerkx et al. [23] and with a slightly narrower 95% CI, i.e., an interval of (0.70, 0.82) from the hybrid GLMM versus (0.66, 0.79) from the bivariate random effects method. The posterior median is 0.84 for the overall specificity, which is 0.03 lower than the bivariate model estimates. In addition, our approach allows the estimation of disease prevalence and possible correlations among prevalence, Se and Sp. We also presented posterior estimates of PPV, NPV, LR+ and LR− in Table 3.3. In this case-study, the estimates from hybrid GLMM and from model 3 are very similar as only 6 of the 32 studies are case-control studies, e.g., the median sensitivity is estimated as 0.762 in hybrid model and 0.770 in model 3.3. The quantile contours of posterior estimates Se versus π, Sp versus π, Se versus Sp and NPV versus PPV at quantile levels 0.25, 0.5, 0.75, 0.90 and 0.95 are presented in Figure 3.1 A-D, respectively. Figure 3.1A indicates slightly positive correlation between Se and π. Negative correlation can be observed between Sp and π and between Se and Sp in Figure 3.1B and 3.1C. This observation agrees with the posterior estimates of correlation coefficients in Table 3.3: posterior $\rho_{\varepsilon\mu}$, $\rho_{\varepsilon\nu}$ and $\rho_{\mu\nu}$ has median estimates as 0.08, -0.42 and -0.47. Slightly negative correlation is shown in Figure 3.1D between NPV and PPV. The observed estimates of Se and Sp for each study and the posterior estimates from the hybrid GLMM and model 3.2 are ploted in Figure 3.2. The plot shows that different approaches can lead to different posterior estimates.

Sensitivity analysis to prior distributions for $\mathbf{\Sigma}^{-1}$

In addition to the scaled Wishart prior, an unscaled Wishart prior is commonly used in which no scale parameter is imposed on the precision matrix components. For the unscaled Wishart prior for $\mathbf{\Sigma}^{-1}$, there are several applicable selections of matrix $\mathbf{R}$: the identity matrix [74], or a diagonal matrix with diagonal entries chosen to be close to the diagonal elements of posterior precision matrix [62]. In the latter option, previous estimates of the precision matrix can serve as a prior for further estimations. As the scaled Wishart prior in previous paragraph gives posterior variance parameter estimates close to $(0.32^2, 0.55^2, 0.91^2)$, we choose the Wishart prior parameter $\mathbf{R}$ to have diagonal entries close to $(0.32^2, 0.55^2, 0.91^2)^{-1} \approx (9.8, 3.3, 1.2)$. Thus, to study whether the posterior estimates are sensitive to different prior assumptions, we fit the data via two

Figure 3.1: Quantile contours of posterior densities from estimates of the meta-analysis of gadolinium-enhanced MRI in detecting lymph node metastases assuming scaled Wishart prior. A-D plot posterior Se versus prevalence (π), Sp versus π, Se versus Sp and PPV versus NPV, respectively, at quantile levels 0.25, 0.5, 0.75, 0.9 and 0.95.

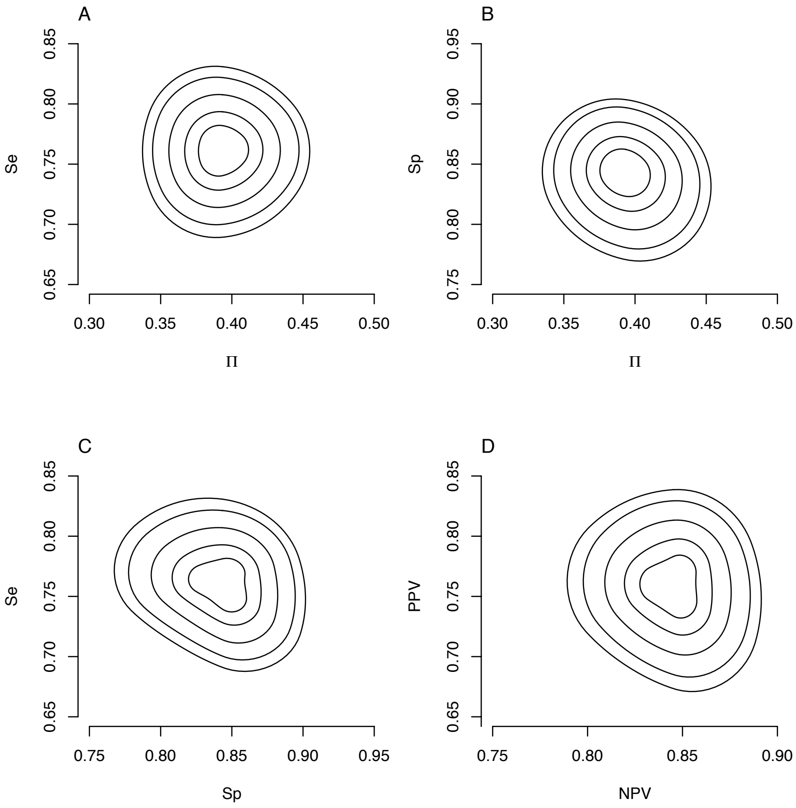
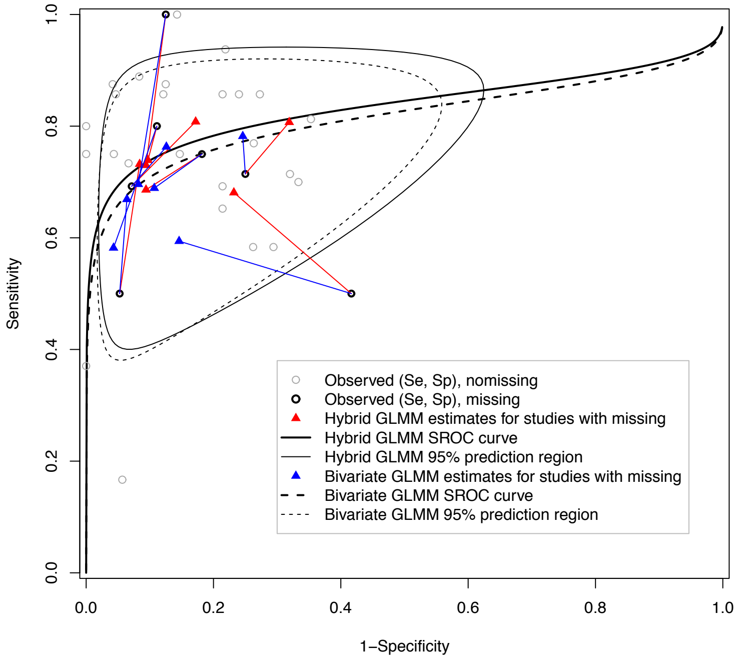


Figure 3.2: SROC curves from the Hybrid GLMM and the bivariate GLMM using MLE approach. Solid lines are the SROC curve from the hybrid GLMM estimates and the 95% prediction region for the summary point estimates of Se and Sp. Dashed lines are the SROC curve from the bivaraite estimates and the 95% prediction region for the summary point estimates of Se and Sp. Black and gray circles are the observed Se and Sp from studies with and without missing counts, respectively. Red and blue triangles are the posterior estimates of Se and Sp from the Hybrid GLMM and the Bivariate GLMM ignoring partial verification, respectively.



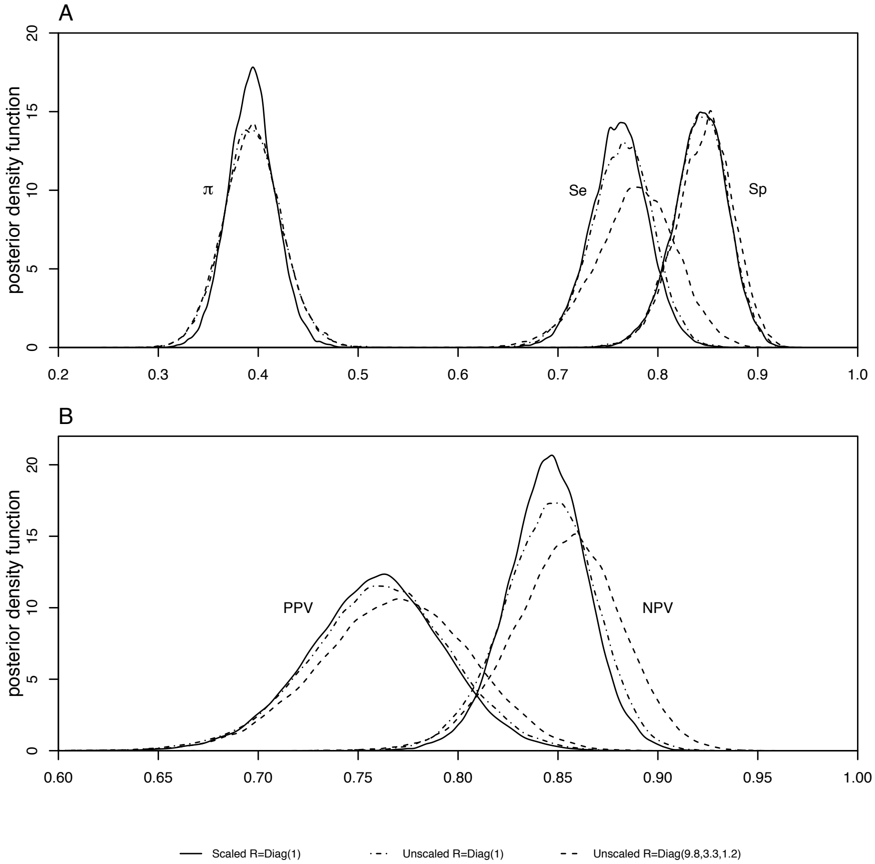
unscaled Wishart priors: the identity matrix and a diagonal matrix with elements as $(9.8, 3.3, 1.2)$. The fitted results are shown in Table 3.3 under unscaled methods. It shows that different priors have little impact on the posterior median Sp or π estimates.

To visually study the impact of different priors on posterior estimates, panel A of Figure 3.3 plots posterior densities of Se, Sp and π and panel B of Figure 3.3 plots posterior densities of PPV and NPV under different prior assumptions. Figure 3.3 shows that different priors have little impact on the posterior Sp or π estimates. The unscaled R = diag(9.8, 3.3, 1.2) prior gives negligibly larger Se, PPV and NPV posterior estimates than the other two priors. The small impact of prior assumption is consistent with intuition and the literature. For example, Lambert et al. pointed out that in a univariate setting that relatively large study sizes (15 or 30 in their simulation settings) would be less influenced by the prior of the scale parameter than small study size (5 in their simulation settings) [75].

An alternative maximum likelihood (MLE) approach

A referee has suggested considering a frequentist MLE approach as an alternative to obtain parameter estimates. Simulation studies comparing the Bayesian and MLE approaches are available in the literature [18]. We present here the estimates of MRI meta-analysis study via MLE approach, which was carried out by SAS NLMIXED procedure. The median estimate (95% confidence interval) is 0.39 (0.30, 0.45) for disease prevalence, 0.77 (0.70, 0.83) for sensitiviy and 0.85 (0.80, 0.90) for specificity. The bivariate GLMM [5, 10, 4] ignoring partial verification was also fitted via SAS NLMIXED procedure, where sensitivity is estimated to be 0.72 (0.66, 0.79) and specificity is estimated to be 0.87 (0.82, 0.92). The estimates are close to our posterior estimates from model 1 and model 3.2 via the Bayesian approach (Table 3.3). The summary receiver operating characteristic (SROC) curves was first proposed by Moses et al. [54] to reflect the trade-off between sensitivity and specificity caused by implicit thresholds and bigger area under curve (AUC) suggests better test performance. SROC curves using the MLE estimates from the hybrid GLMM and the bivariate GLMM approaches are plotted for comparison [2, 11, 45] (Figure 3.2). AUC are estimated to be 0.83 and 0.81 from the hybrid GLMM and the bivariate GLMM, respectively. The posterior Se and Sp estimates and AUC estimates from the hybrid GLMM and the bivariate GLMM ignoring

Figure 3.3: Density plots of posterior estimates of the meta-analysis of gadolinium-enhanced MRI in detecting lymph node metastases under different prior assumptions. Panel A plots posterior densities of Se, Sp and prevalence (π). Panel B plots posteriors densities of PPV and NPV.



partial verification are different, indicating that ignoring partial verification can lead to different conclusions on test accuracy. Thus, it is important to account for partial verification in a meta-analysis of diagnostic tests.

3.3.2 Meta-analysis of adrenal fluorine-18 fluorodeoxyglucose (FDG) positron emission tomography (PET) in characterizing adrenal masses

Boland et al. conducted a systematic review and meta-analysis of 21 cohort studies about test accuracy of FDG-PET in characterizing adrenal masses[76]. The reference standard tests used in the 21 cohort studies include surgery, percutaneous biopsy and follow-up CT. FDG-PET is concluded to be highly accurate in detecting and differentiating malignant adrenal disease. The authors applied the bivarate random effects model and reported that the mean sensitivity, specificity of FDG-PET are estimated to be 0.97 (95% confidence interval: 0.93, 0.98) and 0.91 (95% confidence interval: 0.87, 0.94), respectively[76]. However, the authors evaluated the methodologic quality of the included studies by the QUADAS criterias and 18 out of the 21 studies were at risk of partial verification bias. Among the 18 studies with missing counts, we were able to extract the total missing counts for 8 studies from the original papers. The cell counts of each study are reported in Table A.2 of Appendix A. Again, we impose a strong assumption on studies with only total missing counts available that the missing subjects were all tested negative by FDG-PET. We make this assumption here to creat a violation of the missing completely at random situation to show difference in estimates from the hybrid GLMM and from model 3.2. Under this assumption, sensitivity estimates will be conservative. Again, in practice, missing values should be recovered as much as possible and studies with missing values should be discarded to avoid bias.

We fit this data by the hybrid GLMM and model 3.2. In both models we use the same priors and number of posterior samples as in the meta-analysis of MRI data (section 3.3.1). We do not fit this example by model 3.3, because all the included studies in this meta-analysis are cohort studies. The estimates of interesting parameters are presented in Table 3.4. The hybrid GLMM estimates the overall median (95% CI) sensitivity, specificity and prevalence as 0.94 (95% CI: 0.91, 0.97), 0.93 (95% CI: 0.90, 0.95) and 0.39 (95% CI: 0.31, 0.47), respectively. The overall sensitivity, specificity

Table 3.4: Median estimates and 95% CI for meta-analysis of FDG PET: comparing the hybrid GLMM and model 3.2 where partial verification is ignored

Parameter	Hybrid GLMM	Model 3.2
π	0.39 (0.31, 0.47)	0.45 (0.37, 0.53)
σ_ε	0.68 (0.47, 1.01)	0.63 (0.43, 0.95)
Se	0.94 (0.91, 0.97)	0.96 (0.93, 0.98)
σ_μ	0.68 (0.23, 1.51)	0.71 (0.23, 1.54)
Sp	0.93 (0.90, 0.95)	0.90 (0.87, 0.94)
σ_ν	0.54 (0.22, 1.08)	0.51 (0.22, 1)
$\rho_{\mu\nu}$	0.78 (-0.37, 0.97)	0.80 (-0.28, 0.97)
$\rho_{\varepsilon\mu}$	-0.07 (-0.76, 0.74)	-0.05 (-0.80, 0.73)
$\rho_{\varepsilon\nu}$	-0.46 (-0.89, 0.37)	-0.31 (-0.85, 0.49)
NPV	0.96 (0.93, 0.98)	096 (0.93, 0.98)
PPV	0.89 (0.84, 0.93)	0.89 (0.84, 0.93)
LR+	16.83 (9.94, 37.97)	21.77 (12.85, 49.75)
LR−	0.06 (0.03, 0.10)	0.05 (0.02, 0.08)

and prevalence estimates from model 3.2 are 0.96 (95% CI: 0.93, 0.98), 0.90 (95% CI: 0.87, 0.94) and 0.45 (95% CI: 0.37, 0.53), respectively. The trivariate GLMM ignoring partial verification overestimate sensitivity by 0.03, underestimate specificity by 0.03 and overestimate prevalence by 0.06. Again, this example shows that ignoring partial verification bias can give different estimates for the test accuracy parameters.

3.4 Discussion

In this chapter we proposed a hybrid Bayesian hierarchical model to combine cohort and case-control studies in meta-analysis of diagnostic tests to account for disease prevalence and to correct partial verification bias. In general, this approach improves the precision of the estimates of test accuracies and predictive values by using all available information, and can be easily applied in practice using free downloadable software R [77] and WinBUGS [70].

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

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

An important question is what is an approriate sample size for such meta-analysis? Our simulation settings assumed sample size of 10 and 30 studies and lead to nearly unbiased estimates. As we have taken a full Bayesian approach, this becomes an even more intriguing question as the needed sample size may depend on whether there are informative priors for some parameters to improve estimation. In practice, sample size of meta-analysis varies largely. Davey et al. [78] summarized that among 22,453 meta-analyses with at least two studies, the median number of studies is three and inter-quartile ranges from 2 to 6. As our hybrid GLMM is a random effects model, larger sample sizes may be needed.

In this chapter, we assume that the reference test is a gold standard. In practice, however, the reference test may be imperfect and subject to misclassification. Extensions to relax the assumption of perfect reference test are currently under investigation. In such settings, every subjects true disease status is unknown and the imperfect tests may be correlated conditional on the latent disease status, inducing additional complexity for the estimation of test performance. Effort has been devoted in this regard. For example, Chu et al. [18] talked about adjusting for missing data with imperfect reference test. Dendukuri et al. [20] proposed a Bayesian approach to access overall sensitivity and specificity under absence of gold standard assumption, extending the hierarchical summary receiver operating characteristic method by Rutter and Gatsonis [2]. Both approaches included conditional dependence between the two tests through additional covariance terms. However, restrictions on the covariance terms have to be imposed to ensure well-defined probability models.

Another assumption to be relaxed in future research is the MAR assumption. We consider the MAR assumption to be practical because in many studies whether a subject is being tested by the reference test is merely dependent on the outcome of the diagnostic test and other observed characteristics. However, in some studies such as longitudinal studies the MNAR assumption may be more appropriate. Baker [38] discussed maximum likelihood estimates for the situation with multiple tests and Kosinski and Barnhart [37] presented a general likelihood-based regression approach, based on the conditional selection model by Little [79], that can flexibly account for covariates and model different missing data mechanisms. Future development is needed to incorporate these approaches in meta-analysis settings.

Chapter 4

A trivariate meta-analysis of diagnostic studies accounting for prevalence and non-evaluable subjects: re-evaluation of the meta-analysis of coronary CT angiography studies

Chapter 3 focuses on adjusting partial verification bias caused by subjects with unverified disease status. In this chapter, we discuss the situation where subjects are non-evaluable by the index test. A recent paper proposed an intent-to-diagnose approach to handle non-evaluable index test results and discussed three alternative approaches, with an application to the meta-analysis of coronary CT angiography diagnostic accuracy studies [41]. For ease of presentation, we name the three alternative approaches as Model 4.1 (non-evaluable subjects are excluded from the study), Model 4.2 (non-evaluable diseased subjects are taken as true positives and non-diseased subjects are taken as false positives) and Model 4.3 (non-evaluable diseased subjects are taken as false negatives and non-diseased subjects are taken as true negatives). In this chapter we propose an extended TGLMM [80] to handle non-evaluable index test results in Section 4.1. The performance of the intent-to-diagnose approach, the three alternative approaches and the extended TGLMM approach is examined by extensive simulation studies in Section 4.2. The meta-analysis of coronary CT angiography diagnostic accuracy studies is re-evaluated by the extended TGLMM in Section 4.3. Finally, we conclude this chapter with some discussions in Section 4.4.

4.1 Methods

We generalize the TGLMM approach to account for missing index test outcomes by extending the "classic" 2×2 table (Table 2.2) to Table 4.1, using same notations in Chapter 3. Each cell in Table 4.1 reports the cell count and cell probability corresponding to a combination of index test and disease outcomes in study i. The missing probabilities and disease prevalence are incorporated in the cell probabilities in Table 4.1.

Table 4.1: 3 × 2 table accounting for prevalence and missing index test results.

	Gold standard		
Index Test	+	−	Total
+	n_{i11} $(1-\omega_{im1})\pi_i Se_i$	n_{i10} $(1-\omega_{im0})(1-\pi_i)(1-Sp_i)$	n_{i1+} $(1-\omega_{im1})\pi_i Se_i + (1-\omega_{im0})(1-\pi_i)(1-Sp_i)$
−	n_{i01} $(1-\omega_{im1})\pi_i(1-Se_i)$	n_{i00} $(1-\omega_{im0})(1-\pi_i)Sp_i$	n_{i0+} $(1-\omega_{im1})\pi_i(1-Se_i) + (1-\omega_{im0})(1-\pi_i)Sp_i$
Missing	n_{im1} $\omega_{im1}\pi_i$	n_{im0} $\omega_{im0}(1-\pi_i)$	n_{im+} $\omega_{im1}\pi_i + \omega_{im0}(1-\pi_i)$
Total	n_{i+1} π_i	n_{i+0} $1-\pi_i$	n_{i++} 1

Each cell reports the cell count and cell probability corresponding to a combination of index test and disease outcomes in study i.

n_{itd} denotes the cell counts in study i with index test outcome $T = t$ and reference test outcome $Dis = d$, where $t = 1, 0, m$ stands for positive, negative and missing, and $d = 1, 0$ denotes positive and negative.

Se_i, Sp_i and π_i are sensitivity, specificity and prevalence of study i, respectively. ω_{imd} denotes the missing probability of index test given disease status d in study i.

Assuming a multinomial distribution, the likelihood for $\boldsymbol{\theta}_i = (Se_i, Sp_i, \pi_i)$ and $\boldsymbol{\omega}_i = (\omega_{im1}, \omega_{im0})$ given data (cell counts) is:

$$\begin{aligned} L(\boldsymbol{\theta}_i, \boldsymbol{\omega}_i|\text{Data}) \propto\ & \{(1-\omega_{im1})\pi_i Se_i\}^{n_{i11}}\{(1-\omega_{im0})(1-\pi_i)(1-Sp_i)\}^{n_{i10}} \\ & \{(1-\omega_{im1})\pi_i(1-Se_i)\}^{n_{i01}}\{(1-\omega_{im0})(1-\pi_i)Sp_i\}^{n_{i00}} \\ & (\pi_i\omega_{im1})^{n_{im1}}\{(1-\pi_i)\omega_{im0}\}^{n_{im0}} \end{aligned} \tag{4.1}$$

It is straight forward to tell from (4.1) that $L(\boldsymbol{\theta}_i, \boldsymbol{\omega}_i|\text{Data}) \propto \text{L}(\boldsymbol{\theta}_\text{i}|\text{Data}) \times \text{L}(\boldsymbol{\omega}_\text{i}|\text{Data})$, where the log-likelihood of $\boldsymbol{\theta}_i$ is:

$$\begin{aligned} \text{log}L(\boldsymbol{\theta}_i|\text{Data}) =\ & n_{i11}\{\log(\pi_i)+\log(Se_i)\} + n_{i10}\{\log(1-\pi_i)+\log(1-Sp_i)\} \\ & + n_{i01}\{\log(\pi_i)+\log(1-Se_i)\} + n_{i00}\{\log(1-\pi_i)+\log(Sp_i)\} \\ & + n_{im1}\log(\pi_i) + n_{im0}\log(1-\pi_i) \end{aligned}$$

Let $\boldsymbol{\theta} = \{\boldsymbol{\theta}_i\}$. Assuming independence among studies conditional on $\boldsymbol{\theta}_i$, the total log likelihood of $\boldsymbol{\theta}$ is:

$$\text{log}L(\boldsymbol{\theta}|\text{Data}) = \sum_{i=1}^{N} \text{log}L(\boldsymbol{\theta}_i|\text{Data}) \tag{4.2}$$

Let $\text{logit}(\pi_i) = \eta + \varepsilon_i$, $\text{logit}(Se_i) = \alpha + \mu_i$ and $\text{logit}(Sp_i) = \beta + \nu_i$. (η, α, β) are the fixed effect parameters such that median π, Se and Sp can be approximated as $\text{logit}^{-1}(\hat{\eta})$, $\text{logit}^{-1}(\hat{\alpha})$ and $\text{logit}^{-1}(\hat{\beta})$, respectively. The random effect vector $(\varepsilon_i, \mu_i, \nu_i)$ is assumed to be trivariate normally distributed:

$$(\varepsilon_i, \mu_i, \nu_i)^T \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\varepsilon^2 & \rho_{\varepsilon\mu}\sigma_\mu\sigma_\varepsilon & \rho_{\varepsilon\nu}\sigma_\nu\sigma_\varepsilon \\ & \sigma_\mu^2 & \rho_{\mu\nu}\sigma_\nu\sigma_\mu \\ & & \sigma_\nu^2 \end{bmatrix},$$

where the diagonal elements in $\boldsymbol{\Sigma}$ account for between-study variations of π, Se and Sp and the off-diagnonal elements take care of potential correlations among the three parameters.

Median PPV, NPV, LR+ and LR− and median AUC (AUC_M) can be approximated as [45]:

$$PPV = \frac{\text{logit}^{-1}(\eta)\text{logit}^{-1}(\alpha)}{\text{logit}^{-1}(\eta)\text{logit}^{-1}(\alpha) + \{1-\text{logit}^{-1}(\eta)\}\{1-\text{logit}^{-1}(\beta)\}},$$

$$NPV = \frac{\{1 - \text{logit}^{-1}(\eta)\}\text{logit}^{-1}(\beta)}{\{1 - \text{logit}^{-1}(\eta)\}\text{logit}^{-1}(\beta) + \text{logit}^{-1}(\eta)\{1 - \text{logit}^{-1}(\alpha)\}},$$

$$LR{+} = \text{logit}^{-1}(\alpha)/\{1 - \text{logit}^{-1}(\beta)\}, \quad LR{-} = \{1 - \text{logit}^{-1}(\alpha)\}/\text{logit}^{-1}(\beta),$$

$$\text{AUC}_M = \int_0^1 \text{logit}^{-1}\{(\alpha - \rho_{\mu\nu}\beta\sigma_\mu)/\sigma_\nu + \rho_{\mu\nu}\sigma_\mu/\sigma_\nu[\text{logit}(1 - \text{Sp})]\, d\text{Sp}.$$

The extended TGLMM can be fitted by standard software like SAS NLMIXED procedure, which implements an adaptive Gaussian quadrature to approximate the log-likelihood in (4.2) integrated on random effects with dual quasi-Newton optimization techniques. The NLMIXED procedure directly outputs fixed effects estimates $\hat{\eta}$, $\hat{\alpha}$ and $\hat{\beta}$ and can provide median prevalence, Se, Sp, PPV, NPV, LR+, LR− estimates and their confidence intervals through the "estimate" statements. Sample SAS code is available in the Appendix.

4.2 Simulations

4.2.1 Simulation Scenarios

We conduct simulation studies under three missing scenarios to systematically evaluate the performance of the proposed extended TGLMM approach and the approaches discussed in Schuetz et al. [41]: missing probabilities for diseased and non-diseased subjects are same (0.1), or missing probability of diseased group (0.1) is smaller than non-diseased group (0.2), or missing probability of diseased group (0.2) is larger than non-diseased group (0.1). All three scenarios satisfy the MAR assumption, and the first scenario is in fact MCAR [46]. True sensitivity and specificity are 0.7 and 0.9, disease prevalence is 0.25 and variances of Se, Sp and prevalence are 1 on logit scale. These assumptions mimic a diagnostic test with relatively low sensitivity, high specificity and a disease with moderate prevalence. A moderate positive correlation of 0.3 is assumed between Se and π, and moderate negative correlations of -0.3 are assumed between Sp and π and between Se and Sp, on logit scales. Such correlation directions were observed in some meta-analysis studies[60, 43]. Intuitively, a population with higher prevalence may have more diseased cases with clear disease symptoms, leading to increased sensitivity. Under each setting, 5000 meta-analysis data sets are simulated with 30 studies in each data set. π_i, Se_i and Sp_i for each study were generated according to the trivariate assumption described in Section 4.1. True and false positives, true and false negatives and non-evaluable counts are sampled from the multinomial distribution in Table 4.1. For each simulated meta-analysis data set, the extended TGLMM, Model 4.1-4.3 and the intent-to-diagnose approach are fitted. Bias in percentage, mean standard error (SE) and 95% confidence interval coverage probability (CP) are collected and compared for estimates of sensitivity, specificity, prevalence, PPV, NPV, LR+ and LR−. Bias in percentage is calculated by $(\hat{\delta} - \delta) \times 100/\delta$, where δ is the true value and $\hat{\delta}$ is the estimator.

4.2.2 Simulation Results

Table 4.2 shows the simulation results under different scenarios. When MCAR ($\omega_{m1} = \omega_{m0} = 0.1$), disease prevalence estimates from all five models are nearly unbiased (bias less than 1%). The extended TGLMM and Model 4.1 both give nearly unbiased estimates (bias less than 1.6%) and nominal coverage probabilities around 93% for Se, Sp, PPV, NPV, LR+ and LR− estimates. Model 4.2 over-estimates sensitivity and under-estimates specificity: bias of sensitivity estimate is 4.6% and bias of specificity estimates is 11.9%. Estimates of PPV and LR+ are more biased (22.6% bias for PPV and 49.2% bias for LR+). Using Model 4.3 sensitivities are largely under-estimated (12.6% bias) and specificities are over-estimated (1.1% bias). The intent-to-diagnose approach largely under-estimates both sensitivity and specificity (12.6% and 11.9% bias, respectively). The CPs for some estimates from Model 4.2 and 4.3 and the intent-to-diagnose approach can be as low as 0 (e.g., specificity estimates from Model 4.2), indicating that none of the confidence intervals cover the true values. When missing probability of the diseased group is smaller than the non-diseased group ($\omega_{m1} = 0.1, \omega_{m0} = 0.2$), the extended TGLMM and Model 4.1 both give nearly unbiased estimates (bias around 0.1%) of sensitivity and specificity. However, Model 4.1 over-estimates disease prevalence (9.6% bias) while the extended TGLMM gives nearly unbiased (bias within 1%) estimate of prevalence. As a consequence, Model 4.1 gives biased estimates of PPV and NPV (3.1% and 1.3%, respectively), while the extended TGLMM provides nearly unbiased estimates for all parameters (within 2%). Again, under this scenario, the intent-to-diagnose approach largely under-estimates sensitivity, specificity, PPV, NPV and LR+ and over-estimates LR−, with CPs less than 40% and some as low as 0. On the other hand, when $\omega_{m1} = 0.2$ and $\omega_{m0} = 0.1$, the extended TGLMM and Model 4.1 again give nearly unbiased estimates (bias around 0.1%) of sensitivity and specificity. Model 4.1 under-estimates disease prevalence (8.4% bias) while the extended TGLMM provides nearly unbiased estimates. The intent-to-diagnose approach largely under-estimates sensitivity, specificity, PPV, NPV and LR+ and over-estimates LR− and some CPs are as low as 0. When the missing probabilities for diseased and non-diseased subjects are more unbalanced, we expect the estimates from Model 4.1-4.3 and the intent-to-diagnose approach to have larger bias and smaller CP. In practice, however, depending on the test performance and missing probabilities, the direction and magnitude of the bias from the four approaches discussed in Schuetz et al. [41] can be different from what we observed in these simulation studies.

Table 4.2: Simulation results under three scenarios of MAR assumption: equal or unequal missing probabilities for the diseased and non-diseased groups.

Model	TGLMM			Model 4.1			Model 4.2			Model 4.3			Intent-to-diagnose		
Estimate	Bias%	meanSE	CP	Bias%	meanSE	CP	Bias%	meanSE	CP	Bias%	meanSE	CP	Bias%	meanSE	CP
								$\omega_{m1} = \omega_{m0} = 0.1$							
Se	-0.3	0.041	0.94	-0.3	0.041	0.94	4.6	0.036	0.81	-12.6	0.037	0.33	-12.6	0.036	0.33
Sp	-0.1	0.017	0.93	-0.1	0.017	0.93	-11.9	0.018	0	1.1	0.015	0.84	-11.9	0.017	0
π	0.8	0.034	0.93	0.8	0.034	0.93	0.8	0.034	0.93	0.8	0.034	0.93	0.8	0.034	0.93
PPV	-0.1	0.046	0.94	-0.3	0.046	0.94	-22.6	0.047	0.08	-0.9	0.046	0.94	-29	0.049	0.01
NPV	-0.1	0.018	0.93	-0.1	0.018	0.93	-0.2	0.018	0.93	-2.9	0.020	0.81	-4.6	0.022	0.59
LR+	1.6	1.188	0.92	1.6	1.189	0.93	-49.2	0.307	0	-0.5	1.160	0.92	-57.6	0.271	0
LR$-$	0.9	0.044	0.94	0.9	0.044	0.94	1.5	0.044	0.94	27.9	0.039	0.33	46.8	0.045	0.04
								$\omega_{m1} = 0.1, \omega_{m0} = 0.2$							
Se	-0.1	0.041	0.94	-0.1	0.041	0.94	4.7	0.036	0.80	-12.3	0.036	0.34	-12.3	0.036	0.34
Sp	-0.1	0.017	0.94	-0.1	0.017	0.94	-22.3	0.017	0	2.2	0.014	0.62	-22.3	0.017	0
π	0.4	0.034	0.93	9.6	0.036	0.90	0.4	0.034	0.93	0.4	0.034	0.93	0.4	0.034	0.93
PPV	-0.3	0.046	0.93	3.1	0.044	0.88	-36	0.047	0	2.7	0.044	0.89	-42.1	0.047	0
NPV	-0.1	0.018	0.94	-1.3	0.020	0.93	-1.4	0.020	0.92	-2.7	0.020	0.83	-6.3	0.024	0.36
LR+	1.4	1.195	0.94	1.4	1.194	0.94	-65.1	0.159	0	12.3	1.312	0.95	-70.8	0.147	0
LR$-$	0.6	0.044	0.93	0.6	0.044	0.93	14.7	0.050	0.85	26.1	0.038	0.39	66.1	0.051	0
								$\omega_{m1} = 0.2, \omega_{m0} = 0.1$							
Se	-0.1	0.023	0.93	-0.1	0.023	0.93	8.7	0.018	0.12	-21	0.020	0	-21	0.019	0
Sp	0	0.009	0.93	0	0.009	0.93	-10.6	0.009	0	1.1	0.008	0.74	-10.6	0.009	0
Prev	0	0.018	0.93	-8.4	0.017	0.72	0	0.017	0.91	0	0.017	0.91	0	0.0168	0.89
PPV	-0.1	0.025	0.93	-3.7	0.027	0.83	-19.1	0.025	0	-4	0.026	0.8	-30.6	0.025	0
NPV	0	0.010	0.92	1.1	0.009	0.76	1.1	0.009	0.74	-4.6	0.011	0.05	-6.2	0.012	0
LR+	0.3	0.655	0.93	0.3	0.653	0.93	-44.1	0.196	0	-11.7	0.570	0.62	-59.3	0.154	0
LR-	0.3	0.025	0.93	0.3	0.025	0.93	-10.8	0.022	0.62	47.4	0.021	0	66.7	0.024	0

Bias in percentage(Bias%), mean standard error (meanSE) and 95% confidence interval coverage probability (CP) are summarized for estimates of Se, Sp, π, PPV, NPV, LR+ and LR$-$ from different models.

4.3 Re-evaluation of the meta-analysis of coronary CT angiography studies

Cardiac CT scans can be used to rule out stenoses, however, are found to be subject to non-evaluable results. Schuetz et al.[41] performed a systematic search for diagnostic accuracy studies of coronary CT angiography. The authors searched Medline, Embase and ISI Web of Science databases for prospective studies using conventional coronary angiography as the gold standard and have patients with non-evaluable CT images. Eventually, 26 studies were included that reports cell counts in a 3×2 table as Table 4.1. The authors mentioned that the 3×2 table can be extended to a 3×3 table for non-evaluable results of the gold standard, however such cases were rare (0.1%) in this systematic review. We re-evaluate the 26 studies by the extended TGLMM and compare to the estimates following the four approaches discussed in Schuetz et al.[41].

The fitted median estimates and 95% confidence intervals are reported in Table 4.3. The extended TGLMM accounting for missing subjects gives median sensitivity, specificity, LR+, LR− and AUC estimates close to the estimates when non-evaluable subjects are excluded as in Model 4.1. The median disease prevalence estimated from the extended TGLMM is slightly lower than the estimate from Model 4.1. Model 4.2 gives significantly lower specificity estimate and Model 4.3 gives lower sensitivity estimate. The intent-to-diagnose approach provides lower estimates for sensitivity, specificity and AUC as it is the most conservative approach. Figure 4.1 presents the estimated PPV and NPV with 95% confidence bands versus prevalence, based on the overall sensitivity and specificity estimates from the extended TGLMM and the intent-to-diagnose approach. Figure 4.1 shows that as disease prevalence changes, PPV and NPV estimates from the latter approach are not ever included in the 95% confidence band of the estimates from the extended TGLMM, which suggests potential underestimation of PPV and NPV.

4.4 Conclusions

In this chapter, we propose an extended TGLMM approach to handle non-evaluable index test subjects in meta-analysis of diagnostic tests. The extended TGLMM is compared to an intent-to-diagnose approach and three alternative approaches proposed

Figure 4.1: Overall PPV and NPV plot based on the extended TGLMM (denoted by "TGLMM") and the intent-to-diagnose approach. The solid and dashed lines are the overall estimates of PPV and NPV from the extended TGLMM and the intent-to-diagnose approach corresponding to different prevalences ranging from 0 to 1, respectively. The dotted lines are the 95% confidence intervals of PPV and NPV estimates from the extended TGLMM approach. The vertical dashed line is the overall prevalence estimates from the meta-analysis of coronary CT angiography studies.

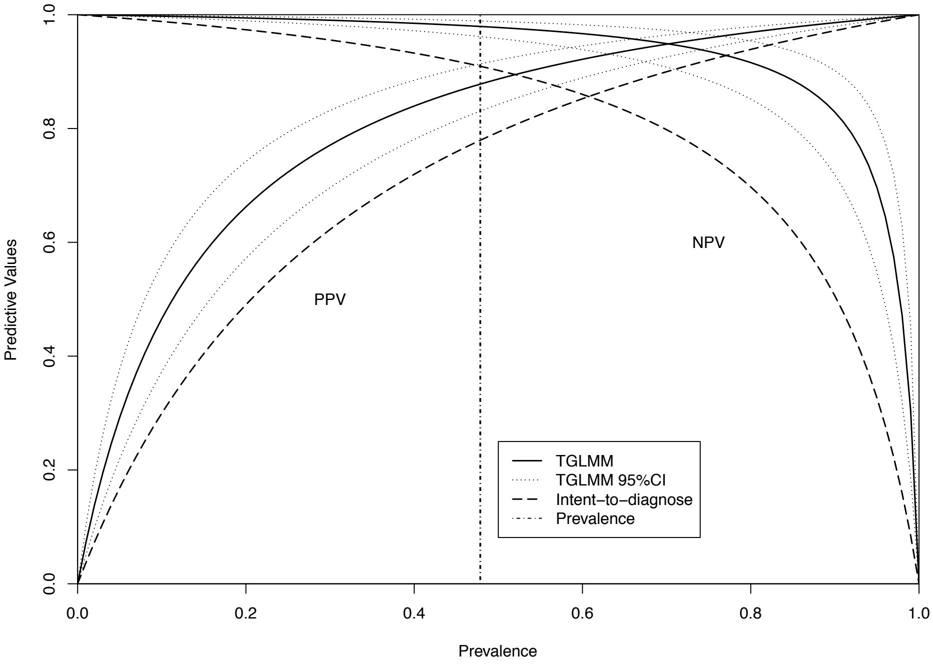


Table 4.3: Median estimates and 95% confidence intervals (in brackets) for parameter estimates using different methods

Method	Sensitivity	Specificity	Prevalence	PPV
TGLMM	98.0 (96.7,99.3)	87.5 (82.7,92.3)	47.8 (37.9,57.7)	87.8 (83.3,92.3)
Model 4.1	98.0 (96.7,99.3)	87.4 (82.5, 92.3)	49.3 (38.9,59.7)	88.4 (84,92.7)
Model 4.2	98.1 (96.9,99.3)	75.9 (69.3,82.5)	47.8 (37.9,57.8)	78.9 (71.9,85.9)
Model 4.3	91.7 (88.1,95.4)	89 (85.4,92.7)	47.8 (37.9,57.7)	88.4 (84.1,92.7)
Intent-to-diagnose	91.7 (88.1,95.3)	76.2 (69.7,82.6)	47.9 (37.9,57.9)	78 (70.2,85.7)

Method	NPV	LR+	LR−	AUC
TGLMM	97.9 (96.4,99.5)	7.8 (4.8,10.9)	0.02 (0.01,0.04)	0.99 (0.96,1)
Model 4.1	97.8 (96.1,99.4)	7.8 (4.8,10.9)	0.02 (0.01,0.04)	0.99 (0.96,1)
Model 4.2	97.8 (96.2, 99.4)	4.1 (2.9,5.2)	0.02 (0.01,0.04)	0.98 (0.97,1)
Model 4.3	92.1 (88.4,95.8)	8.4 (5.5,11.3)	0.09 (0.05,0.14)	0.96 (0.93,0.99)
Intent-to-diagnose	90.9 (86.4,95.5)	3.8 (2.7,5.0)	0.11 (0.06,0.16)	0.93 (0.89,0.96)

by Schuetz et al.[41] through simulation studies and re-evaluaion of the meta-analysis of coronary CT angiography studies.

In summary, by simulation studies we showed that under MAR assumption, excluding index test non-evaluable subjects (Model 4.1) will not lead to biased estimates of sensitivity, specificity, LR+, LR− and AUC. Thus in practice, researchers can be confident to apply Model 4.1 when there is a belief in the MAR assumption. However, when disease prevalence or PPV and NPV are of interest, excluding non-evaluable subjects could lead to biased estimates of these parameters. Under this situation, the extended TGLMM accounting for missingness should be preferred. Even though the extended TGLMM is more theoretically complex than the widely used bivariate random effects model, it is easy to program use SAS NLMIXED procedure. Model 4.2, Model 4.3 and the intent-to-diagnose approach all largely under- or over- estimate sensitivity and specificity, so that they should not be recommended when MAR assumption is not seriously violated.

Adequate reporting of the missing outcomes in study reports is essential to apply the discussed models. As shown in the simulation studies, different missing scenarios can have different impact on how estimates are biased and more importantly, missing mechanism can indicate whether the MAR assumption holds. When the MAR assumption is violated, i.e., the probability of non-evaluation depends on unobserved index test outcomes, the direction and magnitude of bias are hard to predict. Few sensitivity analysis methods using pattern mixture models and selection models are available for this scenario[81, 82]. These approaches can be explored in further research. On the other hand, number of non-evaluable results need to be known in order to apply the proposed methods. However, a recent study shows that they are not consistently or adequately reported in published studies [83].

A reviewer has pointed out that as long as number of non-evaluable subjects are known, disease prevalence can be estimated unbiasedly through an univariate meta-analysis. Consequently, together with unbiased sensitivity and specificity estimates, PPV and NPV estimates are unbiased too. This approach is a simpler method than the proposed extended TGLMM to estimate prevalence, however, can be less efficient by ignoring the potential correlation between prevalence, sensitivity and specificity, which may result in wider confidence intervals.

For an individual patient, different approaches of treating a missing result can have different impact. For example, if index test results are missing due to the same reason of returning a negative result (and thus is MNAR), then treating such patients as disease negatives can yield unbiased estimate of prevalence for a study, and also won't affect the patients' diagnosis. On the contrary, if index test missing patients are treated as positives for reasons such as suspicious of serious disease like cancer [84], it may result in over-estimation of disease prevalence and unnecessary medial cost for the patient. For another example, if index test is repeatable and repeated for subjects with non-evaluable results, then it is appropriate to ignore missing results.

Chapter 5

A Bayesian Hierarchical Model for Network Meta-analysis of Diagnostic Tests

In this chapter, we extend our discussion from NMA to NMA-DT and develop a NMA-DT framework from the perspective of missing data analysis to address the challenges discussed in Section 1.2. The proposed framework is motivated from the literature on network meta-analysis of randomized clinical trials (NMA-CT), which extends the scope of traditional pairwise meta-analysis by synthesizing both direct and indirect comparisons of multiple treatments across randomized controlled trials [85, 86, 87, 88, 89, 90]. 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 include only a subset of the whole set of tests. The test outcomes from non-included tests must be considered as missing data. By simultaneously comparing all candidate tests and the 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. This effectively handles four critical challenges in the traditional MA-DT [54, 91, 61, 8] by: 1) combining information from studies with all three designs; 2) pooling both studies with or without a gold standard;

3) allowing different sets of candidate tests in different studies, or different subsets of subjects within a study; and 4) accounting for potential heterogeneity across studies, due to differences in study populations, design and lab technical issues, or complex correlation structures among multiple diagnostic tests.

In the rest of this chapter, we start by introduing two motivating case studies in Section 5.1: a NMA of deep vein thrombosis tests, and a NMA of latent tuberculosis tests. We then present the proposed NMA-DT model, and the Bayesian inference method in Section 5.2 and apply the proposed method to the two motivating studies in Section 5.3. In Section 5.4, the performance of the proposed method is evaluated through simulation studies, and compared to the naive separate MA-DT approach. Finally, Section 5.5 provides a brief discussion.

5.1 Motivating studies

5.1.1 NMA of deep vein thrombosis (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 a pulmonary embolus, and possibly resulting in death [92]. Consequently, correct diagnosis of DVT plays an important role in its early-detection and treatment. The gold standard diagnostic test for DVT, contrast venography, is an invasive procedure and can introduce allergic reactions [93]. Therefore, ultrasonography is a commonly used surrogate test, because it is non-invasive 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 can also be used to diagnose DVT. Moreover, the test for D-dimer concentration can be easily performed in a screening blood test. Several studies have shown that D-dimer has high sensitivity, high negative predictive value (NPV), and moderate specificity [94].

A recent paper by Kang et al.[51] presented a meta-analysis that included 12 studies comparing the accuracy of diagnostic tests for DVT. Among the 12 studies, four studies compared D-dimer test to venography, three studies compared ultrasonography to venography, and five studies compared the D-dimer to ultrasonography [51]. None of the studies compared the three tests together. Kang et al.[51] applied a mixed-effects log-linear model, with random effects 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 (e.g. sensitivity and specificity) made it difficult to interpret the model parameters, and hard to generalize to comparing more tests.

5.1.2 NMA of latent tuberculosis (TB) tests

Tuberculosis is a fatal disease that causes two million deaths per year globally [95]. The tuberculin skin test (TST) have been used to detect latent TB for many years, but has been found to have low and highly variable specificity [48]. Two interferon-γ-release assays, QuantiFERON-TB gold (QFG) and T-SPOT.TB (TSPOT), are now available, and they are considered to be attractive alternative tests with operational advantages and possibly increased specificity.

Sadatsafavi et al.[19] studied the performance of the three candidate tests for diagnosing latent TB, namely QFG, TSPOT and TST, through a meta-analysis of 22 studies. Among them, two studies compared the three tests on the same group of participants, one study compared TSPOT to TST, and the rest of the studies compared QFG to TST. Cross-tabulated results comparing the three tests or subsets of the three tests were reported [19]. In this NMA-DT dataset, none of the studies included the gold standard test and some studies compared three tests together. Sadatsafavi et al. [19] used latent class random-effects models to account for varying test performances and prevalences across studies, where only one random effect for sensitivity or specificity of one of the tests can be included. In addition, potential correlations between disease prevalence and the test accuracy parameters were not taken into account.

5.2 A unified statistical framework

We present a Bayesian hierarchical NMA-DT model to compare multiple tests simultaneously. In this paper, we focus on modeling a commonly used pair of test accuracy indicies, Se and Sp, where sensitivity(Se) is the probability of a candidate test being positive given a diseased subject, and specificity(Sp) is the probability of a candidate test being negative given non-diseased status[1]. In addition, disease prevalence is also modeled such that by estimating the overall prevalence, other test accuracy indicies such as PPV and NPV can be calculated. In this section, we first present the model given the random effects, and then describe the distributions of the random effects and prior distributions of the fixed effects.

5.2.1 Likelihood function

We view different studies as if they were all potentially 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 [46]. 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 in our case means that missingness is independent of sensitivity and specificity [46].

Let $T = \{T_0, T_1, \ldots, T_K\}$ be a set of $K+1$ binary diagnostic tests, where T_0 denotes the gold standard and $T_1, \ldots, T_K$ stand for the candidate tests under evaluation. Suppose we have a collection of $i = 1, \ldots, N$ studies, where each reports outcomes of tests in a subset of T. In the ith study, 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 on the jth subject, and 0 if not). Let π_i be the study-specific disease prevalence: $\pi_i = P(y_{ij0} = 1)$, $i = 1, \ldots, N$. Let Se_{ik} and Sp_{ik} denote the study specific sensitivity and specificity for the kth test ($k = 1, \ldots, K$), 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 tests conducted on subject j in the ith study, and $\boldsymbol{y}_{ij} = \{y_{ijk} : k \in K_{ij}\}$ be the collection of test outcomes for this subject. We note that $\boldsymbol{y}_{ij}$ can be equivalently written as $\boldsymbol{y}_{ij} = \{\delta_{ijk} y_{ijk} : k = 1, \ldots, K\}$.

To derive the likelihood for the jth subject in the ith study, first consider a subject that is tested by the gold standard test ($\delta_{ij0} = 1$) so that the true disease status is known. Conditional independence is assumed such that candidate test results are independent given disease status. This assumption has been widely used in latent class models assessing accuracy of diagnostic tests without a gold standard [17, 18]. Thus, the likelihood of the test outcomes for a diseased subject is given by

$$P(\boldsymbol{y}_{ij}, y_{ij0} = 1) = P(y_{ij0} = 1) P(\boldsymbol{y}_{ij}|y_{ij0} = 1) = \pi_i \prod_{k \in K_{ij}} (Se_{ik})^{y_{ijk}} (1 - Se_{ik})^{(1-y_{ijk})} = \pi_i h_{ij1},$$

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

$$P(\boldsymbol{y}_{ij}, y_{ij0} = 0) = P(y_{ij0} = 0)P(\boldsymbol{y}_{ij}|y_{ij0} = 0) = (1-\pi_i) \prod_{k \in K_{ij}} (Sp_{ik})^{(1-y_{ijk})}(1 - Sp_{ik})^{y_{ijk}} = (1-\pi_i)h_{ij0},$$

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

Second, for subject j who is not tested by the gold standard, the likelihood of the test outcomes is given by

$$P(\boldsymbol{y}_{ij}) = P(\boldsymbol{y}_{ij}, y_{ij0} = 1) + P(\boldsymbol{y}_{ij}, y_{ij0} = 0) = \pi_i h_{ij1} + (1 - \pi_i)h_{ij0}.$$

In general, the likelihood of test outcomes for subject j (tested by the gold standard or not, indicated by δ_{ij0}) can be written as

$$P(\boldsymbol{y}_{ij}) = (\pi_i h_{ij1})^{\delta_{ij0} y_{ij0}} [(1 - \pi_i)h_{ij0}]^{\delta_{ij0}(1-y_{ij0})} [\pi_i h_{ij1} + (1 - \pi_i)h_{ij0}]^{(1-\delta_{ij0})}. \quad (5.1)$$

5.2.2 Random effects and prior specifications

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

$$\pi_i = \Phi(\eta + \varepsilon_i),\ Se_{ik} = \Phi(\alpha_k + \mu_{ik}), \text{ and } Sp_{ik} = \Phi(\beta_k + \nu_{ik}),\ i = 1, \ldots, N,\ k = 1, \ldots, K, \quad (5.2)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function for probit transformations. The parameter η is the fixed effect for prevalence, and α_k and β_k are the fixed effects for sensitivity and specificity of T_k, respectively. Median disease prevalence, sensitivity and specificity of T_k can be estimated as $\pi = \Phi(\eta)$, $Se_k = \Phi(\alpha_k)$ and $Sp_k = \Phi(\beta_k)$, respectively. The random effects ε_i, μ_{ik} and ν_{ik} are the study-specific effects for prevalence, sensitivity and specificity of T_k, respectively. It is straightforward to incorporate meta-regression covariates in equation (5.2) as

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

for $i = 1, \ldots, N$ and $k = 1, \ldots, K$, where $\boldsymbol{X}_i$, $\boldsymbol{W}_i$ and $\boldsymbol{Z}_i$ 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 [21, 16]:

$$(\varepsilon_i, \mu_{i1}, \nu_{i1}, \dots, \mu_{iK}, \nu_{iK})^T \sim MVN(\mathbf{0}, \mathbf{\Sigma}),\ i = 1, \dots, N.$$

The covariance matrix $\mathbf{\Sigma}$ can be written as $\mathbf{\Sigma} = S\Omega S$, 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 whose off-diagonal elements measure potential correlations among disease prevalence and the test accuracy parameters. We assume the same correlation structure for all studies (with missing tests or not). Therefore, studies reporting all test outcomes of T contribute to estimating $\mathbf{\Sigma}$, and studies with missing test outcomes directly contribute to estimating a submatrix of $\mathbf{\Sigma}$. Furthermore, missing test accuracies can be predicted from $\mathbf{\Sigma}$. 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 correct inferences when comparing the test performances.

A conjugate Wishart prior is assumed for the precision matrix: $\mathbf{\Sigma}^{-1} \sim \text{Wishart}(\boldsymbol{R}, v)$. Taking the degrees of freedom v equal to the dimension of $\mathbf{\Sigma}$, $2K+1$, will assign an approximately uniform prior on the correlation coefficients. Different choices of $\boldsymbol{R}$ can give relatively informative or non-informative priors on the variance parameters; specific choices of $\boldsymbol{R}$ are discussed in the case studies. Non-informative normal priors ($N(0, 10)$) are assumed for η, α_k and β_k ($k = 1, \dots, K$), which correspond to 95% prior credible intervals (CI) of approximately (0,1) for π, Se_k and Sp_k, $k = 1, \dots, K$.

5.2.3 Model implementation and a Bayesian ranking procedure

We use the JAGS software via the rjags package in R to sample from the joint posterior distribution using Markov chain Monte Carlo (MCMC) methods [70, 96]. The posterior samples are drawn by Gibbs and Metropolis-Hastings algorithms. Naturally, the posterior estimates are similar to the MLEs when the priors are non-informative. On the other hand, the Bayesian approach allows for full posterior inference, so that asymptotic approximations are not required. Posterior samples of median disease prevalence, sensitivity and specificity of T_k can be achieved by transformation of the MCMC samples: $\pi = \Phi(\eta)$, $Se_k = \Phi(\alpha_k)$ and $Sp_k = \Phi(\beta_k)$. Posterior samples of PPV, NPV, LR+ and LR− of T_k can also be obtained from the MCMC samples of sensitivities, specificities and prevalences:

$$PPV_{ik} = \frac{Se_{ik}\pi_i}{Se_{ik}\pi_i + (1 - Sp_{ik})(1 - \pi_i)}, \quad NPV_{ik} = \frac{Sp_{ik}(1 - \pi_i)}{Sp_{ik}(1 - \pi_i) + (1 - Se_{ik})\pi_i},$$

$$LR{+}_{ik} = \frac{Se_{ik}}{1 - Sp_{ik}}, \quad \text{and } LR{-}_{ik} = \frac{1 - Se_{ik}}{Sp_{ik}}.$$

In NMA-DT, the Bayesian approach can also facilitate ranking the tests using a Bayesian ranking procedure based on the posterior samples. Specifically, the “best” tests can be identified by calculating the posterior probability of T_k being among the best C tests (not counting T_0):

$$Pr(\text{rank}(Se_k) \leq C, \text{rank}(Sp_k) \leq C|\text{Data}), C = 1, \ldots, K. \tag{5.3}$$

5.2.4 A measure of inconsistency

The NMA-DT model relies on an important “consistency” assumption, which assumes that candidate tests would have been performed consistently on subjects assigned and not assigned to the test. However, inconsistency could occur, for example when, studies that do not include T_1 include a population for which T_1 is inappropriate, and hence whose performance may differ systematically from studies that do include T_1. In this situation, the MAR assumption is questionable, and borrowing information from studies that do not provide direct estimates of test accuracies must be done with caution. The concern of inconsistency is also discussed in contrast-based NMA methods [86, 87], wherein indirect evidence may be inconsistent with direct evidence. Lu and Ades [87] 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 iti is restricted to relative effects (e.g., log odds ratios), while NMA-DT estimates marginal test accuracies (e.g., Se and Sp). A new measurement is proposed for arm-based NMA methods [97].

For T_k, to measure inconsistency in sensitivity (specificity), we propose to calculate the discrepancy between posterior study-specific sensitivity (specificity) estimates in observed studies and unobserved studies, i.e.,

$$IC_k^{Se} = \sum_{k \in K_{ij}} Se_{ik} \Big/ \sum I(k \in K_{ij}) - \sum_{k \notin K_{ij}} Se_{ik} \Big/ \sum \{1 - I(k \in K_{ij})\},$$

$$IC_k^{Sp} = \sum_{k \in K_{ij}} Sp_{ik} \Big/ \sum I(k \in K_{ij}) - \sum_{k \notin K_{ij}} Sp_{ik} \Big/ \sum \{1 - I(k \in K_{ij})\}, \tag{5.4}$$

Posterior estimates of IC_k^{Se} and IC_k^{Sp} can be calculated from MCMC posterior samples and their 95% CIs can be used to assess whether they differ from 0, which would indicate significantly inconsistent test performance between studies with and without T_k outcomes.

5.3 Case study results and sensitivity analyses

5.3.1 NMA of DVT tests

We analyze the NMA of DVT tests in Section 5.1.1 by the proposed NMA-DT model. In this study, we have $K = 2$. We adopt a moderately informative Wishart prior with $v = 5$ and $\boldsymbol{R}$ 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 $(\sigma_\pi, \sigma_{Se_1}, \sigma_{Sp_1}, \sigma_{Se_2}, \sigma_{Sp_2})$. We fit the model by assuming vague $N(0, 10)$ priors for η, α_k and β_k.

After 10,000 burn-in samples, 1,000,000 posterior samples were obtained. Table 5.1 shows the results from the proposed NMA-DT model under the "all studies" column. Figure 5.1 plots posterior distributions and study-specific posterior medians and 95% CIs for the prevalence, sensitivity, and specificity parameters. We write posterior medians followed by 95% CI in brackets for the rest of this paper. The NMA-DT model concludes that ultrasonography has a median Se of 0.90 (0.77, 0.96) and a 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). From equation (5.3), the posterior probabilities of ultrasonography ranking first in terms of sensitivity is 0.84, and the D-dimer test ranking first in terms of specificity is 0.74. The posterior probability that ultrasonography (D-dimer) ranks first in terms of both sensitivity and specificity is 0.20 (0.13), respectively. Overall, ultrasonography is favored in detecting

Table 5.1: Meta-analysis of DVT tests: median estimates and 95% CIs. Estimates from models using all studies under "All studies" are compared to estimates excluding an "outlier" study 5 under "Without outlier".

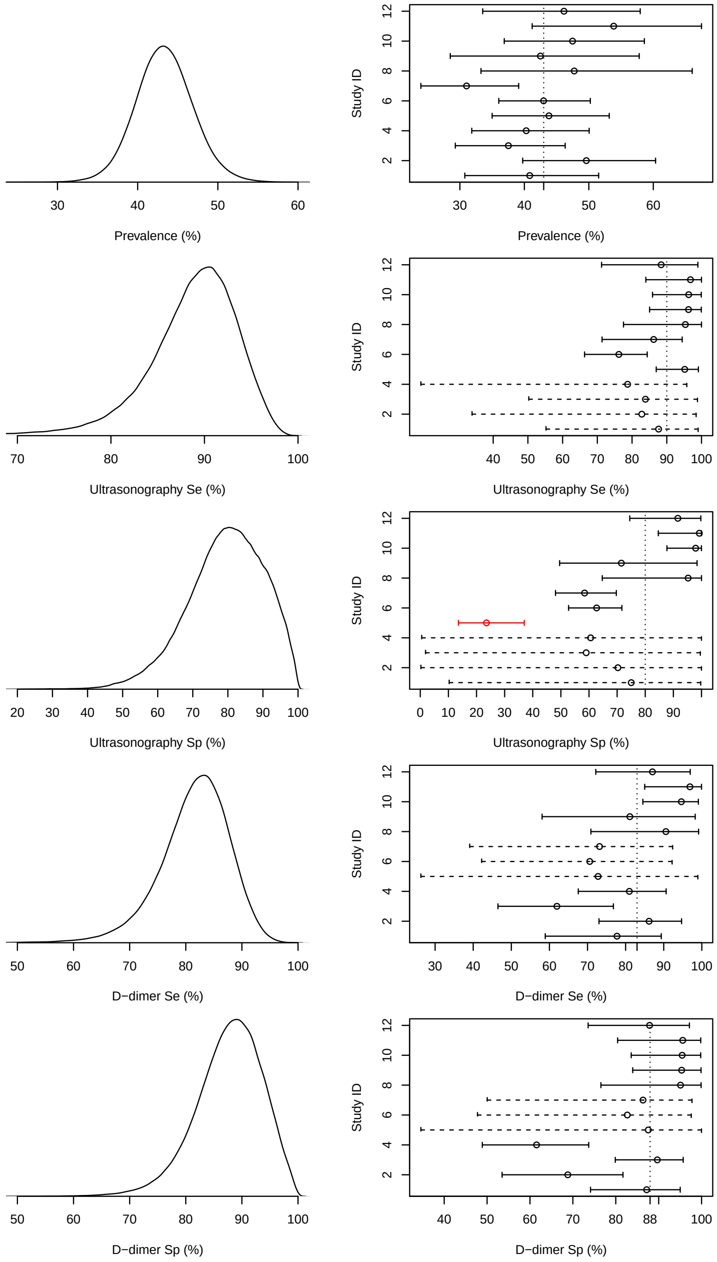
	All studies	Without outlier
Prevalence	0.43 (0.36, 0.50)	0.43 (0.35, 0.51)
	Ultrasonography	
Sensitivity	0.90 (0.77, 0.96)	0.88 (0.74, 0.96)
Specificity	0.80 (0.54, 0.97)	0.83 (0.63, 0.97)
PPV	0.84 (0.68, 0.96)	0.80 (0.58, 0.96)
NPV	0.91 (0.80, 0.97)	0.90 (0.79, 0.97)
LR+	4.39 (1.89, 27.90)	5.22 (2.19, 29.44)
LR−	0.13 (0.05, 0.33)	0.15 (0.05, 0.36)
	D-dimer	
Sensitivity	0.83 (0.68, 0.92)	0.82 (0.67, 0.91)
Specificity	0.88 (0.75, 0.97)	0.87 (0.75, 0.97)
PPV	0.84 (0.68 0.96)	0.83 (0.67, 0.96)
NPV	0.87 (0.77, 0.94)	0.86 (0.77, 0.93)
LR+	7.00 (3.10, 33.49)	6.43 (3.04, 28.4)
LR−	0.20 (0.09, 0.38)	0.22 (0.10, 0.40)

disease status with higher sensitivity, whereas D-dimer performs better in ruling out the non-diseased with higher specificity. The posterior inconsistency measurements are 0.13 (−0.06, 0.45), 0.04 (−0.06, 0.34), 0.06 (−0.12, 0.42) and 0.07 (−0.2, 0.45) for D-dimer Se, ultrasonography Se, D-dimer Sp and ultrasonography Sp, respectively. None of these measurements suggest the presence of significant inconsistency.

Sensitivity analyses to prior distribution on $\boldsymbol{\Sigma}^{-1}$

Sensitivity analyses to the prior distributions on $\boldsymbol{\Sigma}^{-1}$ were conducted to evaluate the effect of the prior distribution on the posterior prevalence, sensitivity and specificity. A relatively informative Wishart prior with $v = 5$ and $\boldsymbol{R}$ having diagonal elements equal to 20 and off-diagonal elements equal to 0.05 is used in this repeat analysis. This Wishart

Figure 5.1: Meta-analysis of DVT tests: posterior densities and study-specific posterior estimates. The left column plots posterior densities and the right column plots study-specific posterior medians and their 95% CIs for prevalence, sensitivities and specificities of ultrasonography and D-dimer tests. Circles represent study-specific posterior medians and solid and dashed lines denote the corresponding 95% CIs when the test is included in the study and not included (imputed by MCMC sampling), respectively. A red line indicates a potential outlier study. Dotted lines indicate overall posterior medians across studies.



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 0.89 (0.78, 0.96) and specificity 0.79 (0.54, 0.96). The D-dimer test has posterior median sensitivity 0.82 (0.67, 0.92) and specificity 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 $v = 5$ and $\boldsymbol{R}$ having 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 5.1, this prior leads to wider CIs for all parameters and slightly higher posterior medians for ultrasonography Sp.

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

Sensitivity analysis to an "outlier" study

In the right column of Figure 5.1, heterogeneity is observed in the posterior estimates across studies. For example, Study 5 has an extremely low estimate of ultrasonography Sp (red), suggesting that it may be an "outlier" study. Thus, to investigate the implications of a potential outlier, we repeat the analysis excluding Study 5. Table 5.1 summarizes the posterior estimates under column "without outlier". Compared to the analysis using all studies, when Study 5 is excluded, the estimate of ultrasonography specificity is increased from 0.80 (0.54, 0.97) to 0.83 (0.63, 0.97) and the posterior median sensitivity is lowered from 0.90 (0.77, 0.96) to 0.88 (0.74, 0.96). As a result, estimates of PPV and LR+ of ultrasonography are also changed. Estimates of disease prevalence and D-dimer sensitivity and specificity remain similar to the estimates using all studies. This is because Study 5 compares ultrasonography to the gold standard, so that the estimates of D-dimer accuracy are not directly affected when it is excluded.

Sensitivity analysis to the MAR assumption

The NMA-DT model is built upon the assumption of MAR. However, in practice, nonrandom missingness (MNAR) can happen when, 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. In this subsection, we conduct a sensitivity analysis to explore the influence on parameter estimates when the MAR assumption is violated. Let the $N \times K$ matrix $\boldsymbol{M}$ denote the study-level missingness of a NMA-DT dataset containing N studies and K candidate tests. The entries of $\boldsymbol{M}$ are m_{ik}, for $i = 1, \ldots, N$ and $k = 1, \ldots, K$, such that $m_{ik} = 1$ if T_k is missing in study i and $m_{ik} = 0$ otherwise. Bernoulli distributions are assumed for the missingness indicators: $m_{ik} \sim Ber(p_{ik})$, where p_{ik} is the probability of a 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 its accuracy is low. When $\gamma_{1k} = 0$ ($\gamma_{0k} = 0$), the outcomes of T_k are MAR with respect to its sensitivity (specificity). We incorporate the model of missingness in the likelihood in (5.1) under different values of γ_{1k} and γ_{0k}: 0, -1 and -2, which correspond to MAR, and an odds ratio of missingness of 0.37 and 0.13, respectively (with respect to 1 unit increase in the logit scale of accuracy parameters).

Table 5.2: Meta-analysis of DVT tests: median parameter estimates and 95% CIs under different missingness assumptions.

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	-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 (00.74,0.98)	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)**

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.

The posterior medians of prevalence, sensitivity and specificity are presented in Table 5.2 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 our different assumptions. When the missingness probabilities are negatively correlated with one of the tests only, the posterior estimates of Se and Sp are lower while the accuracy estimates of the other tests are higher due to the correlation structure between the test accuracy parameters. For example, when $\gamma_{11} = \gamma_{01} = -1$, the posterior estimate of D-dimer Se is 0.78 compared to 0.83 under MAR, and Sp is 0.80 compared to 0.88 under MAR, while the ultrasonography Se is 0.95 compared to 0.90 under MAR and Sp is 0.83 compared to 0.80 under MAR. When the missingness probabilities are negatively correlated with sensitivities, the estimate of D-dimer Se is slightly higher, whereas the estimate of ultrasonography Se is slightly lower. When the missingness probabilities are negatively correlated with specificities, specificity estimates of D-dimer and ultrasonography are both lower than the estimates under MAR. When the missing probabilities are negatively correlated with sensitivities and specificities of both tests, assuming MAR generally provides higher estimates of the test accuracies (except for sensitivity estimates when $\gamma_{1k} = \gamma_{0k} = -1, k = 1, 2$). The differences between the estimates under the MAR and MNAR assumptions are generally enlarged when $\gamma_{1k} = \gamma_{0k} = -2, k = 1, 2$, than when they take values of -1. In general, when missingness is negatively correlated with the test accuracy parameters, ignoring the model of missingness will overestimate 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 under-estimated when one of the tests must cope with MNAR.

5.3.2 NMA of latent TB tests

The meta-analysis of latent TB studies in Sadatasfavi et al.[19] is re-analyzed by the proposed NMA-DT model. We use a Wishart prior with $v = 7$ (K=3) and $\boldsymbol{R}$ having diagonal elements equal to 5 and off diagonal elements equal to 0.05, which corresponds to a 95% prior CI of (0.2, 17) for the standard deviations. The same priors as in Section 5.3.1 are used for η, α_k and β_k. 1,000,000 MCMC samples were collected after

Table 5.3: Meta-analysis of latent TB tests: posterior medians and 95% CI's

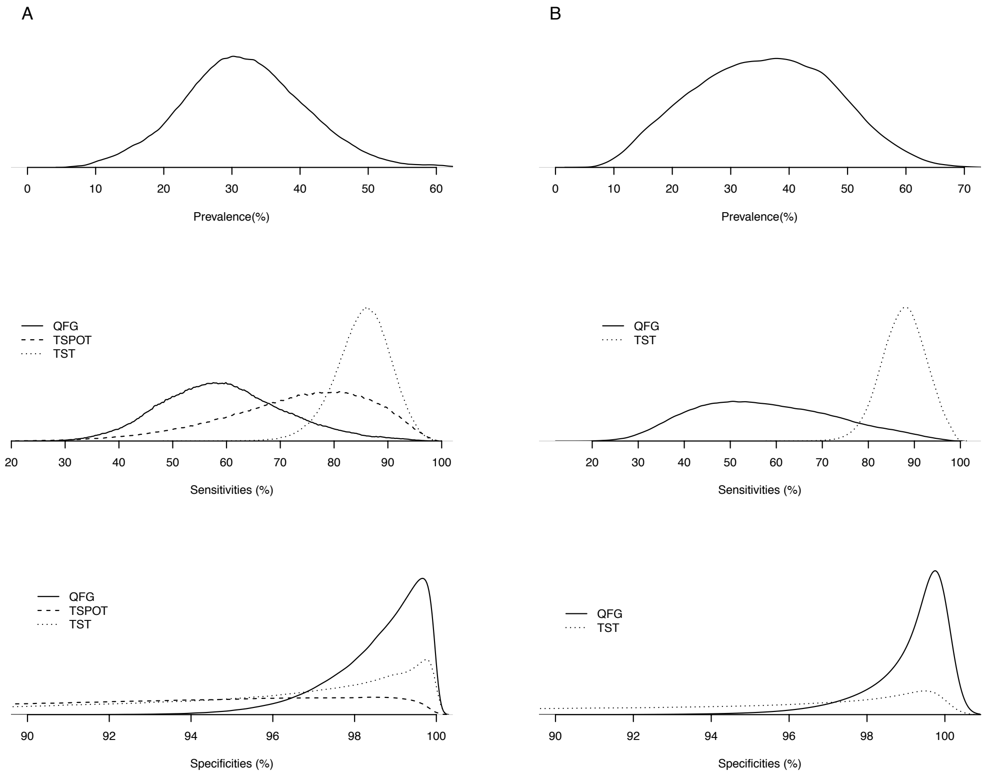
Parameter	Median	(95% CI)	Median	(95% CI)	Median	(95% CI)
Prevalence			0.31	(0.16, 0.49)		
	QFG		**TSPOT**		**TST**	
Sensitivity	0.58	(0.38, 0.82)	0.74	(0.41, 0.94)	0.85	(0.73, 0.94)
Specificity	0.99	(0.95, 1)	0.91	(0.62, 0.99)	0.96	(0.78, 1)
PPV	0.96	(0.81, 1)	0.82	(0.39, 0.99)	0.93	(0.50, 1)
NPV	0.84	(0.67, 0.96)	0.89	(0.69, 0.98)	0.93	(0.86, 0.98)
LR+	49.5	(12.4, 830.2)	10.2	(1.7, 164.8)	28.2	(3.7, 944.9)
LR-	0.42	(0.18, 0.63)	0.29	(0.07, 0.71)	0.16	(0.06, 0.29)

10,000 burn-in samples. The posterior estimates are summarized in Table 5.3, with marginal posterior density estimates are plotted in Figure 5.2, panel A. The overall disease prevalence posterior median is 0.31 (0.16, 0.49). Comparing the three tests, TST has highest sensitivity of 0.85 (0.73, 0.94) and specificity of 0.97 (0.77, 1). TSPOT has sensitivity of 0.74 (0.41, 0.94) and the lowest specificity, 0.93 (0.63, 1). QFG has the lowest sensitivity, 0.58 (0.38, 0.82), but the highest specificity, 0.99 (0.95, 1). The posterior probabilities of the TST, TSPOT and QFG test ranking first are 0.78, 0.2 and 0.02 in terms of sensitivity, and 0.29, 0.06 and 0.65 in terms of specificity, respectively. Therefore, the results suggest that the TST outperforms the other two tests considering sensitivity, and QFG performs the best in terms of specificity. Inconsistency estimates are -0.35 (-0.53, 0.15), 0.16 (-0.17, 0.58), 0.01 (-0.04, 0.19) and 0.01 (-0.10, 0.48) for QFG Se, TSPOT Se, QFG Sp and TSPOT Sp, respectively. Significant inconsistency is not detected in these estimates.

Sensitivity analyses to prior distributions of $\boldsymbol{\Sigma}^{-1}$

Sensitivity analyses to the prior distribution of $\boldsymbol{\Sigma}^{-1}$ are also considered to evaluate the effect of the prior distribution on the posterior estimates of prevalence, sensitivity and specificity. Compared with Table 5.3, when the Wishart prior takes $\boldsymbol{R}$ with diagonal elements equal to 20 and off-diagonal elements equal to 0.05, the TSPOT test and TST have lower Sp estimates: 0.87 (0.56, 0.99) and 0.93 (0.71, 1), respectively. Estimates of prevalence and other sensitivities and specificites are close to those in Table 5.3. When

Figure 5.2: Meta-analysis of latent TB tests: panel A plots the posterior densities of parameters using all studies; Panel B plots the posterior densities when studies reporting the TSPOT test are excluded.



the Wishart prior takes $\boldsymbol{R}$ with diagonal elements equal to 1 and off-diagonal elements equal to 0.05, similar posterior medians and CIs as in Table 5.3 are obtained.

Sensitivity analysis to TSPOT test

In Figure 5.2 panel A, the posterior densities of TSPOT sensitivity and specificity are relatively flat compared to the other two tests. This can be explained by the fact that only three out of 22 studies report TSPOT meaning there is not much information about TSPOT in these NMA-DT data. We therefore exclude the three studies with TSPOT (two comparing all tests and one comparing TSPOT to TST) to examine whether the exclusion of an individual candidate test affects the results of an NMA-DT. The posterior densities of the QFG and TST tests after excluding studies reporting TSPOT are presented in panel B of Figure 5.2. Posterior median disease prevalence is estimated as 0.36 (0.14, 0.58). Compared to the estimates from Table 5.3, the TST test is estimated to have a higher sensitivity of 0.88 (0.78, 0.96) but a lower specificity of 0.91 (0.63, 0.99). The QFG test has a slightly lower estimate of sensitivity of 0.56 (0.32, 0.88), and the same estimate of specificity of 0.99 (0.96, 1). The probabilities of the TST and QFG test ranking first are 0.96 and 0.04 in terms of sensitivity, and 0.09 and 0.91 in terms of specificity, respectively. The inference about TST is affected more by the exclusion of the TSPOT test than the QFG test. This is an interesting observation, indicating that excluding one of the tests, i.e., ignoring the correlation structure among all tests, could have nontrivial effects on the accuracy of the estimates of the other tests. Similar conclusions are made in Mills et al., in which excluding treatments in NMA can have substantial changes of other treatment effect estimates [98]. Therefore, it is important to account for such correlations, and to combine information from all available studies in NMA-DT.

Sensitivity analysis to the MAR assumption

The model of missingness in Section 5.3.1 is fitted to the latent TB data to explore the influence on the posterior estimates under MNAR. Again, non-positive γ_{1k} and γ_{0k} are used so that T_k is prone to be missing when its accuracy is low and γ_{1k} and γ_{0k} can take values 0 or -1. The posterior medians of prevalence, sensitivity and specificity are presented in Table 5.4 under different missingness assumptions: MAR, missingness

Table 5.4: Meta-analysis of latent TB tests: posterior estimates under different missingness assumptions.

MNAR	**None**	**Se**	**Sp**	**TSPOT**
Parameter	*Median (95% CI)*	*Median (95% CI)*	*Median (95% CI)*	*Median (95% CI)*
Prevalence	0.31 (0.16,0.49)	0.36 (0.06,0.6)	0.37 (0.18,0.6)	0.31 (0.15,0.53)
Se: QFG	0.58 (0.38,0.82)	**0.47 (0.29,0.99)**	0.53 (0.28,1)	0.56 (0.38,0.96)
Se: TSPOT	0.74 (0.41,0.94)	**0 (0,0.36)**	0.99 (0.06,1)	**0 (0,0.72)**
Se: TST	0.85 (0.73,0.94)	**0.86 (0.74,1)**	0.84 (0.71,0.94)	0.89 (0.74,1)
Sp: QFG	0.99 (0.91,1)	1 (0.95,1)	**1 (0.98,1)**	0.98 (0.94,1)
Sp: TSPOT	0.91 (0.62,0.99)	0.87 (0.01,1)	**0.04 (0,0.34)**	**0.86 (0,1)**
Sp: TST	0.96 (0.78,1)	1 (0.7,1)	**1 (0.81,1)**	1 (0.81,1)

MNAR="None" is equivalent to MAR; MNAR="Se"("Sp") means missingness related to the sensitivities (specificities) of all tests; MNAR="TSPOT" means missingness related to sensitivity and specificity of TSPOT test. Bold numbers indicate parameters directly related to missingness.

related to sensitivities of all three tests, missingness related to specificities of all tests, and missingness related to the TSPOT test only.

When MNAR is assumed for sensitivities, the QFG and TSPOT tests have lower posterior sensitivity estimates, where the sensitivity of the TSPOT test is estimated to be 0. Intuitively, because TSPOT is missing in 19 out of the 22 studies, it has a high missingness probability, and thus should have low sensitivity if assuming that a test is likely to be missing when it has bad performance (negative values of γ_{1k} in this case). For the same reason, when MNAR is assumed for specificities, the TSPOT test has specificity estimated as 0.04. When MNAR is assumed for the TSPOT test, its sensitivity is estimated to be 0 and its specificity as 0.86, both lower than the estimates under MAR.

5.4 Simulation Studies

5.4.1 Simulation setups

Simulation studies were conducted to test how the NMA-DT model performs under different assumptions. 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: $(\sigma_{\pi}, \sigma_{Se_1}, \sigma_{Sp_1}, \ldots, \sigma_{Se_K}, \sigma_{Sp_K}) = 0.3$. We test the model performance under different assumptions on the magnitude of the correlations on the probit-transformed parameters: weak (0.3), medium(0.5) or strong (0.8) positive correlations between prevalence and sensitivities and negative correlations between prevalence and specificities and between sensitivies and specificities. 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 5.2, with the mean vector and covariance matrix specified above. For each subject, test outcomes are generated according to the likelihood equation (5.1). Finally, missingness indicators for each of the three tests are assigned 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. Cross-classified cell counts can be collected for each study to present the observed data as in the case studies. Each simulated dataset is fitted by the proposed NMA-DT method.

We compare the performance of the NMA-DT model with a “naive” approach. The “naive” method applies the trivariate generalized linear mixed model (TGLMM) [16] 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. 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. The “naive” analysis is not applied to T_2 because T_1 and T_2 are exchangeable.

5.4.2 Simulation results

Table 5.5 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 correlation assumptions, the NMA-DT model is shown to provide nearly unbiased estimates for all parameters with small MSE. Generally, as the correlation becomes stronger, the estimates are more biased. For example, the estimate of β_1 is biased by 0.02 when there are strong correlations. The coverage probabilities remain close to the nominal level of 0.95 when there is weak or medium correlation, and increase to around 0.97 when there is strong correlation.

The fixed effect estimates for prevalence, sensitivity and specificity for T_1 from the "naive" approach are also summarized in Table 5.5, column "Naive". Comparing the two models, this approach provides generally larger bias and consistantly larger MSE than the NMA-DT approach. This suggests that the NMA-DT model is less biased and more efficient than the "naive" approach. It provides evidence that the NMA-DT model gains efficiency by combing information from more studies and by taking into account the correlation structure. Moreover, when there are weak or strong correlations, the "naive" approach has large CPs and are greater than 0.97. Overall, the NMA-DT model is shown to outperform the "naive" approach.

5.5 Discussion

There is a growing interest in simultaneously comparing the performance of multiple diagnostic tests in a meta-analysis. However, due to the mixture of different study designs, the variety of reported test outcomes across studies, the inherent heterogeneity in a meta-analysis, and the complex correlation structure of multiple test outcomes on the same individuals, the methodological development for NMA-DT remains challenging. In this chapter, we presented a Bayesian hierarchical NMA-DT framework that addresses these challenges. An important feature of this proposed framework is that it unifies all three types of study designs into the multiple test comparison design using a missing data framework. In addition, the proposed framework can provide ranks of diagnostic tests, which can be used to guide clinical decision making. Through simulation

Table 5.5: 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.

	NMA-DT			Naive		
Parameter (true)	Bias	MSE	CP	Bias	MSE	CP
Weak Correlation						
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			
Medium Correlation						
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						
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			

studies, we have shown that the proposed NMA-DT method can provide unbiased estimates for overall prevalence, test sensitivities, and specificities. In addition, it is more efficient than a commonly used "naive" approach, in which separate meta-analyses are implemented to evaluate one candidate test at a time.

The importance of checking the evidence consistency assumption was discussed. However, current inconsistency measures in NMA-CT, such as the inconsistency degrees of freedom [87] and the use of consistency equations [99, 100], cannot be directly applied in NMA-DT because they are built upon relative effects. In this paper, we proposed to measure inconsistency by computing the mean difference between posterior estimates from studies reporting and not reporting a specific test. This measurement may suffer from insufficient power when study sample sizes are small. Further research shall focus on developing a formal test out of this measurement, and validate its power.

Another critical assumption is independent test results within each disease class. In traditional MA-DT, this assumption is required under non-gold standard situations where latent class models are used [18]. However, conditional dependence can exist, such as when two candidate tests are based on a similar biological phenomeno. Several methods have been developed to adjust for this dependence either through a correlation parameter [18], an additional latent class random effect [101], or using multivariate probit models [102]. However, they cannot be directly applied to our NMA-DT model, because correlation parameters are suitable only for pairwise comparisons, and only a small portion of the studies utilizing three different kinds of designs may be subject to conditional dependence. Specifically, for studies adopting the randomized design, each candidate test is compared to the gold standard, thus the conditional independence assumption is not required. For studies adopting the multiple test comparison design, conditional dependence may only become a concern when a gold-standard test is missing. For similar reasons, non-comparative designs may be subject to conditional dependence when a gold-standard test is not involved, and subjects are tested by multiple candidate tests. As a result, how to adjust for conditional dependence in NMA-DT is a subject for future studies.

A concern brought by combining studies in a systematic review is how to correctly measure between-study heterogeneity, which plays an important role in choosing the appropriate statistical model (i.e., fixed versus random effects model). In this chapter, as well as in some of the “classic” random effects meta-analysis methods, generalized linear mixed models were used to account for heterogeneity in a Bayesian framework. The posterior estimates of the covariance matrix for the random effects can provide some information on the extent of heterogeneity. An inverse Wishart prior is often used for the covariance matrix, but is limited in that the posterior variance components are then always positive [103]. Another limitation brought by the inverse Wishart prior is that when the correlation matrix grows, it imposes an unstructured covariance matrix while a structured correlation assumption may be preferred to improve estimation and convergence. Several attampts have been made to find alternative priors for the covariance matrix [104, 71, 103], and their application in NMA-DT merits further research.

Chapter 6

Conclusions

6.1 Summary of major findings

This thesis explored novel statistical methods in the context of meta-analysis and network meta-analysis of diagnostic tests. Our contribution lies in improving accuracy and efficiency in evaluating diagnostic tests, which ultimately enable better decision-making for patients, health practitioners and policy makers. In practice, this thesis serves as a practical guide on how to better evaluate diagnostic test performance by presenting detailed examples. Most importantly, the application of our novel NMA-DT methods can be broadly applied in summarizing clinical trials related to diagnosing various diseases. Health care decision makers can be better informed from this method than from the traditional MA-DT, as they will now be able to rank each test and estimate the probabilities of being the best diagnostic test.

In Chapter 2, we provided a systematic review of both traditional and advanced models in MA-DT, for settings with and without a gold standard. We made careful comparisons and summarizations of different methods, and gave recommendations for their application in practice.

After reviewing current development in MA-DT, Chapter 3 tackled some of the remaining limitations. We proposed a novel Bayesian hybrid GLMM that addressed two important problems at the same time: combining case-control and cohort studies in one meta-analysis; and adjusting for partial verification bias. We evaluated and compared its performance with the current method, and illustrated its application using two interesting case studies. This method fills in the gap of correctly dealing with partial verification in meta-analysis settings and allowing for utilization of much information as possible from combing different study designs.

In Chapter 4, we then investiagated the challenge of accounting for non-evaluable subjects. Built upon the TGLMM [16] approach, we proposed an extended TGLMM method to account for this. Furthermore, we discovered that some current approaches and conclusions in the literature can be misleading, and conducted simulation studies to compare the performance of the proposed method. Simulation findings support the extended TGLMM in that they give unbiased estimates of sensitivity and specificity, as well as prevalence and other prevalence related accuracy indices (NPV and PPV). Also, recommendations were made to aid readers in choosing appropriate models in practice.

Finally, in Chapter 5 we proceeded from MA-DT to NMA-DT, where the topic was extended from evaluating a single candidate test to simultaneously evaluating multiple candidate tests. We developed a missing data framework and a Bayesian hierarchical model that offers important advantages over the traditional MA-DT: 1) it combines studies using all three designs; 2) it pools both studies with and without a gold standard; 3) it combines studies with different sets of candidate tests; and 4) it accounts for heterogeneity across studies and complex correlation structure among multiple tests. We provided two examples to illustrate the proposed model, with discussion of the MAR assumption, effect of outlying studies, choice of prior for covariance matrix etc. Our work broadens the scope of meta-analysis of diagnostic tests from single test to comparison of multiple tests, addressing the keen need of cost-effectiveness research generated from the exploding number of available diagnostic instruments for a disease condition.

6.2 Limitations and extensions to future work

Limitations remain that suggest future work. The hybrid GLMM in Chapter 3 and extended TGLMM in Chapter 4 both rely on the assumption of MAR, as well as the main model of NMA-DT in Chapter 5. We considered MNAR situations in the discussion sections of these chapters, and presented models of missingness to adjust for MNAR as sensitivity studies in the two examples in Chapter 5. We can take a closer look at MNAR situations in future research. Another limitation is the choice of priors for covariance matrices. In Bayesian multivariate GLMMs (Chapter 3 and 5), Wishart priors were used for the covariance matrices to ensure positive definite matrices. However, this approach has some limitations. On the one hand, the choice of parameters for the Wishart prior needs more discussion. We attempted to study the effect of prior selection on posterior estimates through scaled Wishart priors and sensitivity analysis. Further studies can focus on conducting simulation studies to systematically compare these effects. On the other hand, more importantly, this issue is closely related to estimating the degree of heterogeneity in MA-DT and NMA-DT. All the random effects models introduced in this thesis aim at explaining heterogeneous performance of tests across studies, thus correctly estimating the extent of this heterogeneity can justify our choice of random effects models over fixed effects model. Wishart prior is limited in that the posterior variance components are always positive, which restricted the development of a test for zero variances. Application of alternative priors deserves future study. For NMA-DT, evidence inconsistency can become an important concern when applying the proposed model. We proposed a measurement of inconsistency in Chapter 5. Based on this measurement, we can proceed to develop a formal test and validate its power.

This thesis motivates some interesting practical topics as well. First of all, to help researchers applying the proposed approaches, developing computation software, such as an R package, can ease the application of the models. Second, to step outside the diagnostic test framework, potential applications can be meta-analysis of safety studies in clinical research, where limited safety data in single studies can be combined across studies. The correlation between toxicity and efficacy estimates shares some similarity in the correlation between accuracy indices of diagnostic tests. Last but not least, the hybrid GLMM and the Bayesian NMA-DT model both incorporate the idea of combing different study designs in one meta-analysis, which can be potentially extended to combing information observational studies and from randomized clinical trials.

References

[1] Margaret Sullivan Pepe. *The statistical evaluation of medical tests for classification and prediction*, chapter 2. Oxford University Press, Oxford, 2003.

[2] Carolyn M Rutter and Constantine A Gatsonis. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine*, 20(19):2865–2884, 2001.

[3] Fujian Song, Khalid S Khan, Jacqueline Dinnes, and Alex J Sutton. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *International Journal of Epidemiology*, 31(1):88–95, 2002.

[4] Hans C Van Houwelingen, Lidia R Arends, and Theo Stijnen. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, 21(4):589–624, 2002.

[5] Roger M Harbord, Jonathan J Deeks, Matthias Egger, Penny Whiting, and Jonathan AC Sterne. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics*, 8(2):239–251, 2007.

[6] Petra Macaskill. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *Journal of Clinical Epidemiology*, 57(9):925–932, 2004.

[7] Susan Mallett, Jonathan J Deeks, Steve Halligan, Sally Hopewell, Victoria Cornelius, and Douglas G Altman. Systematic reviews of diagnostic tests in cancer: review of methods and reporting. *British Medical Journal*, 333(7565):413, 2006.

[8] Aeilko H Zwinderman and Patrick M Bossuyt. We should not pool diagnostic likelihood ratios in systematic reviews. *Statistics in Medicine*, 27(5):687–697, 2008.

[9] LR Arends, TH Hamza, JCv Houwelingen, MH Heijenbrok-Kal, MGM Hunink, and Theo Stijnen. Bivariate random effects meta-analysis of roc curves. *Medical Decision Making*, 2008.

[10] Haitao Chu and Stephen R Cole. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of clinical epidemiology*, 59(12):1331–1332, 2006.

[11] Johannes B Reitsma, Afina S Glas, Anne WS Rutjes, Rob JPM Scholten, Patrick M Bossuyt, and Aeilko H Zwinderman. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*, 58(10):982–990, 2005.

[12] Haitao Chu and Hongfei Guo. Letter to the editor: a unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics*, 10(1):201–203, 2009.

[13] Richard D Riley, Keith R Abrams, Alexander J Sutton, Paul C Lambert, and John R Thompson. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology*, 7(1):3, 2007.

[14] RD Riley, KR Abrams, PC Lambert, AJ Sutton, and JR Thompson. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in medicine*, 26(1):78–97, 2007.

[15] Taye H Hamza, Hans C van Houwelingen, and Theo Stijnen. The binomial distribution of meta-analysis was preferred to model within-study variability. *Journal of clinical epidemiology*, 61(1):41–51, 2008.

[16] Haitao Chu, Lei Nie, Stephen R Cole, and C Poole. Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: Alternative parameterizations and model selection. *Statistics in medicine*, 28(18):2384–2399, 2009.

[17] SD Walter, Les Irwig, and PP Glasziou. Meta-analysis of diagnostic tests with imperfect reference standards. *Journal of Clinical Epidemiology*, 52(10):943–951, 1999.

[18] Haitao Chu, Sining Chen, and Thomas A Louis. 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(486):512–523, 2009.

[19] Mohsen Sadatsafavi, Neal Shahidi, Fawziah Marra, Mark J FitzGerald, Kevin R Elwood, Na Guo, and Carlo A Marra. 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*, 63(3):257–269, 2010.

[20] Nandini Dendukuri, Ian Schiller, Lawrence Joseph, and Madhukar Pai. Bayesian meta-analysis of the accuracy of a test for tuberculous pleuritis in the absence of a gold standard reference. *Biometrics*, 68(4):1285–1293, 2012.

[21] Mariska MG Leeflang, Patrick MM Bossuyt, and Les Irwig. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *Journal of Clinical Epidemiology*, 62(1):5–12, 2009.

[22] Haitao Chu, Lei Nie, Stephen R Cole, and Charles Poole. Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: Alternative parameterizations and model selection. *Statistics in medicine*, 28(18):2384–2399, 2009.

[23] Wenche M Klerkx, Leon Bax, Wouter B Veldhuis, A Peter M Heintz, Willem PThM Mali, Petra HM Peeters, and Karel GM Moons. Detection of lymph node metastases by gadolinium-enhanced magnetic resonance imaging: systematic review and meta-analysis. *Journal of the National Cancer Institute*, 102(4):244–253, 2010.

[24] Susan Liebeschuetz, Sheila Bamber, Katie Ewer, Jonathan Deeks, Ansar A Pathan, and Ajit Lalvani. Diagnosis of tuberculosis in south african children with a t cell-based assay: a prospective cohort study. *The Lancet*, 364(9452):2196–2203, 2004.

[25] Anne WS Rutjes, Johannes B Reitsma, Jan P Vandenbroucke, Afina S Glas, and Patrick MM Bossuyt. Case–control and two-gate designs in diagnostic accuracy studies. *Clinical Chemistry*, 51(8):1335–1341, 2005.

[26] David F Ransohoff and Alvan R Feinstein. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *The New England Journal of Medicine*, 299(17):926–930, 1978.

[27] Joris AH de Groot, Patrick MM Bossuyt, Johannes B Reitsma, Anne WS Rutjes, Nandini Dendukuri, Kristel JM Janssen, and Karel GM Moons. Verification problems in diagnostic accuracy studies: consequences and solutions. *BMJ*, 343, 2011.

[28] Penny Whiting, Anne WS Rutjes, Johannes B Reitsma, Patrick MM Bossuyt, and Jos Kleijnen. The development of quadas: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology*, 3(1):25, 2003.

[29] Anne WS Rutjes, Johannes B Reitsma, Marcello Di Nisio, Nynke Smidt, Jeroen C van Rijn, and Patrick MM Bossuyt. Evidence of bias and variation in diagnostic accuracy studies. *Canadian Medical Association Journal*, 174(4):469–476, 2006.

[30] Jeroen G Lijmer, Ben Willem Mol, Siem Heisterkamp, Gouke J Bonsel, Martin H Prins, Jan HP van der Meulen, and Patrick MM Bossuyt. Empirical evidence of design-related bias in studies of diagnostic tests. *Journal of the American Medical Association*, 282(11):1061–1066, 1999.

[31] Colin B Begg and Robert A Greenes. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*, pages 207–215, 1983.

[32] Xiao-Hua Zhou. Correcting for verification bias in studies of a diagnostic test's accuracy. *Statistical Methods in Medical Research*, 7(4):337–353, 1998.

[33] JA Roldán Nofuentes and JD Luna del Castillo. Comparing the likelihood ratios of two binary diagnostic tests in the presence of partial verification. *Biometrical Journal*, 47(4):442–457, 2005.

[34] Ofer Harel and Xiao-Hua Zhou. Multiple imputation for correcting verification bias. *Statistics in medicine*, 25(22):3769–3786, 2006.

[35] JAH De Groot, KJM Janssen, AH Zwinderman, KGM Moons, and JB Reitsma. Multiple imputation to correct for partial verification bias revisited. *Statistics in medicine*, 27(28):5880–5889, 2008.

[36] A Rogier T Donders, Geert JMG van der Heijden, Theo Stijnen, and Karel GM Moons. Review: a gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091, 2006.

[37] Andrzej S Kosinski and Huiman X Barnhart. Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics*, 59(1):163–171, 2003.

[38] Stuart G Baker. Evaluating multiple diagnostic tests with partial verification. *Biometrics*, pages 330–337, 1995.

[39] Joris AH de Groot, Nandini Dendukuri, Kristel JM Janssen, Johannes B Reitsma, James Brophy, Lawrence Joseph, Patrick MM Bossuyt, and Karel GM Moons. Adjusting for partial verification or workup bias in meta-analyses of diagnostic accuracy studies. *American journal of epidemiology*, 175(8):847–853, 2012.

[40] Colin B Begg, Robert A Greenes, and Boris Iglewicz. The influence of uninterpretability on the assessment of diagnostic tests. *Journal of chronic diseases*, 39(8):575–584, 1986.

[41] Georg M Schuetz and Peter Schlattmann. Use of 3 × 2 tables with an intention to diagnose approach to assess clinical performance of diagnostic tests: meta-analytical evaluation of coronary ct angiography studies. *BMJ*, 345(2):e6717–e6717, 2012.

[42] David L Simel, John R Feussner, Elizabeth R Delong, and David B Matchar. Intermediate, indeterminate, and uninterpretable diagnostic test results. *Medical Decision Making*, 7(2):107–114, 1987.

[43] Xiaoye Ma, Lei Nie, Stephen R Cole, and Haitao Chu. Statistical methods for multivariate meta-analysis of diagnostic tests: An overview and tutorial. *Statistical Methods in Medical Research*, 2013. In press.

[44] Taye H Hamza, Johannes B Reitsma, and Theo Stijnen. Meta-analysis of diagnostic studies: a comparison of random intercept, normal-normal, and binomial-normal bivariate summary roc approaches. *Medical Decision Making*, 28(5):639–649, 2008.

[45] Haitao Chu, Hongfei Guo, and Yijie Zhou. Bivariate random effects meta-analysis of diagnostic studies using generalized linear mixed models. *Medical Decision Making*, 30(4):499–508, 2010.

[46] Roderick JA Little and DB Rubin. *Statistical analysis with missing data, 2nd edn.* John Wiley & Sons, New Jersey, 2002.

[47] Abdel Aziz M Shaheen, Alex F Wan, and Robert P Myers. Fibrotest and fibroscan for the prediction of hepatitis c-related fibrosis: a systematic review of diagnostic test accuracy. *The American journal of gastroenterology*, 102(11):2589–2600, 2007.

[48] Madhukar Pai, Alice Zwerling, and Dick Menzies. Systematic review: T-cell–based assays for the diagnosis of latent tuberculosis infection: an update. *Annals of internal medicine*, 149(3):177–184, 2008.

[49] R Diel, D Goletti, G Ferrara, G Bothamley, D Cirillo, B Kampmann, C Lange, M Losi, R Markova, and GB Migliori. Interferon-γ release assays for the diagnosis of latent mycobacterium tuberculosis infection: a systematic review and meta-analysis. *European Respiratory Journal*, 37(1):88–99, 2011.

[50] Karen R Steingart, Laura L Flores, Nandini Dendukuri, Ian Schiller, Suman Laal, Andrew Ramsay, Philip C Hopewell, and Madhukar Pai. Commercial serological tests for the diagnosis of active pulmonary and extrapulmonary tuberculosis: an updated systematic review and meta-analysis. *PLoS medicine*, 8(8):1–19, 2011.

[51] Jian Kang, Rollin Brant, and William A Ghali. Statistical methods for the meta-analysis of diagnostic tests must take into account the use of surrogate standards. *Journal of Clinical Epidemiology*, 66(5):566–574, 2013.

[52] Yemisi Takwoingi, Mariska MG Leeflang, and Jonathan J Deeks. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Annals of Internal Medicine*, 158(7):544–554, 2013.

[53] TA Trikalinos, DC Hoaglin, KM Small, and CH Schimid. Evaluating practices and developing tools for comparative effectiveness reviews of diagnostic test accuracy: Methods for the joint meta-analysis of multiple tests. methods research report. Technical report, Rockville, MD: Agency for Healthcare Research and Quality, 01 2013.

[54] Lincoln E Moses, David Shapiro, and Benjamin Littenberg. Combining independent studies of a diagnostic test into a summary roc curve: Data-analytic approaches and some additional considerations. *Statistics in Medicine*, 12(14):1293–1316, 1993.

[55] SD Walter. The partial area under the summary roc curve. *Statistics in medicine*, 24(13):2025–2040, 2005.

[56] Alexander J Sutton, Nicola J Cooper, Steve Goodacre, and Matthew Stevenson. Integration of meta-analysis and economic decision modeling for evaluating diagnostic tests. *Medical Decision Making*, 2008.

[57] Les Irwig, Anna NA Tosteson, Constantine Gatsonis, Joseph Lau, Graham Colditz, Thomas C Chalmers, and Frederick Mosteller. Guidelines for meta-analyses evaluating diagnostic tests. *Annals of Internal medicine*, 120(8):667–676, 1994.

[58] Jonathan J Deeks. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *BMJ: British Medical Journal*, 323(7305):157, 2001.

[59] Anna N Angelos Tosteson and Colin B Begg. A general regression methodology for roc curve estimation. *Medical Decision Making*, 8(3):204–215, 1988.

[60] HERMANN Brenner, OLAF Gefeller, et al. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Statistics in medicine*, 16(9):981–991, 1997.

[61] Jialiang Li, Jason P Fine, and Nasia Safdar. Prevalence-dependent diagnostic accuracy measures. *Statistics in Medicine*, 26(17):3258–3273, 2007.

[62] Bradley P Carlin and Thomas A Louis. *Bayesian methods for data analysis.* CRC Press, 2011.

[63] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[64] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087, 1953.

[65] Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

[66] Walter R Gilks, NG Best, and KKC Tan. Adaptive rejection metropolis sampling within gibbs sampling. *Applied Statistics*, pages 455–472, 1995.

[67] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

[68] Stephen P Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998.

[69] Andrew Thomas, Bob OHara, Uwe Ligges, and Sibylle Sturtz. Making bugs open. *R news*, 6(1):12–17, 2006.

[70] David J Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337, 2000.

[71] John Barnard, Robert McCulloch, and Xiao-Li Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4):1281–1312, 2000.

[72] T Tokuda, B Goodrich, I Van Mechelen, A Gelman, and F Tuerlinckx. Visualizing distributions of covariance matrices. Technical report, Technical report, University of Leuwen, Belgium and Columbia University, USA. URL http://www. stat. columbia. edu/˜ gelman/research/unpublished/Visualization. pdf.(Cited on pages 114, 116, 117 and 119.), 2011.

[73] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2003.

[74] WJ Browne and David Draper. Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational statistics*, 15:391–420, 2000.

[75] Paul C Lambert, Alex J Sutton, Paul R Burton, Keith R Abrams, and David R Jones. How vague is vague? a simulation study of the impact of the use of vague prior distributions in mcmc using winbugs. *Statistics in Medicine*, 24(15):2401–2428, 2005.

[76] Giles WL Boland, Ben A Dwamena, Minal Jagtiani Sangwaiya, Alexander G Goehler, Michael A Blake, Peter F Hahn, James A Scott, and Mannudeep K Kalra. Characterization of adrenal masses by using fdg pet: a systematic review and meta-analysis of diagnostic test performance. *Radiology*, 259(1):117–126, 2011.

[77] Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

[78] Jonathan Davey, Rebecca M Turner, Mike J Clarke, and Julian PT Higgins. Characteristics of meta-analyses and their component studies in the cochrane database of systematic reviews: a cross-sectional, descriptive analysis. *BMC medical research methodology*, 11(1):160, 2011.

[79] Roderick JA Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

[80] Xiaoye Ma, Muhammad Fareed K Suri, and Haitao Chu. A trivariate meta-analysis of diagnostic studies accounting for prevalence and non-evaluable subjects: re-evaluation of the meta-analysis of coronary ct angiography studies. *BMC medical research methodology*, 14(1):128, 2014.

[81] Roderick JA Little. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121, 1995.

[82] Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

[83] Bethany Shinkins, Matthew Thompson, Susan Mallett, and Rafael Perera. Diagnostic accuracy studies: how to report and analyse inconclusive test results. *BMJ: British Medical Journal*, 346:f2778, 2013.

[84] Christopher GT Blick, Sarfraz A Nazir, Susan Mallett, Benjamin W Turney, Natasha N Onwu, Ian SD Roberts, Jeremy P Crew, and Nigel C Cowan. Evaluation of diagnostic strategies for bladder cancer using computed tomography (ct) urography, flexible cystoscopy and voided urine cytology: results for 778 patients from a hospital haematuria clinic. *BJU international*, 110(1):84–94, 2012.

[85] Julian Higgins and Simon G Thompson. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11):1539–1558, 2002.

[86] G Lu and AE Ades. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine*, 23(20):3105–3124, 2004.

[87] Guobing Lu and AE Ades. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association*, 101(474):447–459, 2006.

[88] Georgia Salanti, AE Ades, and John Ioannidis. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *Journal of clinical epidemiology*, 64(2):163–171, 2011.

[89] Jing Zhang, Bradley P Carlin, James D Neaton, Guoxing G Soon, Lei Nie, Robert Kane, Beth A Virnig, and Haitao Chu. Network meta-analysis of randomized clinical trials: Reporting the proper summaries. *Clinical Trials*, 11(2):246–262, 2014.

[90] Hwanhee Hong, Haitao Chu, Jing Zhang, and Brad P Carlin. A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Research Synthesis Methods*, 2015. In press.

[91] Tania B Huedo-Medina, Julio Sánchez-Meca, Fulgencio Marin-Martinez, and Juan Botella. Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological methods*, 11(2):193–206, 2006.

[92] Enrique R Venta and Luz A Venta. The diagnosis of deep-vein thrombosis: an application of decision analysis. *Journal of the Operational Research Society*, 38(7):615–624, 1987.

[93] Clive Tovey and Suzanne Wyatt. Diagnosis, investigation, and management of deep vein thrombosis. *British Medical Journal*, 326(7400):1180–1184, 2003.

[94] Enrico Bernardi, Paolo Prandoni, Anthonie WA Lensing, Giancarlo Agnelli, Giuliana Guazzaloca, Gianluigi Scannapieco, Franco Piovella, Fabio Verlato, Cristina Tomasi, Marco Moia, et al. D-dimer testing as an adjunct to ultrasonography in patients with clinically suspected deep vein thrombosis: prospective cohort study. *BMJ: British Medical Journal*, 317(7165):1037–1040, 1998.

[95] J Dinnes, J Deeks, H Kunst, A Gibson, E Cummins, N Waugh, F Drobniewski, and A Lalvani. A systematic review of rapid diagnostic tests for the detection of tuberculosis infection. *Health Technology Assessment*, 11(3):1–196, 2007.

[96] Martyn Plummer et al. 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)*, pages 20–22, 2003.

[97] Hong Zhao, Jim S Hodges, Haijun Ma, Qi Jiang, and Bradley P Carlin. Hierarchical Bayesian approaches for detecting inconsistency in network meta-analysis.

Research Report, Division of Biostatistics, University of Minnesota, 006, 2014. Submitted to *Statistics in Medicine*.

[98] Edward J Mills, Steve Kanters, Kristian Thorlund, Anna Chaimani, Areti-Angeliki Veroniki, and John Ioannidis. The effects of excluding treatments from network meta-analyses: survey. *BMJ: British Medical Journal*, 347:f5195, 2013.

[99] Georgia Salanti, Julian PT Higgins, AE Ades, and John PA Ioannidis. Evaluation of networks of randomized trials. *Statistical methods in medical research*, 17(3):279–301, 2008.

[100] Orestis Efthimiou, Dimitris Mavridis, Richard D Riley, Andrea Cipriani, and Georgia Salanti. Joint synthesis of multiple correlated outcomes in networks of interventions. *Biostatistics*, 16(1):84–97, 2015.

[101] Yinsheng Qu, Ming Tan, and Michael H Kutner. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, pages 797–810, 1996.

[102] Huiping Xu and Bruce A Craig. A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests. *Biometrics*, 65(4):1145–1155, 2009.

[103] Zhen Chen and David B Dunson. A Bayesian approach for assessing heterogeneity in generalized linear models. Technical report, Working Paper, National Institute of Environmental Health Sciences, 2003.

[104] Michael J Daniels and Robert E Kass. Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association*, 94(448):1254–1263, 1999.

Appendix A

Data for MA of gadolimium-enhanced MRI and MA of FDG PET

Table A.1: Data for the meta-analysis of gadolinium-enhanced MRI in detecting lymph node metastases

ID	Study	n_{11}	n_{01}	n_{10}	n_{00}	n_{1m}	n_{0m}
	Cohort Studies						
1	Bley, 2005	7	3	3	6	0	0
2	Drew, 1999	7	5	5	12	0	0
3	Gaa, 1999	18	6	3	19	0	0
4	Hasegawa, 2003	11	1	4	14	0	0
5	Kang, 2000	9	5	3	29	0	0
6	Kaza, 2006[a]	3	2	1	9	1	7
7	Krupski, 2002	6	0	2	7	0	0
8	Kvistad, 2000	15	9	8	33	0	0
9	Low, 2003	15	7	1	25	0	0
10	Wallengren, 1996[ab]	2	1	0	7	2	0
11	Einspieler, 1991[a]	5	1	2	3	9	4
12	Hawighorst, 1998	13	6	3	11	0	0
13	Hallscheidt, 1998[a]	7	1	1	23	NA	NA
14	Luciani, 2004	7	1	1	7	0	0
15	Sheu, 2001[ab]	9	2	4	26	38	0
16	Manfredi, 2004[ab]	1	1	1	18	16	0
17	Murray, 2002	10	0	17	20	0	0
18	Okizuka, 1996	10	5	3	14	0	0
19	Oellinger, 2000	5	8	2	17	0	0
20	Rockall, 2007[a]	4	5	1	40	23	23
21	Barentsz, 1996	12	2	2	41	0	0
22	Ramsay, 2004[ab]	2	5	2	7	9	0
23	Hunerbein, 2000	3	1	1	22	0	0
24	Matsuoka, 2003	5	2	0	12	0	0
25	Thurnher, 1991[a]	6	3	1	11	NA	NA
26	Mumtaz, 1997	36	4	6	29	0	0
	Case-control Studies						
27	Heuck, 1997	16	2	2	22	NA	NA
28	Kim, 2000	91	16	65	45	NA	NA
29	Vorreuther, 1990	4	0	1	31	NA	NA
30	Tempany, 2000	5	8	25	133	NA	NA
31	Matsuoka, 2004	18	6	8	22	NA	NA
32	Medl, 1995	6	6	1	16	NA	NA

[a]: The study has partial verification.

[b]: The numbers of n_{1m} and n_{0m} are arbitrarily assigned such that $n_{0m} = 0$.

Table A.2: Data for the meta-analysis of FDG PET in characterizing adrenal masses

ID	Study	n_{11}	n_{01}	n_{10}	n_{00}	n_{1m}	n_{0m}
	Cohort Studies						
1	Groussin, 2009^{a}	22	5	0	38	0	12
2	Brady, 2009^{a}	36	36	36	44	0	92
3	Boland, 2009^{a}	14	0	0	10	0	32
4	Vikram, 2008	25	12	5	70	0	0
5	Tessonnier, 2008^{ab}	12	0	0	29	0	0
6	Sung, 2008^{a}	26	7	8	19	0	1
7	Okada, 2008^{ab}	16	0	3	16	0	0
8	Park, 2007^{ab}	7	3	1	9	0	0
9	Han, 2007^{a}	60	7	4	34	0	75
10	Caoili, 2007^{ab}	10	5	1	43	0	0
11	Jana, 2006^{ab}	28	2	2	48	0	12
12	Blake, 2006^{ab}	9	2	0	30	0	0
13	Metser, 2005^{a}	67	8	1	99	0	7
14	Zettinig, 2004^{ab}	3	0	0	13	0	0
15	Kumar, 2004^{ab}	67	4	5	37	0	0
16	Frilling, 2004	31	2	0	11	0	0
17	Yun, 2001^{ab}	18	2	0	30	0	0
18	Maurea, 2001	13	0	0	10	0	0
19	Gupta, 2001^{ab}	21	1	1	11	0	0
20	Erasmus, 1997^{ab}	23	2	0	8	0	0
21	Boland, 1995^{a}	14	0	0	10	0	32

^{a}: The study has partial verification.

^{b}: Failed to extract missing counts.

Appendix B

Additional simulation results for hybrid GLMM

Table B.1: Summary of 2000 simulations with data generated from settings with 30 studies and true Se (Sp)=0.9 (0.95).

		Sp			Se			π			PPV			NPV		
Corr[a]	Model[b]	Bias	RE	CP	Bias	RE	CP	Bias	RE	CP	Bias	RE	CP	Bias	RE	CP
0	1	0	1	0.94	0	1	0.93	0.01	1	0.93	0	1	0.94	0	1	0.94
0	2	-0.04	NA	0.52	0.03	NA	0.61	0.19	NA	0.34	0.04	NA	0.84	-0.02	NA	0.75
0	3	0	0.49	0.94	0	0.23	0.94	0.01	1.06	0.95	-0.01	0.76	0.94	0	0.44	0.94
0.5	1	0	1	0.95	0	1	0.94	0	1	0.93	0	1	0.95	0	1	0.95
0.5	2	-0.04	NA	0.5	0.04	NA	0.48	0.15	NA	0.44	0.03	NA	0.91	-0.01	NA	0.9
0.5	3	0	0.5	0.94	0.01	0.28	0.93	0	0.93	0.94	0	0.86	0.96	0	0.48	0.94
0.8	1	0	1	0.94	0	1	0.94	0	1	0.95	0	1	0.96	0	1	0.96
0.8	2	-0.04	NA	0.51	0.04	NA	0.43	0.13	NA	0.46	0.02	NA	0.93	-0.01	NA	0.95
0.8	3	0	0.51	0.93	0.02	0.29	0.92	0	0.81	0.94	0.01	0.85	0.96	0.01	0.48	0.93

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

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

Table B.2: Summary of 2000 simulations with data generated from settings with 10 studies and true Se (Sp)=0.7 (0.8).

		Sp			Se			π			PPV			NPV		
Corr[a]	Model[b]	Bias	RE	CP	Bias	RE	CP	Bias	RE	CP	Bias	RE	CP	Bias	RE	CP
0	1	-0.01	1	0.94	-0.01	1	0.93	0.01	1	0.93	0	1	0.92	-0.01	1	0.93
0	2	-0.13	NA	0.69	0.09	NA	0.76	0.11	NA	0.82	0.03	NA	0.92	-0.04	NA	0.92
0	3	-0.01	0.49	0.93	0.01	0.35	0.94	0.01	1.05	0.94	0	0.87	0.94	-0.01	0.69	0.94
0.5	1	-0.01	1	0.94	0.01	1	0.94	0.01	1	0.93	0	1	0.95	-0.01	1	0.94
0.5	2	-0.14	NA	0.68	0.11	NA	0.72	0.08	NA	0.88	0.01	NA	0.95	-0.02	NA	0.96
0.5	3	-0.01	0.46	0.94	0.03	0.35	0.94	0.01	0.97	0.94	0	0.9	0.96	0	0.66	0.96
0.8	1	-0.01	1	0.94	0.01	1	0.93	0.01	1	0.96	0	1	0.97	0	1	0.98
0.8	2	-0.14	NA	0.68	0.11	NA	0.7	0.07	NA	0.9	-0.01	NA	0.97	0	NA	0.97
0.8	3	-0.01	0.5	0.95	0.05	0.35	0.93	0.01	0.82	0.95	0.01	0.92	0.98	0.01	0.64	0.97

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

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

Table B.3: Summary of 2000 simulations with data generated from settings with 10 studies and true Se (Sp)=0.9 (0.95).

		Sp			Se			π			PPV			NPV		
Corr[a]	Model[b]	Bias	RE	CP	Bias	RE	CP	Bias	RE	CP	Bias	RE	CP	Bias	RE	CP
0	1	0	1	0.93	0	1	0.94	0.01	1	0.92	-0.02	1	0.94	-0.01	1	0.94
0	2	-0.04	NA	0.78	0.03	NA	0.83	0.2	NA	0.63	0.03	NA	0.9	-0.03	NA	0.89
0	3	-0.01	0.46	0.94	0	0.28	0.9	0.01	1.03	0.93	-0.02	0.77	0.94	0	0.53	0.89
0.5	1	0	1	0.93	0	1	0.93	0.01	1	0.93	-0.01	1	0.95	0	1	0.95
0.5	2	-0.04	NA	0.79	0.04	NA	0.78	0.18	NA	0.72	0.03	NA	0.93	-0.02	NA	0.94
0.5	3	0	0.5	0.95	0.02	0.3	0.86	0.01	1	0.95	-0.01	0.86	0.96	0	0.61	0.88
0.8	1	0	1	0.94	0	1	0.95	0.01	1	0.94	0	1	0.98	0	1	0.97
0.8	2	-0.04	NA	0.8	0.04	NA	0.77	0.17	NA	0.74	0.03	NA	0.96	-0.01	NA	0.97
0.8	3	0	0.47	0.95	0.03	0.31	0.86	0.01	0.85	0.94	0.01	0.92	0.99	0.01	0.53	0.88

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

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

Appendix C

Glossary for abbreviations

MA-DT: meta-analysis of diagnostic test

NMA-DT: network meta-analysis of diagnosic test

Se: sensitivity

Sp: specificity

NPV: negative predictive value

PPV: positive predictive value

LR+: positive likelihood ratio

LR- negative likelihood ratio

π: disease prevlaence

AUC: area under the curve

LMM: linear mixed model

GLMM: generalized linear mixed model

MAR: missing at random

MCAR: missing completely at random

MNAR: missing not at random

CI: credible interval

CP: coverage probability

AIC: Akaike informative criteria

MCMC: Markov Chain Monte Carlo