Bivariate Random Effects Meta-Analysis Models for Diagnostic Test Accuracy Studies Using Arcsine–Based Transformations

BIVARIATE RANDOM EFFECTS META-ANALYSIS MODELS FOR DIAGNOSTIC TEST ACCURACY STUDIES USING ARCSINE–BASED TRANSFORMATIONS

BY

ZELALEM FIRISA NEGERI

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2015) McMaster University

(Mathematics & Statistics) Hamilton, Ontario, Canada

TITLE:	Bivariate Random Effects Meta-Analysis Models for Diagnostic Test Accuracy Studies Using Arcsine–Based Transformations
AUTHOR:	Zelalem Firisa Negeri M.Sc., (Statistics) McMaster University, Canada
SUPERVISOR:	Professor Joseph Beyene
NUMBER OF PAGES:	xii, 70

Dedicated to

In the memory of my father Firisa Negeri and lovely brother Amayu Firisa!!!

Abstract

A diagnostic test identifies patients according to their disease status. Different meta-analytic models for diagnostic test accuracy studies have been developed to synthesize the sensitivity and specificity of the test. Because of the likely correlation between the sensitivity and specificity of a test, modeling the two parameters using a bivariate model is desirable. Historically, the logit transformation has been used to model sensitivity and specificity pairs from multiple studies as a bivariate normal.

In this thesis, we propose two transformations, the arcsine square root and the Freeman-Tukey double arcsine transformation, in the context of a bivariate random-effects model to meta-analyze diagnostic test accuracy studies. We evaluated the performance of the three transformations (the commonly used logit and the proposed transformations) using an extensive simulation study in terms of bias, root mean square error and coverage probability. We illustrate the methods using three real data sets.

The simulation study results showed that, for smaller sample size and higher values of sensitivity and specificity, the proposed transformations are less biased, have smaller root mean square error and better coverage probability than the standard

logit transformation regardless of the number of studies. On the other hand, for large sample sizes, the usual logit transformation is less biased and has better coverage probability regardless of the true values of sensitivity, specificity and number of studies. However, when the sample size is large, the logit transformation has better root mean square error for moderate and large number of studies. The point estimates of the two parameters, sensitivity & specificity, from the methods using the three real data sets follow patterns similar to those reported by our simulation.

Acknowledgements

First and foremost, I am thankful to my God, Lord Jesus Christ with whom I attained this stage and grow to be successful.

My deepest gratitude and acknowledgment goes to my supervisor and mentor, Professor Joseph Beyene, whose scientific guidance, constant effort, and insight throughout the research were invaluable in the completion of this thesis. It was a great honor for me that I worked with him. I would also like to thank the thesis examining committee members: Professor Narayanaswamy Balakrishnan and Professor Roman Viveros-Aguilera for their insightful comments and valuable critique.

I am grateful and want to acknowledge all of my colleagues in the statistics for integrative genomics and meta-analysis (SIGMA) research group for the great time we had throughout the thesis writing process and our studies. Particularly, I want to thank Mateen Shaikh, for his excellent guide and support throughout the simulation coding and Binod Neupane for the insightful conceptual discussions.

Last but not least, I would like to thank my professors Dr. Angelo Canty, Dr. Fred Hoppe and Dr. Ryan Browne for their effort and support throughout the course work.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

In this section we present a brief and general framework for the thesis.

1.1.1 Meta-Analysis and Diagnostic Test Accuracy

Meta-Analysis

Meta-analysis (MA) can be defined as the statistical aggregation of harmonious effect sizes assigning large weight to studies having small variability and larger precision (Kovalchik, 2013). MA has roots in the mid 1980s, and evolved alongside systematic reviews (Borenstein et al., 2009). In the past two decades, MA has been applied in a wide array of fields including medicine, pharmacy, education, psychology, criminology, business, and ecology.

Once a research question has been identified, MA should begin with a 'systematic

review'. A systematic review deals with the comprehensive procedure of assessing the literature about the topic of interest, choosing the literature satisfying a preset selection criteria, extracting relevant data from the chosen literature and appraising the quality of the systematic review.

Diagnostic Test Accuracy

A diagnostic test refers to a procedure identifying or categorizing patients in accordance to their disease status (with or without disease). The test is classified accurate if it has achieved its fundamental objective of classifying the patients according to their real health status, which in practice can only be achieved by comparing with either the gold or reference standard test. A test that has the ability to correctly classify patients is called the gold standard test. However, because this test can be expensive or does not even exist, a reference test is used instead.

In medicine, the accuracy of a diagnostic test is crucial as inaccuracies result in wrong treatment for the patient. A diagnostic test is inaccurate when the test has a positive result for a patient without the disease, or when the test has a negative result for a patient with the disease. Although neither scenario is preferred by a clinician, the consequences of the latter is usually more severe than the former as it can delay or completely preclude necessary treatment and this could harm the patient more than being given treatment unnecessarily. Therefore, there is greater emphasis in preventing a mis-diagnosis when the patient truly has the disease. Because of the potential for mis-diagnosis, it is important to quantify the accuracy of a diagnostic test usually in comparison to the reference or gold standard test.

1.1.2 Review of Meta-Analytic Models for Diagnostic Test Studies and Their Assumptions

Several models for MA of diagnostic test accuracy (DTA) have been proposed in the past two decades, all methods having their own limitations and strengths. Here, we will be focusing on a brief discussion of those past and current attempts made to synthesize and model diagnostic test studies. We put more theoretical and mathematical details for the models in the 'Methods' section of the thesis in Chapter 2.

Meta-Analytic Models

The model proposed by Moses et al. (1993) is the oldest and it fits a fixed-effects (FE) linear model aggregating the sensitivities and specificities of a diagnostic test. They also proposed a method of constructing a summary receiver operating characteristic (SROC) curve, a great tool that aids in assessing the performance of a diagnostic test.

On the other hand, a growing number of random-effects (RE) bivariate normal models (Rutter and Gatsonis, 2001; Reitsma et al., 2005; Arends et al., 2008) have emerged in the past 15 years. Their basic assumption of the bivariate normal model to synthesize the logit transformed sensitivity and specificity of a diagnostic test is due to the possible (negative) correlation between the two parameters mainly because of threshold variability. The model by Reitsma et al. (2005) is relatively straightforward to apply than the other two models. In fact the two models (Rutter and Gatsonis, 2001; Reitsma et al., 2005) are re-parameterizations of each other and

produced the same results when applied to the same data (Reitsma et al., 2005; Arends et al., 2008). They differ only in the way they constructed their model and parameter estimation techniques. The models by Reitsma et al. (2005) and Arends et al. (2008) use the classical approach and that of Rutter and Gatsonis uses a Bayesian parameter estimation approach.

Kuss et al. (2013) proposed a bivariate copulas model incorporating the correlation between sensitivity and specificity, assuming the beta-binomial instead of the binomial distribution for the number of true positives and negatives for each study arm. They proposed three copula models: Clayton, Gauss and Plackett. They assessed the performance of the copula models with respect to the standard bivariate normal model. Their simulation study pointed out that when data were simulated from one of the copula models, the Plackett copula model performed better than others and all of the copula models tend to be more robust to the high correlation case than the standard model (Kuss et al., 2013).

A more recent model to meta-analyze the sensitivity and specificity of a diagnostic test was that of Eusebi et al. (2014) 'latent class bivariate model' (LCBM) which is an extension of the bivariate model of Chu and Cole (2006), that assumes a binomial distribution instead of normal for the within study variability. According to their simulation results, the latent class bivariate model tends to have better performance in terms of bias and coverage of confidence intervals than the standard bivariate RE model (Eusebi et al., 2014).

Meta-Analytic Model Assumptions

The above discussed models (Moses et al., 1993; Rutter and Gatsonis, 2001; Reitsma et al., 2005, Arends et al., 2008) have assumed the bivariate normal model for the logit transformed sensitivities and false positive rates rather than for the original values. In this subsection, we will discuss the standard and new transformation-based assumptions made by different researchers in MA of DTA.

A common transformation of a single proportion like sensitivity or specificity is the logit transformation. It has been adopted for the bivariate normal modeling of diagnostic test studies. Therefore, the procedure of logit transforming sensitivity and specificity and fitting a bivariate model is a common practice in the literature (Moses et al., 1993; Rutter and Gatsonis, 2001; Reitsma et al., 2005; Arends et al., 2008) and accepted as a standard method of transformation due to its ease for interpretation.

Recently, other transformations such as the probit and complementary log-log (clog-log) have been proposed and compared to the logit with respect to misspecification based on transformations on the estimation of median sensitivity and specificity (Chu et al., 2010). Chu et al. (2010) analyzed the three link functions, and concluded that misspecification of the three link functions does not alter the estimation of the median sensitivity and specificity, but it does alter standard error of the estimates and the coverage probability of the 95% confidence interval. Moreover, they have reported from their simulation studies that the clog-log is tractable due to its asymmetric property, has the highest median area under the curve (AUC) of the SROC & coverage probability and unaltered by the misspecification of the link functions (Chu et al., 2010).

More recently, the parametric transformation by the name 't_α family of transformations' has been shown to be a better alternative to the usual logit transformation (Doebler et al., 2012). They proposed the new transformation to overcome some of the shortcomings of the standard method. They argued that the 't_α family of transformation' overcomes the 'subjectivity' in the choice of the logit transformation, the 'implicit' assumption of linearity and failure to meet the distributional assumption for large values of sensitivities (Doebler et al., 2012). They have compared the performance of the transformations using both a simulation study and real data by fitting a bivariate normal distribution to the transformed data. They concluded that the normality assumption is well satisfied when using their transformation even when the assumption fails in the case of logit transformation. They have indicated that for data sets with fewer zero cell counts and moderate sensitivities, their transformation appeared to have lower AIC but favorable for other data sets with large sensitivity and small false positive rate (Doebler et al., 2012).

1.2 Significance of the Study

Another popular transformation for proportions to approximate normality is the arcsine square root, and its variant, the Freeman–Tukey double arcsine transformation (Freeman and Tukey, 1950). The arcsine square root transformation has widely been used in the analysis of single proportion.

In the context of meta-analyses of interventions, only a single report by Trikalinos

et al. (2013) recently compared the performance of the arcsine square root transformation to the standard logit transformation and untransformed proportions using a simulation study. According to the results of their study, they have recommended to use the variance stabilizing arcsine square root transformation over the standard logit transformation and untransformed data for meta-analysis of interventions.

In the context of DTA, Fokom-Domgue et al. (2015) employed arcsine square root transformation for pooling sensitivity and specificity to appraise the prevalence of cervical intraepithelial neoplasia grade 2 or worse (CIN2+) and positivity test of these screening methods, and fitted a bivariate RE model of Reitsma et al. (2005) with the aim of comparing the performance of three tests for primary cervical cancer screening in sub-Saharan Africa.

Kocken et al. (2012) employed the double arcsine transformation of Freeman–Tukey using the univariate methods to meta-analysis of diagnostic studies. The aim of their study was to evaluate the performance of 'high-risk human papillomavirus testing', 'Cytology test' and 'Co-testing' for high-grade cervical disease.

Although efforts have been made to use the variance stabilizing transformations in univariate MA of DTA and interventions, to our best knowledge, the variance stabilizing transformations have not been used in the context of diagnostic test studies using the recommended bivariate random-effects model. Hence, the purpose of this thesis is to fill the gap in the literature by proposing two different variations of arcsine transformation –the arcsine square root and the Freeman-Tukey double arcsine transformation in the context of bivariate RE model for DTA studies.

1.3 Motivating Examples

In this section we introduce three data sets that will be further analyzed in Chapter 4 to demonstrate how the transformations introduced perform in real world situations. For now, a forest plot for each of the data is presented. A forest plot is a graphical tool which displays the studies under investigation including each studies average effect measure (sensitivity or specificity) with their corresponding confidence intervals. One of the major uses of the forest plot is that, one can visually inspect for the presence or absence of heterogeneity between the studies so that the appropriate model, fixed - or random-effects, could be fitted to the data. We will discuss in detail about the fixed vs random effects model in Chapter 2. We have chosen the data sets so that they represent the scope of MA in terms of sample size, degree of heterogeneity and number of studies.

1.3.1 The 'Children US' Data

This is a data on ultrasonography (US) test for diagnosis of appendicitis in children and was used in the study of Doria et al. (2006). A forest plot for this data is shown in Figure 1.1 below.

The data consists of 23 studies and the average number of children with and without disease are 77 and 254, respectively. Doria et al. (2006) analyzed the sensitivity and specificity separately by fitting a FE model. As can be observed in Figure 1.1, although there is no substantial heterogeneity between studies, the FE

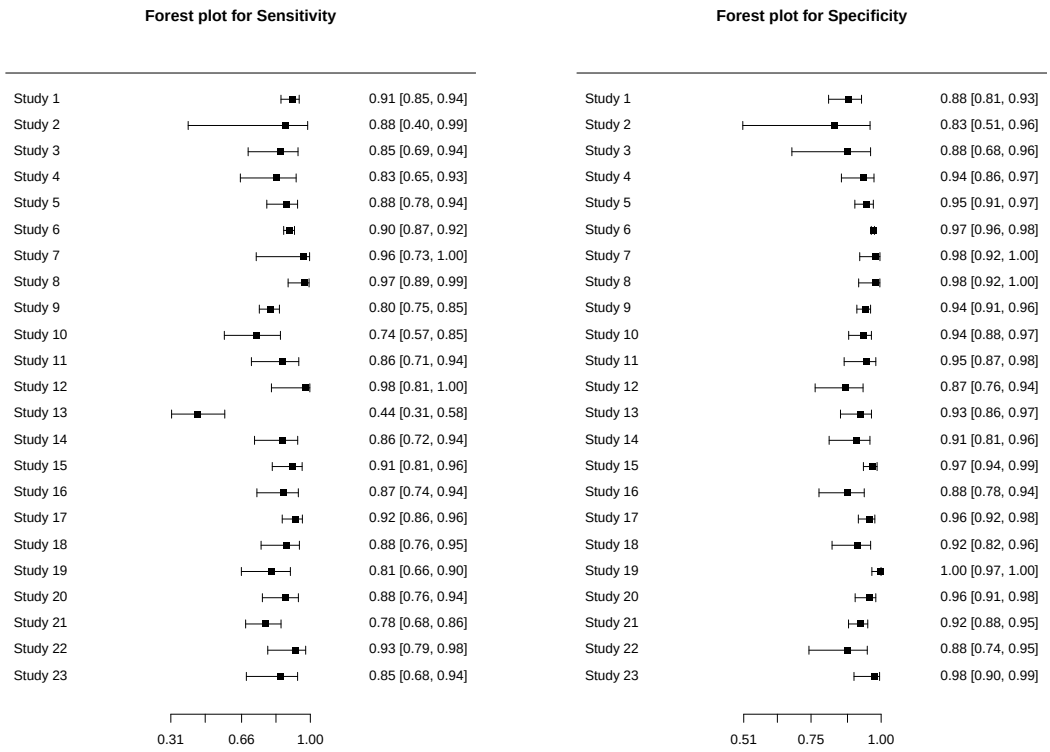


Figure 1.1: Forest plot for sensitivity and specificity of 'Children US' data.

model that Doria et al. (2006) fitted ignores the possible correlation that exists between sensitivity and specificity.

1.3.2 The 'VIA' Data

The 'visual inspection with acetic acid' ('VIA') data is a published data obtained from the study by Fokom-Domgue et al. (2015). There are 10 studies and an average sample size of 145 and 4,778 in each study arm, with & without disease, respectively. A forest plot for this data is shown in Figure 1.2.

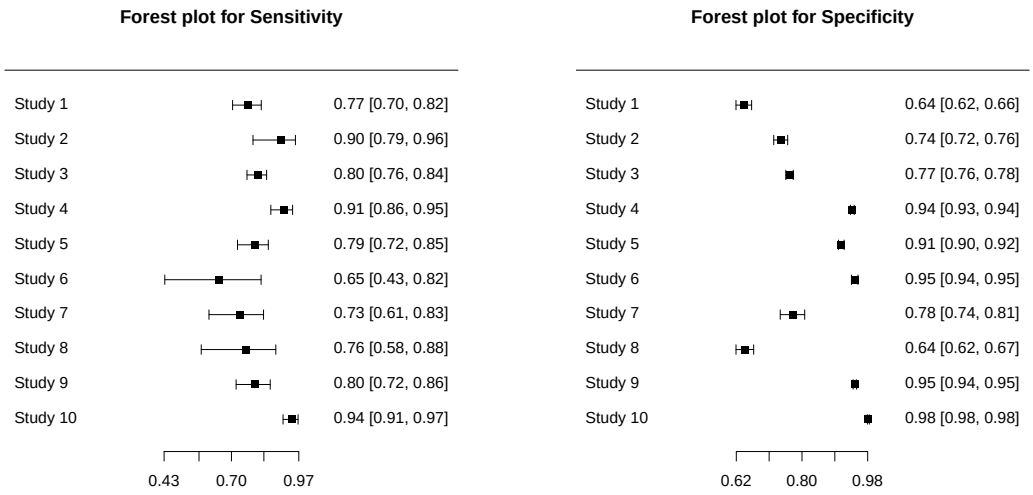


Figure 1.2: Forest plot for sensitivity and specificity of 'VIA' data.

Figure 1.2 clearly shows that there is heterogeneity between studies, particularly for specificity. Fokom-Domgue et al. (2015) recognized this and fitted a bivariate RE model of Reitsma et al. (2005) to compare the accuracy of the 'VIA' testing with 'visual inspection with Lugols iodine' (VILI), and 'human papillomavirus' (HPV) testing for primary cervical cancer screening in sub-Saharan Africa.

1.3.3 The 'Cytology' Data

The third data that we use as a motivating example is the 'Cytology' data of Kocken et al. (2012). It includes eight studies and the average sample size in the diseased

(with disease) and non-diseased (without disease) group is 19 and 170 respectively. A forest plot for this data is displayed in Figure 1.3.

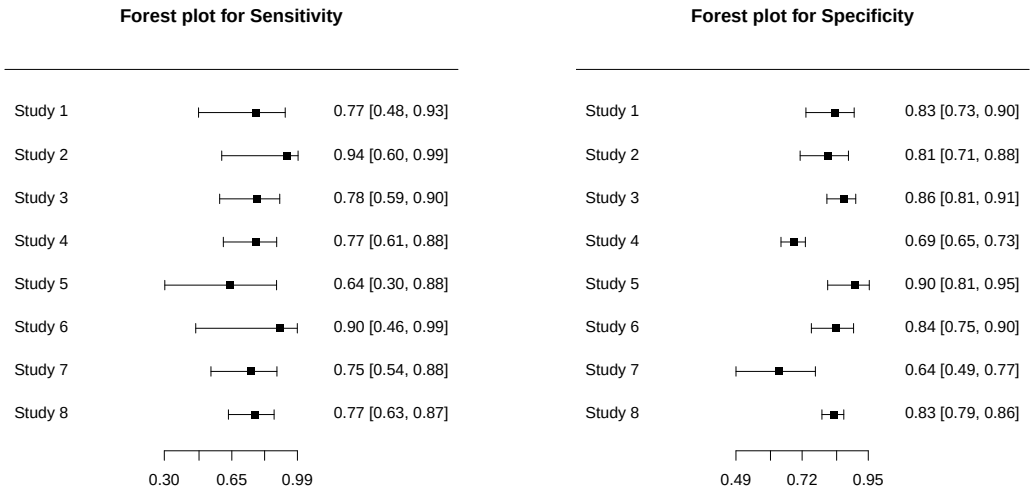


Figure 1.3: Forest plot for sensitivity and specificity of 'VIA' data.

Since the extent of heterogeneity varies for sensitivity and specificity as can be seen from Figure 1.3, Kocken et al. (2012) took that into account by fitting a RE model for specificity and a FE model for sensitivity.

We will investigate the impact of fitting the recommended bivariate RE model on the estimates of the 'Children US' and 'Cytology' data sets, and the effect of the proposed transformations on that of the 'VIA' data later in Chapter 4.

1.4 Scope of the Study

In this thesis, we compare the performances of the proposed transformations with the standard logit transformation while fitting the bivariate RE model for DTA studies. We have carried out an extensive simulation study and used several evaluation criteria including bias, root mean square error (RMSE) and coverage probability. Finally, we have applied the methods to the three real motivating data sets.

The remaining of this thesis is organized as follows: Chapter 2 presents the technical and statistical aspects of meta-analyses of diagnostic test studies. In Chapter 3 we present the core of the thesis —simulation design and results. In Chapter 4, we present results of the proposed methods and the standard approach on real data sets, and we close the thesis with a summary and discussion in Chapter 5.

Chapter 2

Methods

2.1 Diagnostic Test Accuracy Data Structure and Parameters

In this section we present the data structure for the MA of DTA studies and then discuss briefly the parameters often used in measuring the accuracy of a diagnostic test.

2.1.1 Diagnostic Test Accuracy Data Structure

The data structure for a diagnostic test outcome could be presented in a 2×2 table similar to that of the intervention/treatment outcome. However, there are two parameters of interest (sensitivity and specificity) that are commonly used to measure the accuracy of a diagnostic test, unlike the single effect-size employed in MA of interventions. Table 2.1 summarizes the typical data structure for a DTA

studies.

Table 2.1: Data Structure of a Diagnostic Test Study

		Disease Status	
		With disease	Without disease
Test Result	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)
	Total	$n_1 = TP + FN$	$n_2 = FP + TN$

2.1.2 Diagnostic Test Accuracy Parameters

The two most commonly used measures of diagnostic test accuracy (DTA) are test sensitivity (true positive rate (TPR)) and test specificity (true negative rate (TNR)). The conditional probability of obtaining true positive test result given the test is performed on an individual in the diseased group is known as TPR, whereas, TNR refers to the probability of getting true negative test result conditional on the test being performed on a person in the non-diseased group. Another important quantity related to these two measures of test accuracy is the false positive rate (FPR), which is defined as the conditional probability of obtaining false positive test result given that the test is applied on an individual in the non-diseased subject. FPR is usually used in place of TNR in several meta-analytic modeling of DTA literatures and of course the two are related as $FPR = 1 - TNR$.

Mathematically, these two parameters are defined as follows:

1. Sensitivity of a diagnostic test:

$$\textit{Sensitivity} = \Pr(\text{Test Result is Positive}|\text{Diseased}) = \textit{TPR}. \tag{2.1}$$

2. Specificity of a diagnostic test:

$$\textit{Specificity} = \Pr(\text{Test Result is Negative}|\text{Non-Diseased}) = \textit{TNR}. \tag{2.2}$$

Given the observed data for a DTA study of Table 2.1, we can estimate those parameters easily in the following manner;

$$\widehat{TPR} = \frac{TP}{n_1}, \qquad \widehat{TNR} = \frac{TN}{n_2}. \tag{2.3}$$

2.2 Univariate Fixed- and Random-Effects Models for Diagnostic Test Accuracy Studies

Any quantitative research, at least presents a descriptive summary and/or employs a statistical model to make meaningful inference. Similarly, there are different methods proposed for meta-analyzing the two parameters of interest in diagnostic test studies. This section discusses some of these methods for MA of DTA. Even though several models have been proposed in order to meta-analyze sensitivity and specificity of a diagnostic test, we will only review some of the standard methods that have been widely used in the literature.

Before we start the mathematical description of the models, we will make note of the following notations. Let the observed number of true positives (TP), and total count for the diseased group for study i be denoted by u_i and n_{1i}, respectively. Let for the i^{th} study, v_i and n_{2i} denote the observed number of false positives (FP) and total count for non-diseased group, respectively. Then, a 2×2 table similar to the one we have seen earlier could be obtained for study i as follows:

Table 2.2: Data Structure of Diagnostic Test Results for study i

		Disease Status	
		With disease	Without disease
Test Result	Positive	u_i	v_i
	Negative	$n_{1i} - u_i$	$n_{2i} - v_i$
	Total	n_{1i}	n_{2i}

If we denote the true sensitivity (TPR) and 1-specificity (FPR) for the i^{th} study by p_i and q_i respectively, then their estimates are:

$$\hat{p}_i = \frac{u_i}{n_{1i}}, \quad \text{and} \quad \hat{q}_i = \frac{v_i}{n_{2i}}. \tag{2.4}$$

Now, if θ_{1i} and θ_{2i}, $i = 1, 2, ..., k$, represent the true logit transformed sensitivities and false positive rates ($\theta_{1i} = \text{logit}(p_i) = \log(\frac{p_i}{1-p_i})$ and $\theta_{2i} = \text{logit}(q_i) = \log(\frac{q_i}{1-q_i})$) for each study, then the within-study variances of the logit transformed estimated sensitivities ($\hat{\theta}_{1i}$) and false positive rates ($\hat{\theta}_{2i}$), respectively, are modeled by the fact that ($\hat{\theta}_{1i}$) and ($\hat{\theta}_{2i}$) are independent and approximately normally distributed at study

level. That is:

$$\hat{\theta}_{1i}|\theta_{1i} \approxeq N(\theta_{1i}, s_{1i}^2), \text{ and } \hat{\theta}_{2i}|\theta_{2i} \approxeq N(\theta_{2i}, s_{2i}^2) \tag{2.5}$$

where for large sample sizes (n_{1i} and n_{2i}), the within-study variances are estimated as:

$$s_{1i}^2 = \frac{1}{u_i} + \frac{1}{n_{1i} - u_i}, \text{ and } s_{2i}^2 = \frac{1}{v_i} + \frac{1}{n_{2i} - v_i}. \tag{2.6}$$

2.2.1 Univariate Fixed-Effect Model for Diagnostic Test Studies

Hedges and Vevea (1998) described the FE model approach as a 'conditional analysis' in a sense that the research outcome from this type of model will only be inferred subject to a constraint —the number of studies included in the model. This is different from the definitions surrounding the model for several years—which were assuming that the model is good only under homogeneity assumption. However, Hedges and Vevea (1998) argued that the model would still be applicable under the assumption of heterogeneity as long as the inference is made constraint to the number of studies included in the meta-analysis.

If we have k studies to synthesize, then the fixed-effects model assumes that

$$\hat{\theta}_{1i} = \theta_1 + \varepsilon_{1i}, \text{ or } \hat{\theta}_{2i} = \theta_2 + \varepsilon_{2i}, i = 1, 2, ..., k, \tag{2.7}$$

where $\hat{\theta}_{1i}$ and $\hat{\theta}_{2i}$ are the observed values of the logit transformed sensitivity and 1-specificity, respectively, each assumed to be normally distributed with true but unknown common means θ_1 & θ_2 and known within-study variances s_{1i}^2 and s_{2i}^2,

respectively. ε_{1i} and ε_{2i} are independent random variables for sampling error both assumed to be normal with mean 0 and variances ν_{1i} and ν_{2i}, respectively.

The unknown parameters, θ_1 and θ_2, of this model could be estimated using either the weighted least squares method —using the reciprocal of the within study variability as a weight or the unweighted average method. The former finds the estimator that minimizes the variance and is equivalent to the maximum likelihood (ML) estimator (Hedges and Vevea, 1998). Accordingly, the weighted least squares estimator of θ_j, $j = 1, 2$ is given by;

$$\hat{\theta}_{1w} = \frac{\sum_{i=1}^{k} w_{1i}\hat{\theta}_{1i}}{\sum_{i=1}^{k} w_{1i}} \text{ and } \hat{\theta}_{2w} = \frac{\sum_{i=1}^{k} w_{2i}\hat{\theta}_{2i}}{\sum_{i=1}^{k} w_{2i}}, \text{ respectively} \tag{2.8}$$

where $w_{1i} = \frac{1}{s_{1i}^2}$ and $w_{2i} = \frac{1}{s_{2i}^2}$ as described above.

The weighting scheme thus assigns higher weight to the effect measures having smaller variance (or higher precision), and vice versa.

In order to make statistical inference, obtaining the point estimates alone is not sufficient unless we associate a measure of variability to the estimator. Since the k studies are assumed to be independent and the weights, w_{ji}, $j = 1, 2$, are constants, it is straightforward to show that the variance of the fixed-effects estimators $\hat{\theta}_{jw}$ is $var(\hat{\theta}_{jw}) = \frac{1}{\sum_{i=1}^{k} w_{ji}}$. That is:

$$\begin{aligned} var(\hat{\theta}_{jw}) &= var\left(\frac{\sum_{i=1}^{k} w_{ji}\hat{\theta}_{ji}}{\sum_{i=1}^{k} w_{ji}}\right) = \frac{\sum_{i=1}^{k} var(w_{ji}\hat{\theta}_{ji})}{(\sum_{i=1}^{k} w_{ji})^2} = \frac{\sum_{i=1}^{k} w_{ji}^2 var(\hat{\theta}_{ji})}{(\sum_{i=1}^{k} w_{ji})^2} = \frac{\sum_{i=1}^{k} w_{ji}^2 s_{ji}^2}{(\sum_{i=1}^{k} w_{ji})^2} \\ &= \frac{\sum_{i=1}^{k} w_{ji}^2 . \frac{1}{w_{ji}}}{(\sum_{i=1}^{k} w_{ji})^2} = \frac{\sum_{i=1}^{k} w_{ji}}{(\sum_{i=1}^{k} w_{ji})^2} = \frac{1}{\sum_{i=1}^{k} w_{ji}}. \end{aligned}$$

Hence the nominal 95% Wald-type confidence interval for θ_j is given by: $\hat{\theta}_{jw} \mp 1.96 \times se(\hat{\theta}_{jw})$; where $se(\hat{\theta}_{jw})$ is the standard error of $\hat{\theta}_{jw}$, obtained by taking the square root of $var(\hat{\theta}_{jw})$.

2.2.2 Univariate Random-Effects Model for Diagnostic Test Studies

Unlike the FE model, the random-effects model makes an unconditional inference using the effect size of interest (Hedges and Vevea, 1998). The basic assumption about the RE model is that the k independent studies might differ in their methods or behavior of the selected samples (Viechtbauer, 2010). This heterogeneity among the studies thus leads to the assumption of randomness about the unknown true effect θ_j and hence the name RE model. That is to say, the true effect size θ_j, by itself is a random sample from a larger population and has its own distribution —the assumption that leads to an unconditional inference about the whole population.

The RE model is mathematically defined as;

$$\hat{\theta}_{1i} = \theta_{1i} + \varepsilon_{1i} = \mu_1 + \delta_{1i} + \varepsilon_{1i}, i = 1, 2, ..., k \text{ for sensitivity,} \tag{2.9}$$

and

$$\hat{\theta}_{2i} = \theta_{2i} + \varepsilon_{2i} = \mu_2 + \delta_{2i} + \varepsilon_{2i}, i = 1, 2, ..., k \text{ for 1-specificity} \tag{2.10}$$

where $\hat{\theta}_{1i}$, $\hat{\theta}_{2i}$ and ε_{ji} are as defined before; θ_{1i} and θ_{2i} are the unknown true logit transformed sensitivity and 1-specificity for each study both assumed to be normally distributed with means μ_1 & μ_2, and the between study variances τ_1^2 & τ_2^2,

respectively. μ_1 & μ_2 are the mean true logit transformed sensitivity & 1-specificity respectively, and δ_{ji} is an error of θ_{ji} as an estimate of μ_j for $j = 1, 2$.

Unlike the FE model which estimates only one parameter (θ_j), the RE models estimate two parameters, the mean true effect (μ_j) and the between study heterogeneity (τ_j^2). Once the heterogeneity component τ_j^2 is estimated using the method of moments, MLE or restricted maximum likelihood (REML) estimation method, then the mean true effect (μ_j) could be estimated by a weighted least squares method using the reciprocal of the sum of the within- and between-study variances as weights:

$$\bar{\mu}_1 = \frac{\sum_{i=1}^{k} w_{1i}^{*}\hat{\theta}_{1i}}{\sum_{i=1}^{k} w_{1i}^{*}} \text{ and } \bar{\mu}_2 = \frac{\sum_{i=1}^{k} w_{2i}^{*}\hat{\theta}_{2i}}{\sum_{i=1}^{k} w_{2i}^{*}}, \tag{2.11}$$

where $w_{1i}^{*} = \frac{1}{s_{1i}^2+\hat{\tau}_1^2}$ and $w_{2i}^{*} = \frac{1}{s_{2i}^2+\hat{\tau}_2^2}$ are the RE weights to be computed for each study.

Again the variance of the mean true effect estimator $\bar{\mu}_j$, $j = 1, 2$, could easily be shown to be the reciprocal of the sum of the RE weights (w_{ji}^{*}). That is, $var(\bar{\mu}_j) = \frac{1}{\sum_{i=1}^{k} w_{ji}^{*}}$, and, the 95% confidence interval for the mean true effect is given by: $\bar{\mu}_j \mp 1.96 \times se(\bar{\mu}_j)$, where $se(\bar{\mu}_j)$ is the standard error of $\bar{\mu}_j$.

Both the fixed- and random-effects models discussed here are now discouraged in applications since they ignore the possible correlation that exists between sensitivity and specificity due to threshold variability. Instead, new methods either fixed- or random-effects, that combine both effect measures or parameters of DTA have been encouraged to use in practice. In the following section, we will review two of the most widely used methods of this kind and finally discuss our method in the context of the second method.

2.3 Bivariate Fixed- and Random-Effects Models for Diagnostic Test Accuracy Studies

So far we have seen the concepts and methods for univariate MA of DTA studies. In this section we review two bivariate models, the Moses et al. (1993) summary receiver operating characteristic (SROC) curve and Reitsma et al. (2005) bivariate random-effects model.

2.3.1 The Summary Receiver Operating Characteristic (SROC) Curve

The Moses et al. (1993) SROC curve is one of the oldest and standard methods to analyze sensitivity and specificity of a diagnostic test study. The general idea behind the SROC curve is to first plot the difference between logit transformed TPR and FPR against the sum of logit transformed TPR and FPR estimates obtained from the k independent studies on the ROC curve. Then a regression line is fitted to the points using the difference between logit transformed TPR and FPR as a dependent variable. The fitted points from the regression line are then back transformed to the ROC curve for each given value of FPR to complete the SROC curve (Moses et al., 1993).

Mathematically, the regression equation of the SROC curve is given by:

$$D = a + bS, \tag{2.12}$$

where D = logit(TPR) - logit(FPR) = log($\frac{TPR}{1-TPR}$) - log($\frac{FPR}{1-FPR}$) denotes the diagnostic log-odds ratio which identifies the diseased from the non-diseased; and, the measure of diagnostic threshold, S = logit(TPR) + logit(FPR)=log($\frac{TPR}{1-TPR}$) + log($\frac{FPR}{1-FPR}$); a & b are the intercept and slope of the regression line, respectively.

Once a robust or weighted least square regression of D on S is fitted, then the back transformed values of TPR for each values of the FPR, which completes the SROC curve is given by (Walter, 2002):

$$TPR = \frac{\exp(\frac{a}{1-b})(\frac{FPR}{1-FPR})^{(1+b)/(1-b)}}{1+\exp(\frac{a}{1-b})(\frac{FPR}{1-FPR})^{(1+b)/(1-b)}}. \tag{2.13}$$

Moses et al. (1993) proposed the Q^* statistic, the point at which sensitivity of a test equals its specificity, as a global measure of the SROC curve. This point is represented by a straight line from the top right corner to bottom left corner on the ROC curve and interpreted as the constant diagnostic threshold of all the studies such that the test definitely discriminates the diseased from the non-diseased. On the other hand, Walter (2002) discussed in detail the importance of another global measure of SROC curves known as the area under the curve (AUC). It has been discussed by Walter that a test that differentiates between the diseased and non-diseased cases randomly, has an AUC = 0.5, an ideal test has AUC of 1 and an imperfect test has AUC of 0.

Though the SROC method was widely used and straightforward to apply, it has been criticized (Rutter and Gatsonis 2001; Reitsma et al., 2005; Arends et al., 2008) and alternative methods have been proposed. One of the criticism is, it ignores the possible heterogeneity among studies, assuming the threshold values and sampling

variability to be the only sources of variation and hence fits the FE model. However, as it has been indicated in different literatures (Rutter and Gatsonis 2001; Reitsma et al., 2005; Arends et al., 2008), there should be some kind of heterogeneity to be considered between studies that might result from the possible negative correlation between TPR and TNR, disease prevalence, laboratory errors, study design, and patient selection, to mention few.

Another criticism is that (Arends et al., 2008), the SROC curve does not take into consideration the 'measurement error' for the variable S in the regression equation of (2.12) which leads to bias in both the intercept (a) and slope (b), ignores the possible within-study correlation between the two regression variables D and S, does not use the ideal weighting scheme while fitting the weighted least squares regression as it takes the inverse of the variance of diagnostic odds ratio (D) as a weight, and suffers from the unnecessarily adding of 0.5 to all cells of the 2×2 table for continuity correction.

Below, we will discuss the bivariate RE model that overcomes almost all of the shortcomings of the Moses et al. (1993) method.

2.3.2 The Bivariate Random-Eeffects Models for Diagnostic Test Accuracy Studies

The bivariate RE models have the advantage over the standard SROC curve method of Moses et al. (1993) of taking into consideration the possible heterogeneity between studies and thus fit the RE model. Since they assume a distribution for both the TPR and TNR, they will not suffer the problem of measurement error (Arends et al.,

2008). Moreover, the possible negative correlation of TPR and TNR is incorporated by fitting the RE model in the bivariate models. The weighting problem of the SROC curve could also be solved in the bivariate models by using both the within- and between-study variance-covariance matrices as a weight.

However, all the bivariate methods (Rutter and Gatsonis 2001; Reitsma et al., 2005; Arends et al., 2008) themselves do not escape from the problem of shifting the SROC curve from its ideal top left corner position because of the addition of the continuity correction value of 0.5. This will occur as all the methods are based on the trick of taking the logit transformation of sensitivities and specificities which leads to undefined values of $\hat{\theta}_{1i}$ & $\hat{\theta}_{2i}$ and their within study variances for a study with zero cell count. Arends et al. (2008) have suggested as a solution to use the exact (Binomial) distribution for the observed number of positives in both with and without disease cases to overcome this problem.

Reitsma et al. (2005)'s Bivariate RE Model

Among the available bivariate RE models proposed to simultaneously analyze the sensitivity and specificity of diagnostic studies, the Reitsma et al. (2005) approach is relatively easy to understand, widely used, and, most recently has been implemented in a freely available statistical software R (R Core Team, 2015) in a package called 'mada' (Doebler, 2012).

The basic assumption of the Reitsma et al. (2005) RE model is that the true logit transformed sensitivities (θ_{1i}) and false positive rates (θ_{2i}) are distributed as a bivariate normal with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ and between-study covariance

matrix $\mathbf{\Sigma}$. Mathematically,

$$\begin{pmatrix} \theta_{1i} \\ \theta_{2i} \end{pmatrix} \sim N_2 \left[\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma \\ \sigma & \sigma_2^2 \end{pmatrix} \right], \tag{2.14}$$

where μ_1 and μ_2 are the true means for the logit transformed sensitivities and false positive rates, respectively, and, the components of the between-study covariance matrix $\mathbf{\Sigma}$; σ_1^2, σ_2^2, and σ denote, respectively, the true variances of θ_{1i}, θ_{2i} and the possible (positive) covariance between θ_{1i} and θ_{2i}.

At each study level, one can assume the observed variances of the logit transformed estimated sensitivities ($\hat{\theta}_{1i}$) and false positive rates ($\hat{\theta}_{2i}$) as fixed, and, hence the usual FE model would be extended to;

$$\begin{pmatrix} \hat{\theta}_{1i} \\ \hat{\theta}_{2i} \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \theta_{1i} \\ \theta_{2i} \end{pmatrix}, \mathbf{S_i} = \begin{pmatrix} s_{1i}^2 & 0 \\ 0 & s_{2i}^2 \end{pmatrix} \right]. \tag{2.15}$$

Finally, the linear mixed model of Reitsma et al. (2005) is completed by combining the above two models. This defines the bivariate random-effects model given by:

$$\begin{pmatrix} \hat{\theta}_{1i} \\ \hat{\theta}_{2i} \end{pmatrix} \sim N_2 \left[\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \mathbf{\Sigma} + \mathbf{S_i} = \begin{pmatrix} \sigma_1^2 + s_{1i}^2 & \sigma \\ \sigma & \sigma_2^2 + s_{2i}^2 \end{pmatrix} \right]. \tag{2.16}$$

Parameter Estimation

In Reitsma et al. (2005) bivariate RE model, given in (2.16), there are five parameters: μ_1, μ_2, σ_1^2, σ_2^2, and σ, that need to be estimated in order to make statistical inference. The two most commonly used methods are likelihood-based—ML and REML. Basically, the joint probability distribution function (pdf) of $\hat{\theta}_{1i}$ and $\hat{\theta}_{2i}$ from Reitsma et al. (2005) model is:

$$f(\hat{\theta}_{1i}, \hat{\theta}_{2i}; \boldsymbol{\mu}, \boldsymbol{\Sigma}_i) = (2\pi)^{-1}|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}_i^{-1}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\mu})\right), \tag{2.17}$$

where $\hat{\boldsymbol{\theta}}_i$=$(\hat{\theta}_{1i}, \hat{\theta}_{2i})'$ and $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma} + \mathbf{S_i}$.

From the standard statistical theory of distribution, the joint distribution of the untransformed (original) sensitivities (p_i) and false positive rates (q_i) is given by;

$$f(p_i, q_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}_i) = f_{\hat{\theta}_{1i}, \hat{\theta}_{2i}}(p_i, q_i)|\mathbf{J}|, \tag{2.18}$$

where $\mathbf{J}$ is the jacobian matrix of the transformation which in our case is given to be:

$$\mathbf{J} = \begin{vmatrix} \frac{d\theta_{1i}}{dp_i} & \frac{d\theta_{1i}}{dq_i} \\ \frac{d\theta_{2i}}{dp_i} & \frac{d\theta_{2i}}{dq_i} \end{vmatrix} = \begin{vmatrix} \frac{1}{p_i(1-p_i)} & 0 \\ 0 & \frac{1}{q_i(1-q_i)} \end{vmatrix} = \frac{1}{p_i q_i (1-p_i)(1-q_i)}. \tag{2.19}$$

Hence, the joint distribution of p_i and q_i that we need to maximize in order to find the estimates of the five parameters becomes:

$$f(p_i, q_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}_i) = \frac{(2\pi)^{-1}|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}}}{p_i q_i (1-p_i)(1-q_i)} \exp\left(-\frac{1}{2}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}_i^{-1}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\mu})\right). \tag{2.20}$$

Then for k independent studies, observed sensitivities ($\hat{p}_i$) and specificities ($\hat{q}_i$), the likelihood and log-likelihood function are:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{k} \frac{(2\pi)^{-1}|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}}}{\hat{p}_i\hat{q}_i(1-\hat{p}_i)(1-\hat{q}_i)} \exp\left(-\frac{1}{2}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}_i^{-1}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\mu})\right), \tag{2.21}$$

and

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \text{const.} - \frac{1}{2}\sum_{i=1}^{k} \log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{k}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}_i^{-1}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\mu}). \tag{2.22}$$

In order to derive the REML estimates of the parameters, we have to derive the restricted likelihood first. To do so, we can easily derive the ML estimator of $\boldsymbol{\mu}$ from (2.22) first:

$$\hat{\boldsymbol{\mu}} = \left(\sum_{i=1}^{k} \boldsymbol{\Sigma}_i^{-1}\right)^{-1} \left(\sum_{i=1}^{k} \boldsymbol{\Sigma}_i^{-1}\hat{\boldsymbol{\theta}}_i\right). \tag{2.23}$$

Next, the profile log-likelihood used to estimate the covariance matrix is obtained by substituting the value of $\hat{\boldsymbol{\mu}}$ from (2.23) into the full log-likelihood function (2.22):

$$l_{profile}(\boldsymbol{\Sigma}) = \text{const.} - \frac{1}{2}\sum_{i=1}^{k} \log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{k}(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\mu}})'\boldsymbol{\Sigma}_i^{-1}(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\mu}}). \tag{2.24}$$

Jennrich and Schluchter (1986) derived the restricted log-likelihood function to be maximized in order to find the REML estimator of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ as:

$$l_{REML}(\boldsymbol{\Sigma}) = \text{const.} - \frac{1}{2}\sum_{i=1}^{k} \log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{k}(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\mu}})'\boldsymbol{\Sigma}_i^{-1}(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\mu}}) - \frac{1}{2}\log|\sum_{i=1}^{k} \Sigma_i^{-1}|. \tag{2.25}$$

Both the unrestricted and restricted log-likelihoods are clearly defined and suitable to be maximized using standard numerical methods such as the Newton-Raphson

method.

The Newton-Raphson method for finding the ML estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ iteratively is given by:

$$\begin{bmatrix} \hat{\boldsymbol{\mu}}_{ML}^{l+1} \\ \hat{\boldsymbol{\Sigma}}_{ML}^{l+1} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{\mu}}^{l} \\ \hat{\boldsymbol{\Sigma}}^{l} \end{bmatrix} - \begin{bmatrix} \mathbf{H}_{\hat{\boldsymbol{\mu}}^{l}\hat{\boldsymbol{\mu}}^{l}} & \mathbf{H}_{\hat{\boldsymbol{\mu}}^{l}\hat{\boldsymbol{\Sigma}}^{l}} \\ \mathbf{H}_{\hat{\boldsymbol{\Sigma}}^{l}\hat{\boldsymbol{\mu}}^{l}} & \mathbf{H}_{\hat{\boldsymbol{\Sigma}}^{l}\hat{\boldsymbol{\Sigma}}^{l}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}_{\hat{\boldsymbol{\mu}}^{l}} \\ \mathbf{S}_{\hat{\boldsymbol{\Sigma}}^{l}} \end{bmatrix}. \tag{2.26}$$

Where $\hat{\boldsymbol{\mu}}_{ML}^{l+1}$ and $\hat{\boldsymbol{\Sigma}}_{ML}^{l+1}$ are the ML method updated values of the 2×1 mean vector and 3×1 variance-covariance vector of the Newton algorithm for, $\hat{\boldsymbol{\mu}}^{l}$ and $\hat{\boldsymbol{\Sigma}}^{l}$ together form the l^{th} iteration of the 5×1 vector of parameters, $\mathbf{H}_{\hat{\boldsymbol{\mu}}^{l}\hat{\boldsymbol{\mu}}^{l}}$, $\mathbf{H}_{\hat{\boldsymbol{\mu}}^{l}\hat{\boldsymbol{\Sigma}}^{l}}$, $\mathbf{H}_{\hat{\boldsymbol{\Sigma}}^{l}\hat{\boldsymbol{\mu}}^{l}}$, and $\mathbf{H}_{\hat{\boldsymbol{\Sigma}}^{l}\hat{\boldsymbol{\Sigma}}^{l}}$ respectively are the 2×2, 2×3, 3×2, and 3×3 matrices of second derivatives with respect to the components of $\boldsymbol{\mu}$ only, $\boldsymbol{\mu}$ & $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}$ & $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ only —all being evaluated at their current values of the iteration forming the 5×5 hessian matrix, $\mathbf{S}_{\hat{\boldsymbol{\mu}}^{l}}$ and $\mathbf{S}_{\hat{\boldsymbol{\Sigma}}^{l}}$ are the 2×1 and 3×1 score vectors evaluated at the current value of iteration, respectively —together forming the 5×1 score vector for the algorithm.

On the other hand, the iterative Newton-Raphson algorithm to find the solution for the covariance matrix using the REML method would be:

$$\left[\hat{\boldsymbol{\Sigma}}_{REML}^{l+1}\right] = \left[\hat{\boldsymbol{\Sigma}}^{l}\right] - \left[\mathbf{H}_{\hat{\boldsymbol{\Sigma}}^{l}\hat{\boldsymbol{\Sigma}}^{l}}\right]^{-1} \left[\mathbf{S}_{\hat{\boldsymbol{\Sigma}}^{l}}\right]. \tag{2.27}$$

Where $\hat{\boldsymbol{\Sigma}}_{REML}^{l+1}$ is the REML method updated value of the 3×1 variance-covariance vector, $\hat{\boldsymbol{\Sigma}}^{l}, \mathbf{H}_{\hat{\boldsymbol{\Sigma}}^{l}\hat{\boldsymbol{\Sigma}}^{l}}$ and $\mathbf{S}_{\hat{\boldsymbol{\Sigma}}^{l}}$ are as defined earlier.

Once the solution for the covariance matrix is obtained, the REML method updates the mean vector using equation (2.28) —this time substituting $\boldsymbol{\Sigma}_{REML}^{l+1}$ in place

of $\boldsymbol{\Sigma}$:

$$\hat{\boldsymbol{\mu}}_{REML} = \left(\sum_{i=1}^{k} \hat{\boldsymbol{\Sigma}}_{i*}^{-1}\right)^{-1} \left(\sum_{i=1}^{k} \hat{\boldsymbol{\Sigma}}_{i*}^{-1} \hat{\boldsymbol{\theta}}_i\right). \tag{2.28}$$

Where $\hat{\boldsymbol{\Sigma}}_{i*} = \hat{\boldsymbol{\Sigma}}_{REML}^{l+1} + \boldsymbol{S}_i$ and $\hat{\boldsymbol{\theta}}_i$ is as defined before.

In this thesis, however, we will utilize the quasi-Newton method which has been implemented in the multi-purpose optimization algorithm 'optim' of both the multivariate meta-analysis,'mvmeta' (Gasparrini, 2015) and meta-analysis of diagnostic accuracy, 'mada' (Doebler, 2015) R packages (R Core Team, 2015). The quasi-Newton method numerically approximates the Hessian matrix of the regular Newton-Raphson method (Schoenberg, 2001). Both of the R packages mentioned above use the Broyden-Fletcher-Goldfarb -Shanno (BFGS) method, a method regarded as best among other alternatives to the secant updates of quasi-Newton method (Schoenberg, 2001).

2.3.3 Bivariate RE Model Using Two Newly Proposed Transformations

In this thesis, we adopt the Reitsma et al. (2005) model given in (2.17) and propose two new transformations —'arcsine square root' and 'Freeman-Tukey double arcsine' transformation in place of the default 'logit' transformation employed in Reitsma et al. (2005). Other aspects stay the same including the model specification & parameter estimation procedure, and, the only difference comes when computing the within study variances given in (2.6).

For instance, if we are using the arcsine square root transformation of the sensitivities and specificities; that is, if

$$\theta_{1i} = \sin^{-1}(\sqrt{p_i}) \text{ and } \theta_{2i} = \sin^{-1}(\sqrt{q_i}), \ \ i = 1, 2, ...k, \tag{2.29}$$

then, the within study variances for sensitivity and specificity of study i (using the delta method) respectively is:

$$s_{1i}^2 = var\left(\sin^{-1}(\sqrt{p_i})\right) \approx var(p_i)\left(\frac{d}{dp_i}\sin^{-1}(\sqrt{p_i})\right)^2 = \frac{p(1-p)}{n_{1i}}\left(\frac{1}{2\sqrt{p}}\frac{1}{\sqrt{1-p}}\right)^2 = \frac{1}{4n_{1i}} \tag{2.30}$$

$$\text{and } \ s_{2i}^2 = var\left(\sin^{-1}(\sqrt{q_i})\right) \approx var(q_i)\left(\frac{d}{dq_i}\sin^{-1}(\sqrt{q_i})\right)^2 = \frac{q(1-q)}{n_{2i}}\left(\frac{1}{2\sqrt{q}}\frac{1}{\sqrt{1-q}}\right)^2 = \frac{1}{4n_{2i}}. \tag{2.31}$$

On the other hand, the Freeman-Tukey double arcsine transformation (Freeman and Tukey, 1950) of sensitivities & specificities and their corresponding within study variances, respectively, are given by;

$$\theta_{1i} = (1/2)\left[\sin^{-1}\left(\sqrt{\frac{u_i}{n_{1i}+1}}\right) + \sin^{-1}\left(\sqrt{\frac{u_i+1}{n_{1i}+1}}\right)\right], \tag{2.32}$$

$$\theta_{2i} = (1/2)\left[\sin^{-1}\left(\sqrt{\frac{v_i}{n_{2i}+1}}\right) + \sin^{-1}\left(\sqrt{\frac{v_i+1}{n_{2i}+1}}\right)\right]. \tag{2.33}$$

Once again using the delta method, approximate variances are given by:

$$s_{1i}^2 = var(\hat{\theta}_{1i}) = \frac{1}{4n_{1i}+2} \text{ and } s_{2i}^2 = var(\hat{\theta}_{2i}) = \frac{1}{4n_{2i}+2}, \text{ respectively.} \tag{2.34}$$

Another motivation behind the use of the variance stabilizing transformations is that, they do not require the ad hoc addition of the so-called continuity correction to the four cells of the 2×2 data table whenever there is a cell with zero counts. Moses et al. (1993) have discussed the downward bias that this correction has on the estimated ROC curves and also showed that it moves the ROC curve away from its ideal position in the top left corner on the ROC space. Secondly, the variance stabilizing transformations might also be favored for their asymmetric property, as Chu et al., (2010) explained that the asymmetric property of the clog-log transformation led to better goodness of fit than the logit and probit transformations when modeling the pairs of (TPR, TNR), (TPR, 1-TNR), (1-TPR, TNR) or (1-TPR, 1-TNR).

Chapter 3

Simulation Study

In this Chapter, we evaluate the methods we have proposed in Chapter 2 using simulations. Section 3.1 describes how we have designed the simulation study. In Section 3.2, we discuss the techniques we have used in order to evaluate the performance of the simulation experiment; and the results of the simulations are presented in Section 3.3.

3.1 Simulation Design

In this thesis, we have designed a four-step simulation based on the strategy used by Hamza et al. (2008) and Doebler et al. (2012). First, the true logit, arcsine square root (ASR) and Freeman-Tukey double arcsine (FTDA) transformed sensitivities and false positive rates $(\theta_{1i}, \theta_{2i})$ for each study were sampled from a bivariate normal distribution using specified true values of $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, & ρ and then back transformed to their respective unit-interval values (Hamza et al., 2008, Doebler et al., 2012).

Second, the sample sizes for each study arm (with disease and without disease) were generated independently from a Poisson distribution with mean n, as in Doebler et al. (2012). Third, we simulated the number of positive test results in each study arm (TPs and FPs) from the binomial distribution using the two parameters generated in the first two steps above (i.e. using the back-transformed true sensitivities and specificities as probabilities and the sample sizes (n_{1i}, n_{2i}) generated from a Poisson distribution). Finally, the bivariate RE model of Reitsma et al. (2005) was fitted for each of the three transformations —logit, ASR and FTDA transformation. Although we have used both the ML and REML methods to obtain estimates of parameters of the model, we opted to present the result only for the REML case, since we did not observe substantial differences between the two methods of estimation in the results of the two parameters of interest.

Three different packages freely available in the R programming language that we employed during data simulation and analysis are 'metafor', 'mada' and 'mvmeta'. The 'metafor' package (Viechtbauer, 2010) was used to transform and back-transform the simulated sensitivities and specificities within each of the three transformations. The 'mada' package (Doebler, 2015) was used to fit the bivariate RE model using logit transformation. The model fit using the transformations we proposed —ASR and FTDA was done with the 'mvmeta' package (Gasparrini et al., 2015). The simulation of the bivariate data from a normal distribution was performed using the 'mvrnorm' function of the 'MASS' package (Venables and Ripley, 2002). The simulation is replicated 1000 times and different seed has been set for each of the 1000 simulation replications for the sake of parallelizing and reproducibility.

We have summarized the simulation scenarios considered in this thesis in Table 3.1 below, and, a total of 360 scenarios for each of the three transformations have been examined as shown in the table.

Table 3.1: Summary of Parameters Varied in the Simulation Study.

Parameters	Description	Values
(μ_1^*, μ_2^*)	true back-transformed sensitivity and 1-specificity pairs	(95%, 10%) (95%, 30%) (62%, 30%) (80%, 20%)
(σ_1^2, σ_2^2)	true between-study variance of sensitivity and false positive rate pairs	(0.5, 0.5) (1.2, 1.2) (0.5, 1.2)
ρ	true between-study correlation coefficient	0.2 0.5
n	true mean sample size for each study arm	40 100 500
k	number of studies	5 10 25 50 100

Note: The (μ_1^*, μ_2^*) pairs are the back-transformed values of (μ_1, μ_2) which corresponds to different values in the transformed scale for the three transformations—logit ASR and FTDA.

3.2 Performance Evaluation

Once the desired simulated data is obtained and the model is fitted, one must evaluate results. Burton et al. (2006) have discussed evaluation approaches in medical

statistics and based on their recommendations, we have computed measure of accuracy (absolute bias), precision (root mean square error) and coverage (coverage probability). Although we computed these performance evaluation measures for all of the estimated parameters, we present the results only for the two parameters of interest, sensitivity & 1-specificity. We have used the term 'absolute bias' to differentiate between the one we calculated here and other types of bias including relative, percentage, and standardized bias discussed by Burton et al. (2006), and, not to mean we are taking the non-negative value of the bias.

$$\text{Absolute Bias} = E(\hat{\mu}_i) - \mu_i = \bar{\hat{\mu}}_i - \mu_i, i = 1, 2. \tag{3.1}$$

$$\begin{aligned} \text{Root mean square error (RMSE)} = \sqrt{MSE} &= \sqrt{E(\hat{\mu}_i - \mu_i)^2} \\ &= \sqrt{E(\hat{\mu}_i - \bar{\hat{\mu}}_i)^2 + (E(\hat{\mu}_i) - \mu_i)^2} \\ &= \sqrt{\text{var}(\hat{\mu}_i) + \text{Bias}(\hat{\mu}_i)^2}, i = 1, 2. \end{aligned} \tag{3.2}$$

For each of the 1000 simulation replications, the proportion of times the 95% confidence interval: $\hat{\mu}_i \mp 1.96 \times se(\hat{\mu}_i)$, i=1,2 include the true value μ_i is called the 95% coverage probability.

Besides the scalar absolute bias and MSE computed for the estimators $\hat{\mu}_1$ and $\hat{\mu}_2$, we have also included bias and MSE for the vector valued estimator $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \hat{\mu}_2)'$ computed as:

$$Bias = ||Bias(\hat{\boldsymbol{\mu}})|| = \sqrt{(\hat{\mu}_1 - \mu_1)^2 + (\hat{\mu}_2 - \mu_2)^2}. \tag{3.3}$$

$$\begin{aligned} MSE = E(||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2) &= \text{trace}(var(\hat{\boldsymbol{\mu}})) + ||\text{Bias}(\hat{\boldsymbol{\mu}})||^2 \\ &= \sum_{i=1}^{2} var(\hat{\mu}_i) + \sum_{i=1}^{2} (\hat{\mu}_i - \mu_i)^2. \end{aligned} \tag{3.4}$$

3.3 Simulation Results

In this section we present the simulation study results for each of the specific scenarios considered in this thesis. We will use tables to present results in terms of absolute bias, RMSE, and coverage probability for some selected scenarios. Figures have been also employed to compare the proposed methods with the existing one in terms of the vector valued bias & MSE, in addition to the three scalar performance measures mentioned above.

Although six different combinations of the true between-study variances and correlations are considered, we did not observe any significant effect that they have on the results of the simulation performance measures of the parameters of interest, which is consistent with the findings of Hamza et al. (2008). Therefore, we will present one table and figure just for one combination of the true between-study variance and correlation throughout the presentation of results.

3.3.1 Results in terms of absolute bias

Table 3.2 displays the results of the absolute bias for all combinations of true parameters described in Table 3.1. The absolute bias for all the three methods generally decreases as the sample size increases and that it always underestimates the TPR and overestimates the FPR, which is consistent with previous simulation studies (Doebler et al., 2012; Kuss et al., 2013) when they used the logit transformation. But, the absolute bias is not decreasing as the number of study increases, rather, it tends to increase with the number of studies in most of the scenarios which is, again,

in agreement with the study of Doebler et al. (2012) and Eusebi et al. (2014).

Table 3.2: Absolute bias for sensitivity and false positive rate when $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.5$, and $\rho = 0.2$.

True μ_1^* & μ_2^*	Transformation	k = 5		k = 10		k = 25		k = 50		k = 100	
		TPR	*FPR*	*TPR*	*FPR*	*TPR*	*FPR*	*TPR*	*FPR*	*TPR*	*FPR*
n = 40											
95%, 10%	logit	-0.33	0.21	-0.44	0.26	-0.49	0.28	-0.50	0.29	-0.50	0.29
	ASR	-0.22	0.17	-0.22	0.17	-0.22	0.18	-0.22	0.18	-0.22	0.17
	FTDA	-0.20	0.16	-0.20	0.16	-0.20	0.17	-0.20	0.16	-0.20	0.17
95%, 30%	logit	-0.31	0.05	-0.43	0.06	-0.49	0.06	-0.50	0.07	-0.50	0.07
	ASR	-0.22	0.07	-0.22	0.07	-0.22	0.07	-0.22	0.07	-0.22	0.07
	FTDA	-0.20	0.07	-0.20	0.07	-0.20	0.07	-0.20	0.07	-0.20	0.07
62%, 30%	logit	-0.01	0.03	-0.01	0.04	-0.02	0.04	-0.03	0.06	-0.03	0.06
	ASR	-0.05	0.07	-0.05	0.07	-0.04	0.07	-0.04	0.07	-0.04	0.07
	FTDA	-0.05	0.07	-0.04	0.07	-0.04	0.07	-0.04	0.07	-0.04	0.07
80%, 20%	logit	-0.05	0.07	-0.08	0.09	-0.11	0.10	-0.12	0.12	-0.13	0.13
	ASR	-0.12	0.11	-0.12	0.12	-0.11	0.12	-0.12	0.12	-0.12	0.12
	FTDA	-0.12	0.11	-0.11	0.11	-0.11	0.11	-0.11	0.11	-0.11	0.11
n = 100											
95%, 10%	logit	-0.13	0.07	-0.18	0.08	-0.23	0.11	-0.25	0.12	-0.25	0.13
	ASR	-0.21	0.16	-0.21	0.16	-0.20	0.16	-0.20	0.16	-0.20	0.16
	FTDA	-0.18	0.14	-0.18	0.14	-0.18	0.15	-0.18	0.15	-0.18	0.15
95%, 30%	logit	-0.12	0.02	-0.18	0.02	-0.22	0.02	-0.24	0.02	-0.25	0.03
	ASR	-0.21	0.06	-0.21	0.06	-0.20	0.07	-0.20	0.07	-0.20	0.06
	FTDA	-0.18	0.06	-0.18	0.06	-0.18	0.06	-0.18	0.06	-0.18	0.06
62%, 30%	logit	0.00	0.02	-0.00	0.01	-0.01	0.01	-0.01	0.01	-0.01	0.01
	ASR	-0.05	0.06	-0.05	0.06	-0.04	0.07	-0.04	0.07	-0.04	0.06
	FTDA	–0.04	0.06	-0.04	0.06	-0.04	0.06	-0.04	0.06	-0.04	0.06
80%, 20%	logit	-0.01	0.02	-0.02	0.02	-0.03	0.02	-0.03	0.03	-0.04	0.04
	ASR	-0.11	0.10	-0.11	0.10	-0.10	0.11	-0.11	0.11	-0.11	0.11
	FTDA	-0.11	0.10	-0.10	0.10	-0.10	0.10	-0.10	0.10	-0.10	0.10
n = 500											
95%, 10%	logit	-0.01	0.01	-0.02	0.01	-0.03	0.01	-0.03	0.01	-0.03	0.01
	ASR	-0.20	0.15	-0.20	0.15	-0.19	0.15	-0.19	0.15	-0.19	0.15
	FTDA	-0.16	0.13	-0.16	0.13	-0.16	0.14	-0.16	0.13	-0.16	0.14
95%, 30%	logit	-0.01	0.00	-0.02	0.00	-0.03	-0.00	-0.03	0.00	-0.03	0.00
	ASR	-0.20	0.06	-0.20	0.06	-0.19	0.06	-0.19	0.06	-0.19	0.06
	FTDA	-0.16	0.05	-0.16	0.05	-0.16	0.06	-0.16	0.05	-0.16	0.06
62%, 30%	logit	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00	0.00
	ASR	-0.04	0.06	-0.04	0.06	-0.03	0.06	-0.04	0.06	-0.04	0.06
	FTDA	-0.04	0.05	-0.04	0.05	-0.04	0.06	-0.03	0.05	-0.03	0.06
80%, 20%	logit	-0.00	0.01	-0.00	0.00	-0.00	0.00	-0.01	0.00	-0.01	0.01
	ASR	-0.11	0.10	-0.11	0.10	-0.10	0.10	-0.10	0.10	-0.10	0.10
	FTDA	-0.10	0.08	-0.09	0.08	-0.09	0.09	-0.09	0.08	-0.09	0.09

Note: The (μ_1^*, μ_2^*) pairs are the back-transformed values of (μ_1, μ_2) which corresponds to different values in the transformed scale for the three transformations—logit ASR and FTDA.

In terms of bias, both of our proposed methods of transformation performed better than the standard logit transformation for all of the number of studies when the sample size for each study arm is small (n = 40) and the true (TPR, FPR) pair is very high (95%, 10%). The new methods also outperformed their competitor logit transformation for sensitivity when the true (TPR, FPR) pair is (95%, 30%) for all of the five study numbers and small sample size. For the same scenario, however, all the methods have very similar performance for 1-specificity even though the logit

transformation has slight advantage.

For moderate sample size (n = 100), the proposed methods have dominantly over-performed the standard logit transformation when the true pairs of sensitivity and false positive rate are (95%, 10%) and (95%, 30%) for all moderate and large number of studies (k = 25, 50, 100). In contrast, for the same scenarios explained above, but when the true pairs of sensitivity and false positive rate are (62%, 30%) and (80%, 20%), the logit transformation performed better than the proposed methods.

Although the absolute bias for all the three transformations has asymptotically decreased, the standard logit transformation has outperformed the others in terms of absolute bias for all pairs of sensitivity & false positive rate and all number of studies when sample size is large (n = 500). A better performance for the standard logit transformation has been observed for false positive rate in all scenarios except when the true pair of sensitivity & 1-specificity is (95%, 10%) and the sample size is small.

Below, we present a panel of graphs representing the absolute bias for sensitivity and 1-specificity separately. In terms of pattern, Figure 3.1 reveals that for the logit transformation, the bias is generally decreasing in magnitude for both sensitivity and 1-specificity as the sample size increases, it increases for small number of studies (k = 5, 10) and keeps moving on a constant path for the rest of the number of studies which is consistent with the simulation studies of Eusebi et al. (2014) and Doebler et al. (2012).

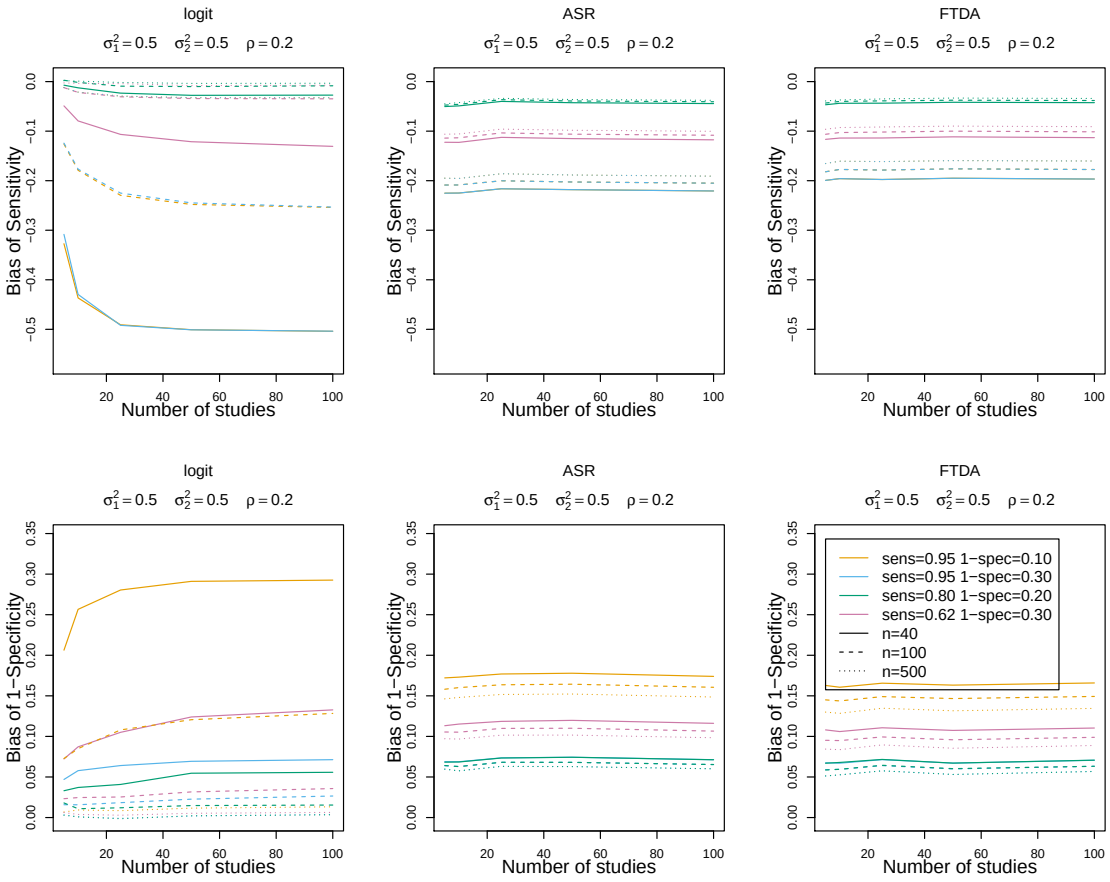


Figure 3.1: Absolute bias for sensitivity (top panel) and 1-specificity (bottom panel) for logit, ASR and FTDA when $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.5$ and $\rho = 0.2$.

For both of the proposed transformations, the bias decreases with increasing sample size, slightly decreases for small number of studies and is constant for the remaining values of k. Another interesting pattern uniformly observed for three

of the transformations is that, even though the rate varies, for approximately the same specificity, it seems the bias for sensitivity increases monotonically with true sensitivity.

Figure 3.1 illustrates that the proposed transformations performed better in sensitivity than the standard logit when; the sample size is small and the true pair of sensitivity & 1-specificity is (95%, 10%) for all number of studies, the sample size is small and the sensitivity & 1-specificity pair is (95%, 30%) for moderate and large number of studies. Another situation where the proposed methods performed better than the logit for sensitivity is when the sample size is small and the true sensitivity & 1-specificity pair is (62%, 30%) for moderate and large number of studies. For small number of studies, the logit transformation performs better for the same scenario. The only time the logit transformation has fully dominated the proposed methods in terms of sensitivity for all sample size and number of study is when the true sensitivity & 1-specificity pair is (80%, 20%).

In terms of 1-specificity, the proposed methods performed better than the logit transformation when the sample size is small, the true pair of sensitivity & 1-specificity is (95%, 10%) for all number of studies. Additionally, when the true sensitivity & 1-specificity pair is (62%, 30%), the sample size is small and the number of studies is large, the proposed methods have a slight advantage over their competitor the logit transformation. Otherwise, for most of the simulation scenarios, the logit transformation has dominated the proposed methods in terms of bias for 1-specificity.

In general, in terms of absolute bias, the proposed methods have performed better

than the logit transformation for small and moderate sample sizes when the true sensitivity is very high. For the same scenario, and when the sensitivity is lower, the logit transformation is preferable. Additionally, the logit transformation is better than the proposed methods for large sample sizes in spite of the simulation scenario.

3.3.2 Results in terms of RMSE

We have tabulated the RMSE of the estimators from the three transformations in Table 3.3 below. Except for some selected scenarios, in most of the cases considered in this simulation study, the two newly proposed transformations —the ASR and the FTDA transformation outperformed the standard logit transformation for both parameters of interest. The logit transformation performed better for all the sensitivity & 1-specificity pairs only when the sample size is large and number of studies vary from moderate to large. The logit transformation also has slight advantage in terms of RMSE when the pair of sensitivity & 1-specificity is (80%, 20%), the sample size is moderate and the number of studies is large.

In the rest of the scenarios, the proposed methods have outperformed the standard one for all the different combinations of σ_1^2, σ_2^2. and ρ. Particularly, huge performance dominance of RMSE for the proposed methods has been observed when the true sensitivity and false positive rate pairs are (95%, 10%) and (95%, 30%) for all the sample sizes and number of studies. This out-performance of the new methods in terms of MSE has been also the case for the rest of the variance-covariance parameter combinations as it can be clearly seen from the figures presented below.

Table 3.3: RMSE for sensitivity and false positive rate when the true $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.5$, and $\rho = 0.2$.

True μ_1^* & μ_2^*	Transformation	k = 5		k = 10		k = 25		k = 50		k = 100	
		TPR	*FPR*	*TPR*	*FPR*	*TPR*	*FPR*	*TPR*	*FPR*	*TPR*	*FPR*
n = 40											
95%, 10%	logit	0.49	0.40	0.50	0.34	0.51	0.32	0.51	0.31	0.51	0.30
	ASR	0.30	0.26	0.26	0.22	0.23	0.20	0.23	0.19	0.22	0.18
	FTDA	0.27	0.25	0.24	0.21	0.21	0.19	0.20	0.17	0.20	0.17
95%, 30%	logit	0.48	0.35	0.49	0.24	0.51	0.16	0.51	0.12	0.51	0.10
	ASR	0.30	0.23	0.26	0.16	0.23	0.12	0.23	0.10	0.22	0.08
	FTDA	0.27	0.22	0.24	0.16	0.21	0.12	0.20	0.09	0.20	0.08
62%, 30%	logit	0.35	0.35	0.24	0.25	0.15	0.15	0.11	0.11	0.08	0.09
	ASR	0.23	0.23	0.17	0.16	0.11	0.12	0.08	0.10	0.07	0.08
	FTDA	0.22	0.22	0.16	0.16	0.10	0.12	0.08	0.09	0.06	0.08
80%, 20%	logit	0.35	0.36	0.25	0.25	0.18	0.18	0.16	0.16	0.15	0.15
	ASR	0.25	0.24	0.20	0.18	0.15	0.15	0.13	0.14	0.13	0.12
	FTDA	0.23	0.23	0.19	0.18	0.15	0.14	0.13	0.12	0.12	0.12
n = 100											
95%, 10%	logit	0.37	0.34	0.29	0.24	0.27	0.18	0.27	0.15	0.26	0.14
	ASR	0.29	0.26	0.25	0.21	0.22	0.19	0.21	0.18	0.21	0.17
	FTDA	0.27	0.25	0.23	0.20	0.20	0.17	0.19	0.16	0.18	0.16
95%, 30%	logit	0.37	0.33	0.29	0.23	0.27	0.15	0.26	0.10	0.26	0.07
	ASR	0.29	0.23	0.25	0.16	0.22	0.12	0.21	0.10	0.21	0.08
	FTDA	0.27	0.23	0.23	0.16	0.20	0.12	0.19	0.09	0.18	0.08
62%, 30%	logit	0.33	0.34	0.23	0.24	0.15	0.14	0.10	0.10	0.07	0.07
	ASR	0.24	0.23	0.18	0.16	0.11	0.12	0.08	0.10	0.07	0.08
	FTDA	0.23	0.23	0.16	0.16	0.11	0.12	0.08	0.09	0.06	0.08
80%, 20%	logit	0.33	0.34	0.23	0.23	0.15	0.15	0.10	0.10	0.08	0.08
	ASR	0.25	0.24	0.20	0.18	0.14	0.15	0.13	0.13	0.12	0.12
	FTDA	0.24	0.23	0.18	0.17	0.14	0.14	0.12	0.12	0.11	0.11
n = 500											
95%, 10%	logit	0.32	0.33	0.23	0.23	0.14	0.14	0.10	0.10	0.08	0.07
	ASR	0.28	0.26	0.25	0.21	0.21	0.18	0.20	0.17	0.20	0.16
	FTDA	0.27	0.25	0.22	0.20	0.19	0.16	0.17	0.15	0.17	0.14
95%, 30%	logit	0.32	0.33	0.23	0.23	0.14	0.14	0.10	0.10	0.08	0.07
	ASR	0.28	0.24	0.25	0.17	0.21	0.12	0.20	0.10	0.20	0.08
	FTDA	0.27	0.23	0.22	0.17	0.19	0.12	0.17	0.09	0.17	0.08
62%, 30%	logit	0.31	0.33	0.23	0.23	0.14	0.14	0.10	0.10	0.07	0.07
	ASR	0.24	0.24	0.18	0.17	0.11	0.12	0.08	0.10	0.06	0.08
	FTDA	0.23	0.23	0.17	0.17	0.11	0.12	0.08	0.09	0.06	0.08
80%, 20%	logit	0.32	0.33	0.23	0.23	0.14	0.14	0.10	0.10	0.07	0.07
	ASR	0.25	0.24	0.20	0.18	0.14	0.14	0.12	0.12	0.11	0.11
	FTDA	0.24	0.24	0.18	0.17	0.14	0.14	0.11	0.11	0.10	0.10

Note: The (μ_1^*, μ_2^*) pairs are the back-transformed values of (μ_1, μ_2) which corresponds to different values in the transformed scale for the three transformations—logit ASR and FTDA.

Figure 3.2 reveals that the pattern of RMSE is decreasing both in terms of the sample size and number of studies. The fact that the RMSE decrease as the number of studies increase reflects that the estimates of the transformed sensitivity and 1-specificity become more precise as the number of studies gets larger as expected. However, for the logit transformation, this property does not hold for sensitivity when the sample size is small for large true sensitivity and specificity values. We suspect the significant increment of the bias with the number of studies for sensitivity may contribute to this result for the specified scenarios.

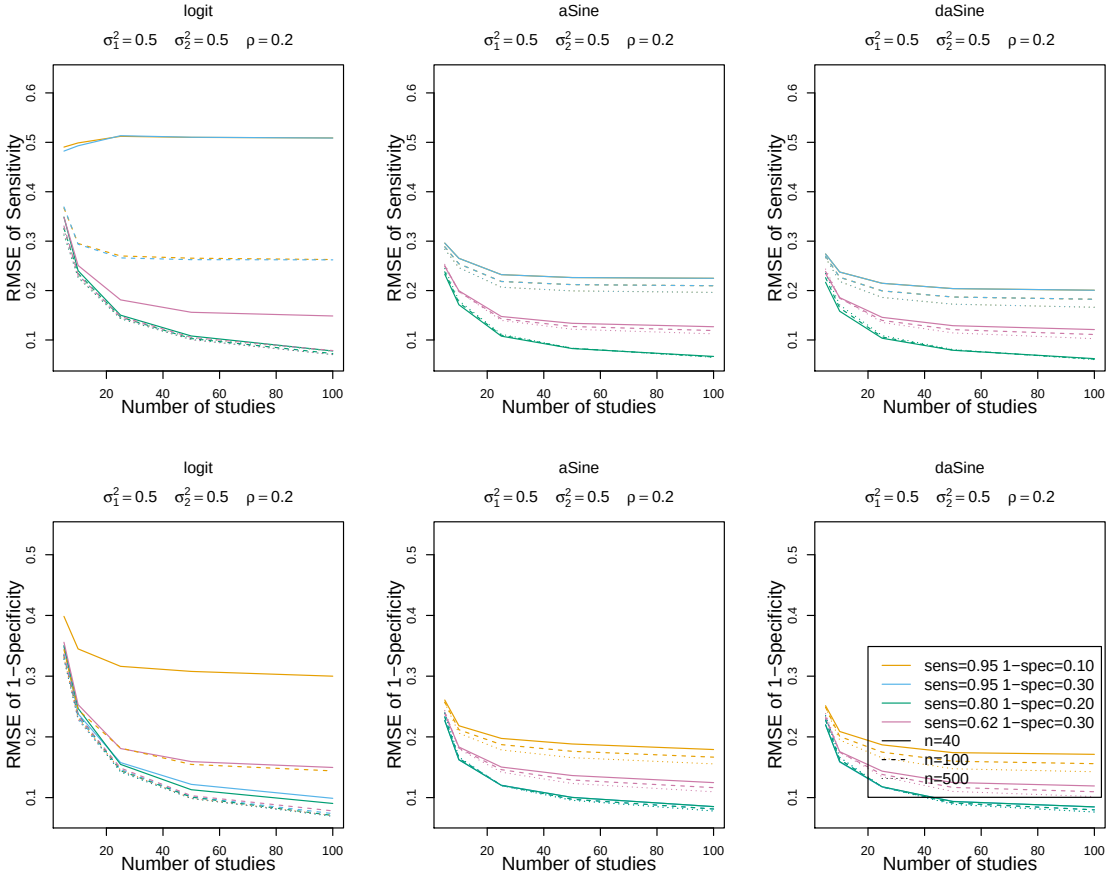


Figure 3.2: Root mean square error for sensitivity (top panel) and 1-specificity (bottom panel) for logit, ASR and FTDA when $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.5$, and $\rho = 0.2$.

As the sample size gets larger, the RMSE for the standard logit transformation is rapidly decreasing although it does not get better than the proposed methods. It can also be observed from Figure 3.2 that the proposed methods best performance

in RMSE notably is when the sample size is small and the sensitivity & 1-specificity pairs are (95%, 10%) & (95%, 30%) for all number of studies considered in this simulation study. The RMSE is in favor of the standard logit when both the sample size and number of studies are large for all pairs of true sensitivity & 1-specificity as it was also observed in Table 3.3. For the other scenarios, it is the proposed methods that dominate the standard logit in terms of RMSE as a measure of performance.

In sum, the proposed methods of transformations are preferable in terms of RMSE for most of the scenarios considered in this thesis except for large sample size and number of studies, in which the logit transformation performed better.

3.3.3 Results in terms of coverage probability

Regarding coverage probability, Table 3.4 presents the results for a selected scenario when $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.5$ and $\rho = 0.2$. An increasing trend of coverage probability for all of the scenarios in terms of sample size, and for selected scenarios in terms of number of studies has been observed, which agrees with the study by Hamza et al. (2008) when they used mean log diagnostic odds ratio as parameter of interest. For instance, the coverage probability has increased with the number of studies for both parameters and logit transformation when the sensitivity and false positive rate pairs are (62%, 30%) & (80%, 20%) and sample size varies from moderate to large.

For small sample size and true pairs of sensitivity & false positive rate of (95%, 10%) & (95%, 30%), the proposed methods have better performance for both parameters of interest in terms of coverage when the number of study varies from small to moderate. However, the logit transformation has outperformed the other two

Table 3.4: The 95% Coverage probability for sensitivity and false positive rate when the true $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.5$, and $\rho = 0.2$.

True μ_1^* & μ_2^*	Transformation	k = 5		k = 10		k = 25		k = 50		k = 100	
		TPR	*FPR*	*TPR*	*FPR*	*TPR*	*FPR*	*TPR*	*FPR*	*TPR*	*FPR*
n = 40											
95%, 10%	logit	0.81	0.85	0.57	0.77	0.12	0.50	0.00	0.22	0.00	0.02
	ASR	0.80	0.84	0.66	0.83	0.21	0.50	0.01	0.15	0.00	0.01
	FTDA	0.84	0.85	0.74	0.84	0.35	0.54	0.06	0.21	0.00	0.02
95%, 30%	logit	0.83	0.89	0.58	0.92	0.12	0.92	0.00	0.89	0.00	0.85
	ASR	0.80	0.87	0.66	0.92	0.21	0.89	0.01	0.83	0.00	0.70
	FTDA	0.84	0.87	0.74	0.91	0.35	0.89	0.06	0.83	0.00	0.69
62%, 30%	logit	0.89	0.88	0.92	0.92	0.94	0.94	0.94	0.93	0.94	0.89
	ASR	0.86	0.86	0.90	0.92	0.92	0.89	0.90	0.83	0.87	0.70
	FTDA	0.88	0.88	0.90	0.91	0.92	0.89	0.91	0.84	0.87	0.68
80%, 20%	logit	0.90	0.88	0.90	0.90	0.88	0.87	0.78	0.76	0.57	0.56
	ASR	0.85	0.86	0.87	0.90	0.79	0.79	0.59	0.59	0.28	0.29
	FTDA	0.87	0.87	0.87	0.89	0.80	0.79	0.60	0.64	0.29	0.32
n = 100											
95%, 10%	logit	0.87	0.88	0.82	0.90	0.64	0.85	0.31	0.76	0.06	0.55
	ASR	0.81	0.85	0.73	0.86	0.34	0.62	0.06	0.27	0.00	0.04
	FTDA	0.85	0.86	0.80	0.86	0.50	0.67	0.18	0.37	0.01	0.08
95%, 30%	logit	0.87	0.88	0.82	0.92	0.65	0.93	0.32	0.95	0.07	0.95
	ASR	0.81	0.86	0.73	0.92	0.34	0.90	0.06	0.86	0.00	0.75
	FTDA	0.85	0.87	0.80	0.91	0.50	0.90	0.18	0.88	0.01	0.77
62%, 30%	logit	0.88	0.89	0.92	0.91	0.94	0.94	0.95	0.95	0.95	0.95
	ASR	0.86	0.86	0.90	0.92	0.93	0.90	0.92	0.86	0.88	0.76
	FTDA	0.88	0.87	0.91	0.91	0.93	0.90	0.92	0.87	0.90	0.77
80%, 20%	logit	0.88	0.89	0.92	0.92	0.94	0.94	0.94	0.95	0.93	0.93
	ASR	0.85	0.86	0.88	0.90	0.82	0.82	0.67	0.68	0.39	0.40
	FTDA	0.87	0.88	0.90	0.90	0.83	0.84	0.71	0.72	0.44	0.46
n = 500											
95%, 10%	logit	0.88	0.88	0.92	0.91	0.94	0.93	0.94	0.95	0.94	0.95
	ASR	0.82	0.85	0.77	0.88	0.48	0.68	0.13	0.39	0.00	0.09
	FTDA	0.86	0.87	0.83	0.88	0.64	0.74	0.32	0.52	0.04	0.17
95%, 30%	logit	0.88	0.88	0.92	0.92	0.94	0.93	0.94	0.95	0.94	0.96
	ASR	0.82	0.86	0.77	0.93	0.48	0.90	0.13	0.88	0.00	0.80
	FTDA	0.86	0.87	0.83	0.92	0.64	0.92	0.32	0.90	0.04	0.81
62%, 30%	logit	0.89	0.87	0.91	0.92	0.94	0.93	0.94	0.95	0.95	0.95
	ASR	0.85	0.86	0.90	0.92	0.93	0.90	0.92	0.88	0.89	0.80
	FTDA	0.88	0.88	0.91	0.92	0.93	0.91	0.93	0.90	0.92	0.81
80%, 20%	logit	0.88	0.88	0.91	0.92	0.94	0.93	0.95	0.95	0.95	0.95
	ASR	0.85	0.86	0.88	0.91	0.84	0.85	0.73	0.73	0.48	0.50
	FTDA	0.89	0.88	0.90	0.91	0.86	0.87	0.76	0.78	0.57	0.58

Note: The (μ_1^*, μ_2^*) pairs are the back-transformed values of (μ_1, μ_2) which corresponds to different values in the transformed scale for the three transformations—logit ASR and FTDA.

methods for the same scenario when the sample size is moderate & large.

On the other hand, for large number of studies, it was the logit transformation that has the best performance for all the sample sizes and pairs of sensitivity & 1-specificity; although in sensitivity, all three methods performed poorly when the number of study is very large and the sample size is small or moderate. This result is expected, since for these scenarios, the bias is large (Table 3.2) and the standard error is small (Table 3.3), hence resulting in a narrower interval which does not include the true values. For moderate and large sample sizes, all the three methods performed

almost similarly although the logit transformation is favored for all the four pairs of sensitivity & 1-specificity and all number of studies.

We observed that even for large sample size, large number of studies, high true sensitivity and specificity values, the coverage probability goes down to zero for both of the proposed transformations. One explanation is that the magnitude of the assumed between-study variances that we considered might be unrealistic for the arcsine-based transformations. Since the three transformations have different scales, the assumption of the same between-study variance is unrealistic and a variance proportional to the underlying scale of each transformation would be needed. In practice, the choice of an appropriate between-study variance for different transformation is challenging. One possible approach is to analyze a large number of empirical diagnostic study data and estimate a range of plausible variances that can be used to inform the simulations. As an illustration, we estimated the between-study variances of logit, ASR and FTDA transformed sensitivity and 1-specificity using the three empirical motivating data sets introduced in Chapter 1 and observed that the magnitude of the between-study variances for ASR and FTDA is approximately $(1/10)^{th}, (1/20)^{th}, (1/30)^{th}$ of that of the estimates for the logit transformed sensitivity and 1-specificity. To provide one concrete example, we simulated new data for the ASR transformation corresponding to $n = 500$, $k = 100$, $\mu_1^* = 95\%$, $\mu_2^* = 90\%$, $\rho = 0.2$ by changing the between-study variance from 0.5 to 0.05 ($(1/10)^{th}$), 0.025 ($(1/20)^{th}$) & 0.017 ($(1/30)^{th}$) and observed a dramatically improved coverage probability for sensitivity of 91%, 92% and 94%, respectively, unlike the zero coverage reported in the above table.

From Figure 3.3, one can also observe that the increasing trend of the coverage probability with sample size for all the four pairs of sensitivity & 1-specificity, and, its decreasing pattern as the number of study increases for both sensitivity and 1-specificity pairs except the (80%, 20%) pair. All three methods also performed uniformly poorly when the true pairs of sensitivity & 1-specificity is (95%, 10%) & (95%, 30%) for small & moderate sample sizes as the number of studies increase.

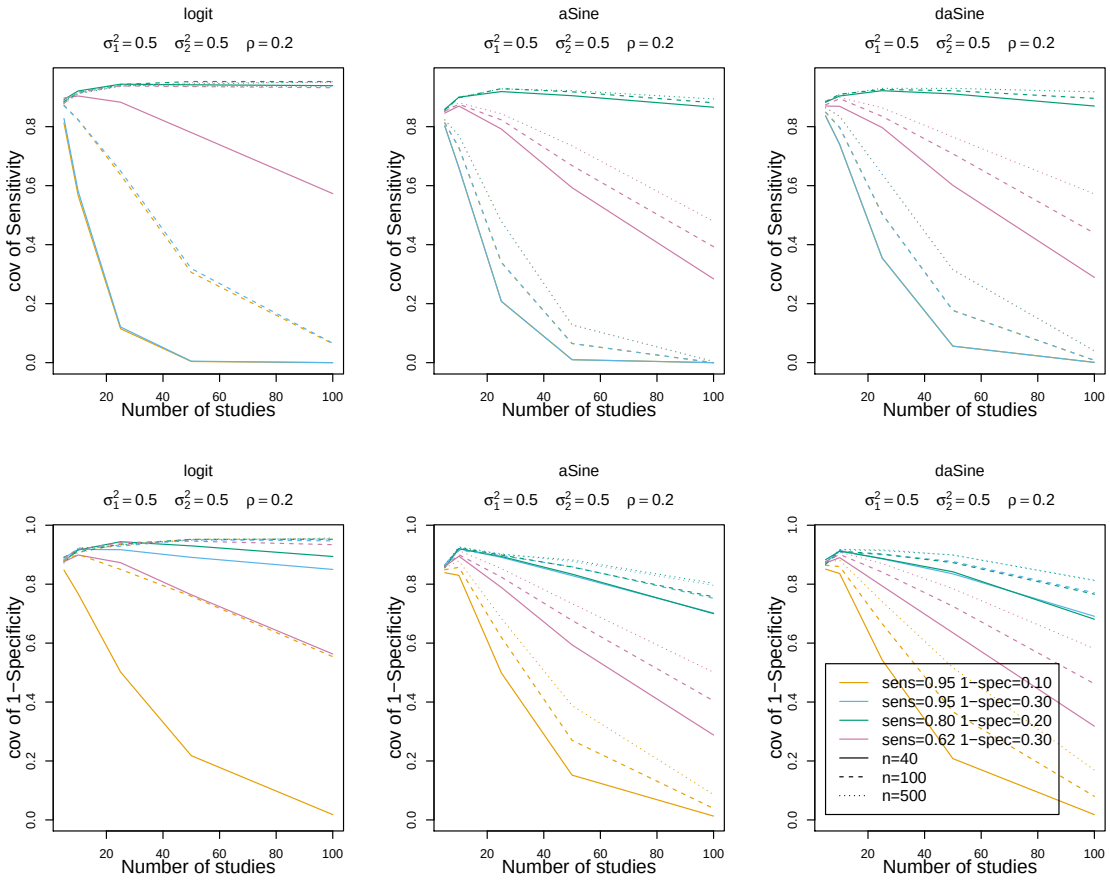


Figure 3.3: 95% coverage probability of sensitivity (top panel) and 1-specificity (bottom panel) for logit, ASR and FTDA when $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.5$ and $\rho = 0.2$.

Among the four pairs of sensitivity & 1-specificity considered in this simulation study, all the methods consistently performed well in terms of coverage probability

only for the (80%, 20%) case and also the coverage probability increases with increasing number of studies, particularly for sensitivity. For this scenario, although all methods performed similarly, the logit transformation has slight but consistent better performance for both parameters of interest.

We have also observed that the only exception where the proposed methods have outperformed the standard one is when the true sensitivity & 1-specificity pairs are (95%, 10%) & (95%, 30%), the sample size is small and the number of study varies from small to moderate. For other scenarios, the standard logit transformation has better performance over the proposed ones in terms of coverage probability.

The surprising coverage probabilities of less than 10% when the true sensitivity & specificity is very high and the number of studies is large is consistent with the findings of Hamza et al. (2008) and Kuss et al. (2013), although the parameter of interest used in the Hamza et al. (2008) was mean log diagnostic odds ratio. We suspect that the relatively smaller standard error estimates for these scenarios is the possible reason for the confidence intervals missing to include the true values.

Generally, for small sample sizes and high values of sensitivity & specificity, the proposed methods have better coverage than the standard method when the number of studies vary from small to moderate. For moderate and large sample sizes, regardless of the sensitivity & specificity pairs, though the logit transformation has slight advantage, all methods have similar coverage in both parameters of interest when the number of studies vary from small to moderate. However, for large number of studies, the logit transformation outperforms the proposed methods in spite of the scenario combinations.

3.3.4 Results in terms of vector valued bias and MSE

In this Section we present the results for absolute bias and MSE computed for the combined estimators of mean sensitivity and 1-specificity as given in (3.3) and (3.4). We have chosen to employ a panel of scatter plots to visualize the pairwise performance of the three methods at a time.

Figure 3.4 presents the scatter plot showing the pairwise comparison of the methods using vector-valued absolute bias and MSE as a performance measure. It can be observed that the logit transformation is better than the proposed methods in absolute bias mostly when both the number of studies and sample size is large for all pairs of sensitivity & 1-specificity except the (62%, 30%) pair. Conversely, as both the sample size and number of studies get smaller, the proposed methods perform better in all of the four sensitivity and 1-specificity pairs considered in the simulation study.

When the vector-valued MSE is used as a measure of performance, Figure 3.4 shows that, although the methods perform similarly for large number of studies and sample size, the logit transformation still performed better than the proposed methods marginally for all pairs of sensitivity & 1-specificity except the (62%, 30%) pair. On the other hand, the proposed methods performed better in terms of MSE when the sample size gets smaller in spite of the size of number of studies and the pairs of the sensitivity & 1-specificity. The results presented here are also in agreement with the RMSE results given under Table 3.3 for sensitivity and 1-specificity individually. When the sensitivity & 1-specificity pair is (62%, 30%), all methods performed similarly both in absolute bias and MSE for large number of studies and sample

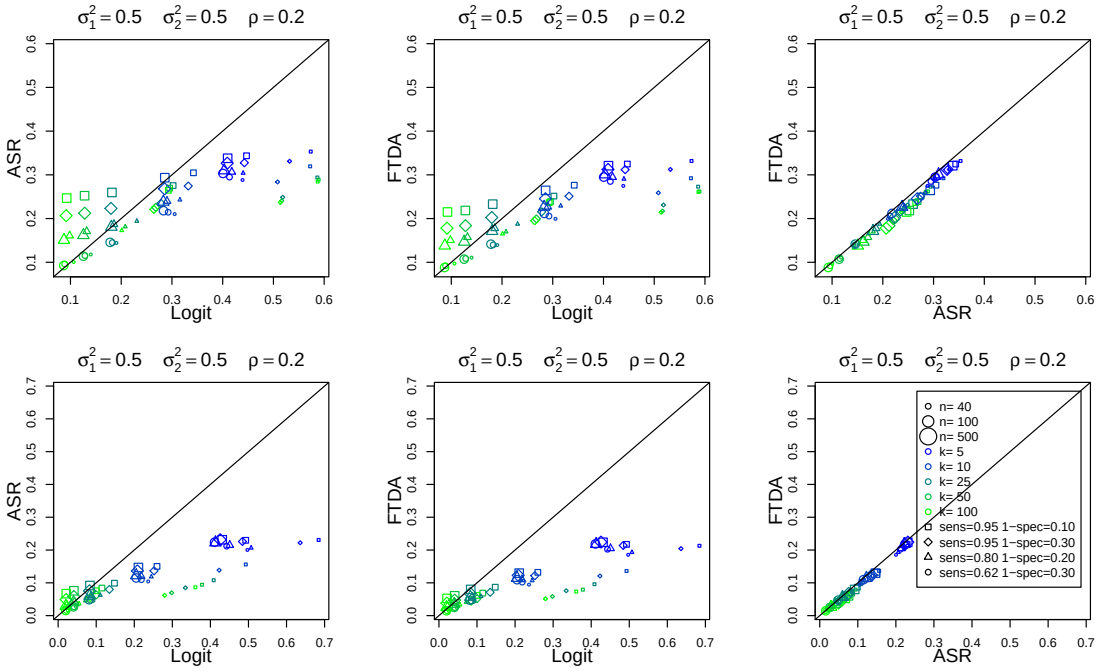


Figure 3.4: Scatter plot of the vector-valued absolute bias (top panel) and MSE (bottom panel) when $\sigma_1^2 = 0.5 = \sigma_2^2$ and $\rho = 0.2$.

size. Finally, we have observed that there is no substantial difference of performance between the proposed methods both in terms of bias and MSE, although the double arcsine transformation of Freeman-Tukey has a marginal advantage.

Chapter 4

Real Data Analysis

In this Chapter, we use three real data sets to illustrate the methods discussed in this thesis. The data sets are chosen so that they are representative of the considered methods/transformations in the simulation study. We compute and present estimates of sensitivity and 1-specificity with their 95% confidence interval using REML method of estimation.

4.1 The 'Children US' Data

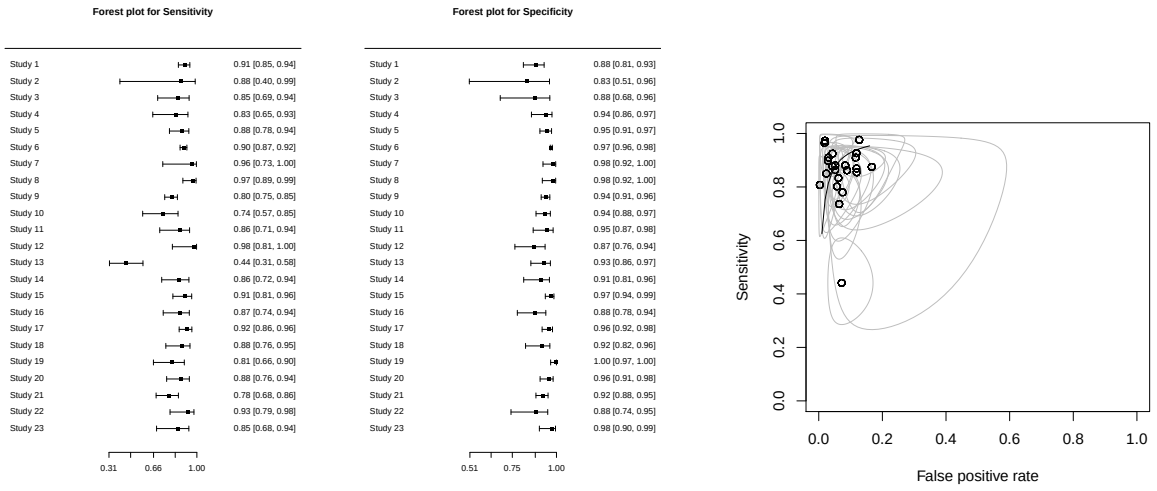
Doria et al., (2006) meta-analyzed the log-transformed sensitivity & specificity separately, and evaluated the accuracy of two diagnostic tests; ultrasonography (US) and computed tomography (CT) for diagnosis of appendicitis in children and adults. We will only consider the children data set which we call 'Children US'. This data consists of 23 studies and the average number of diseased and non-diseased children are 77 and 254, respectively. Table 4.1 shows the data set and Figure 4.1 displays the

forest plot and SROC curve with 95% confidence region. Table 4.2 presents the results of sensitivity & false positive rate estimates and their respective 95% confidence interval.

Table 4.1: The 'Children US' data from Doria et al. (2006) study.

Study	Author	FP	FN	TP	TN
1	bAng	13	14	145	102
2	Cha	1	0	3	7
3	Chang	2	4	26	18
4	Crady	4	4	22	68
5	Davidson	9	8	62	174
6	Hahn	97	50	444	3268
7	Han	1	0	13	79
8	Hayden	1	1	53	75
9	Kaiser	20	48	196	336
10	Karakas	9	9	26	138
11	Lessin	3	4	28	64
12	Lowe	7	0	20	51
13	Pena	6	28	22	83
14	Quillin	5	5	34	56
15	Ramachandran	6	5	55	206
16	Rice	7	5	36	55
17	Ronco	8	8	104	188
18	Rubin	5	5	40	60
19	Siegel	0	7	31	140
20	Sivit 92	5	6	46	123
21	Sivit 00	17	18	65	215
22	Vignault	4	2	31	33
23	Wong ML	1	4	25	61

The estimates of the sensitivity & 1-specificity pairs in percentage value are (85.8%, 6.1%) for logit, (87.2%, 5.0%) for ASR and (88.6%, 4.2%) for FTDA transformation. This result shows that there is sizable difference in the estimates of the



(a) Forest plot for 'Children US' Data

(b) SROC curve for 'Children US' Data

Figure 4.1: Forest Plot (a) and SROC curve (b) for the 'Children US' Data

two parameters of interest from the three methods of transformation. In our simulation study, we have discussed that all the methods underestimate the sensitivity and overestimate the 1-specificity. In this data example, the estimates of the sensitivity for the ASR and FTDA transformation is higher than the logit transformation by 1.4% and 2.8%, respectively, and the estimates of the 1-specificity for the ASR and FTDA transformation is lower than the estimate by logit transformation by 1.1% and 1.9%, respectively. Although it is difficult to identify the sample size as it differs substantially in each study arm for the data, this result agrees with our simulation study when the true pairs of sensitivity & 1-specificity is (95%, 10%) and the sample size is small.

Table 4.2: Estimates of sensitivity & 1-specificity and their respective (95% confidence interval) for the 'Children US' data.

Parameter	Transformation		
	Logit	ASR	FTDA
Sensitivity	85.8% (81.5%, 89.2%)	87.2% (82.9%, 90.9%)	88.6% (84.2%, 92.4%)
1-Specificity	6.1% (4.8%, 7.7%)	5.0% (3.7%, 6.5%)	4.2% (2.9%, 5.7%)

Additionally, the width of the confidence intervals differ substantially between the three methods. Both of the proposed methods have narrower confidence interval width for 1-specificity, and the logit transformation has narrower confidence interval width than the proposed methods for sensitivity. Accordingly, the width of the confidence interval for sensitivity from both proposed methods (ASR & FTDA) is higher than that of the standard logit transformation by 0.3 and 0.5 respectively. For 1-specificity, the widths are narrower for both proposed methods than the logit transformation by 0.1.

Even though Doria et al. (2006) analyzed the log transformed sensitivity and specificity separately by ignoring the possible correlation between the two, our estimated sensitivity of 88.6% by the double arcsine transformation is the nearest to their result. On the other hand, the estimated specificity of 6% by the logit transformation agrees with their.

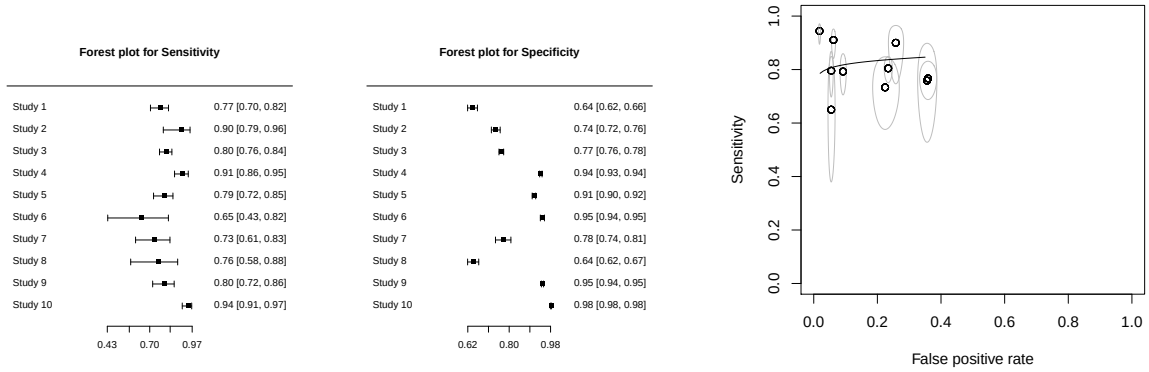
4.2 The 'VIA' Data

The VIA data is one of the three published data from three screening tests: 'visual inspection with acetic acid' (VIA), 'visual inspection with Lugols iodine' (VILI), and 'human papillomavirus' (HPV). These data were considered in Fokom-Domgue et al. (2015) to compare the performance of the three tests for primary cervical cancer screening in sub-Saharan Africa. In their study, they have employed the arcsine square root transformation to pool sensitivity and specificity when assessing the prevalence of cervical intraepithelial neoplasia grade 2 or worse (CIN2+) and positivity rate of these screening methods in sub-Saharan Africa and fit a bivariate RE model to compare the performance of the three tests in the region. The data we use is given in Table 4.3, the forest plot and SROC curve (with a 95% confidence region) of the data is displayed in Figure 4.2 and the results using the three methods discussed in this project are shown in Table 4.4.

Table 4.3: The 'visual inspection with acetic acid' ('VIA') data.

Study	Author	Country	Year	TP	FN	TN	FP
1	University of Zimbabwe	Zimbabwe	1999	158	48	1233	691
2	Sankaranarayan	Burkina Faso	2004	45	5	1485	516
3	Sankaranarayan	Congo	2004	313	76	5014	1532
4	Sankaranarayan	Guinea	2004	153	15	7935	524
5	Sankaranarayan	Mali	2004	130	34	4892	496
6	Sankaranarayan	Niger	2004	13	7	2376	138
7	De Vuyst	Kenya	2005	44	16	460	133
8	Sangwa-Lugoma	DRC	2006	22	7	965	534
9	Muwonge	Angola	2010	105	27	8238	479
10	Ngoma	Tanzania	2010	220	13	9958	183

DRC = Democratic Republic of Congo



(a) Forest plot for 'VIA' Data (b) SROC curve for 'VIA' Data

Figure 4.2: Forest Plot (a) and SROC curve (b) for the 'VIA' Data

The VIA data's average sample size in each study arm for the 10 studies is 145 and 4,778, respectively. The back-transformed estimates of the pairs of sensitivity & 1-specificity in percentage points for logit, arcsine and double arcsine transformation are (82.4%, 12.6%), (82.5%, 15.0%) and (83.0%, 15.0%) respectively.

Table 4.4: Estimates of sensitivity & 1-specificity (95% confidence interval) for the 'VIA' data.

Parameter	Transformation		
	Logit	Arcsine	Double Arcsine
Sensitivity	82.4% (76.2%, 87.3%)	82.5% (77.0%, 87.3%)	83.0% (77.4%, 87.9%)
1-Specificity	12.6% (6.6%, 22.9%)	15.0% (7.8%, 24.0%)	15.0% (7.8%, 24.0%)

In the VIA data example, Table 4.4 shows that the results of the mean sensitivity estimates are similar as they range from 82.4% to 83.0%. The result from the logit

transformation is exactly the same as the one Fokom-Domgue et al. (2015) reported as they have also fitted the same model. The increasing pattern in the estimates of the 1-specificty for the proposed methods than the estimates from the logit transformation reflects our simulation study, as it has been reported in Table 3.2 except when the sample size is small and the true sensitivity & 1-specificity pair is (95%, 10%).

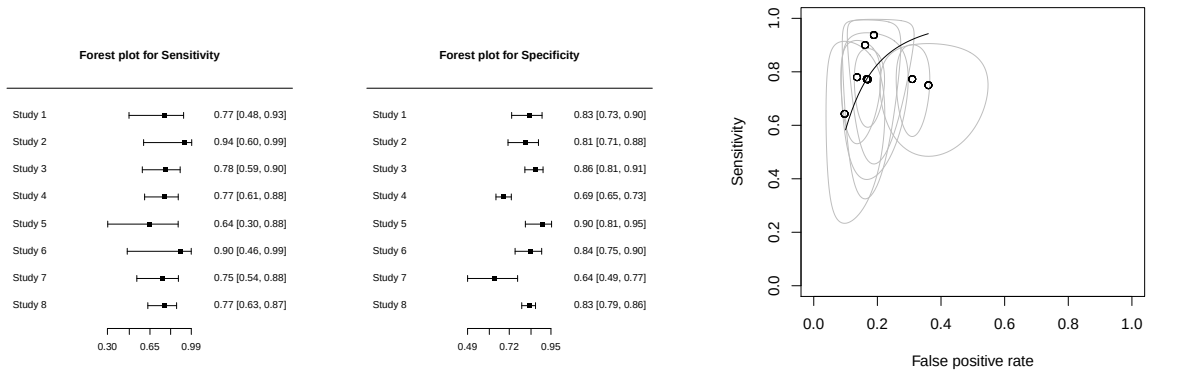
In terms of the width of confidence intervals, the proposed methods appear to have narrower values. The logit transformations width of confidence interval for sensitivity is higher than those from the ASR and FTDA transformation by 0.8 and 0.6 respectively. Similarly for 1-specificity, the proposed methods (ASR and FTDA) have lower confidence interval width by 0.1 than the logit transformation.

4.3 The 'Cytology' Data

Table 4.5 presents another published data ('Cytology') from Kocken et al. (2012). Their study aimed to evaluate the performance of three tests: 'high-risk human papillomavirus testing', 'Cytology test' and 'Co-testing' for high-grade cervical disease. The 'Cytology' data has eight studies & average sample size of 19 and 170 in each study arm. Kocken et al. (2012) analyzed the double arcsine transformed sensitivity and specificity by fitting a univariate FE and RE MA models. We will here give the results of the analysis for this data using bivariate RE model in Table 4.6. Before that, we have also presented the 'Cytology' data's forest plot and SROC curve with a 95% confidence region in Figure 3.3.

Table 4.5: The 'Cytology' data

Study	Author	Year	TP	FP	FN	TN
1	Cecchini	2004	8	12	2	62
2	Sarian	2004	7	15	0	66
3	Alonso	2006	19	24	5	155
4	Kreimer	2006	25	140	7	313
5	Verguts	2006	4	6	2	60
6	Smart	2010	4	15	0	80
7	Heymans	2011	16	15	5	27
8	Kocken	2011	35	63	10	310



(a) Forest plot for 'Cytology' Data (b) SROC curve for 'Cytology' Data

Figure 4.3: Forest Plot (a) and SROC curve (b) for the 'Cytology' Data

According to Table 4.6, the back-transformed estimated values of the pair of sensitivity & 1-specificity in percent are (77.1%, 19.2%), (79.2%, 18.8%) and (82.0%, 16.0%) for logit, ASR and FTDA transformation respectively. The result obtained from the ASR transformation has the closest agreement with the reported estimated value for sensitivity & specificity pair of (79%, 81%) in Kocken et al. (2012) study.

However, the result from FTDA transformation, the same transformation used in their study, has higher estimates of mean sensitivity & specificity pair (82.0%, 84%) when the bivariate RE model is fitted.

Table 4.6: Estimates of sensitivity and 1-specificity (95% confidence interval) for the 'Kocken' data.

Parameter	Transformation		
	Logit	Arcsine	Double Arcsine
Sensitivity	77.1% (69.7%, 83.1%)	79.2% (72.4%, 85.3%)	82.0% (74.6%, 88.5%)
1-Specificity	19.2% (14.2%, 25.4%)	18.8% (13.5%, 24.8%)	16.0% (13.1%, 24.5%)

It can be observed from Table 4.6 that the estimates of the mean sensitivity & 1-specificity pairs differ substantially. The estimated sensitivity from the logit transformation is lower by 2.1% and 4.9% than ASR and FTDA transformation respectively. Contrarily, the estimated 1-specificity from the logit is higher by 0.4% and 3.2% than the ASR and FTDA transformation, respectively.

There is also a notable difference in terms of confidence interval width between the methods. The width of the confidence interval for sensitivity from the logit transformation is higher by 0.5 than the ASR transformation and lower by 0.4 than the FTDA transformation. Whereas, the confidence interval width of the logit transformation for 1-specificty is lower by 0.1 and 0.2 than the ASR and FTDA transformation, respectively.

The results from Table 4.6 are in agreement with our simulation study as it has been reported in Chapter 3 that the logit transformation severely underestimates

sensitivity and overestimates 1-specificity when the sample size is small and the true pairs of parameter of interest is (95%, 10%).

To conclude this Chapter, we have observed that all methods produced similar results with the ones we have reported in our simulation study. Moreover, we have noticed that the use of the recommended bivariate random-effects model resulted in better point estimates of the two parameters of interest, sensitivity & 1-specificity, than the estimates obtained when analyzing the parameters separately.

Chapter 5

Summary, Discussion and Future Directions

Meta-analysis (MA) is the statistical aggregation of effect sizes. It has been used in many fields of study, including medicine. A diagnostic test is a procedure for identifying or categorizing patients in accordance to their disease status (with or without disease). Quantifying the accuracy of a diagnostic test is important as inaccuracies result in mistreatment, particularly, when the patient truly has the disease. Sensitivity and specificity are the two most commonly used methods of measuring the accuracy of a diagnostic test. The sensitivity of a diagnostic test is the ability of the test to correctly detect when a person has the disease from a patient who has it, and specificity is the ability of a test to correctly detect when a person does not have a disease when the patient does not have it.

Several meta-analytic models to synthesize the sensitivity and specificity of a

diagnostic test have been proposed in the literature. In this thesis, we have focused on the Reitsma et al. (2005) bivariate RE model because of its popularity in application and ease of understanding. However, we have adapted the model of Reitsma et al. (2005) by introducing two new transformations —arcsine square root and Freeman-Tukey double arcsine transformation.

We performed an extensive simulation study with the aim of assessing the relative performances of three transformations: the traditional logit, and two transformations we proposed in this thesis—the arcsine square root and the double arcsine transformation of Freeman-Tukey, in terms of bias, RMSE, and coverage probability. In addition to the univariate bias and MSE, we have compared the methods in terms of a vector-valued bias and MSE combining the two parameters, sensitivity and 1-specificity, also known as false positive rate. Finally, we have examined how the methods perform in real data by applying each method to three published data sets. All the analysis for the logit transformation was done using the 'mada' package of the R statistical software, while, for both of the proposed methods, the 'mvmeta' package has been employed.

In our simulation study, we have varied the pairs of sensitivity & 1-specificity, the between-study variance pairs and correlation, the sample size and number of studies. The values considered in the simulation scenario are adopted from two studies: Hamza et al. (2008) and Doebler et al. (2012). We have observed that the true pairs of between-study variances and correlation have no substantial impact on the results of the performance evaluation measures, agreeing with Hamza et al. (2008) study, although they used mean log diagnostic odds ratio as a parameter of

interest. However, the true values of sensitivity & 1-specificity, the mean sample size and number of studies influence the results of the performance measures.

We have found that the bias from the three methods decreases as sample size increases, underestimates the sensitivity and overestimates false positive rate, which is consistent with Doebler et al. (2012) who use the logit transformation. We have also found that for small sample sizes and very large sensitivity and specificity, the proposed methods have better bias in sensitivity regardless of the number of studies. However, when the sample size is small and the sensitivity and specificity are reasonable, the methods perform similarly in terms of bias, both in sensitivity and false positive rate. However, for moderate and large sample sizes, regardless of the number of studies, the logit transformation has performed better regarding bias compared to the proposed transformations for all scenarios of the sensitivity and false positive rate pairs considered. The constant absolute bias over moderate and large number of studies is consistent with the logit transformation's behavior in Eusebi et al. (2014).

The RMSE generally decreases as both the sample size and number of studies increase, which agrees with Doebler et al. (2012) for the logit transformation. This study revealed that in terms of RMSE, both of the proposed transformations outperform the standard logit for most of the scenarios. Contrarily, the logit transformation is slightly favorable in terms of RMSE when the number of studies vary from moderate to large and the sample size is large despite the sensitivity and false positive pairs considered in this study.

Although the coverage probability is increasing with sample size as expected, the methods did not perform well in most of the scenarios as the coverage is often

less than the nominal 95%. The surprisingly less than 10% coverage probability values observed when the sensitivity & specificity are very large and the number of studies is large is consistent with Hamza et al. (2008) study when they used the logit transformation and mean log diagnostic odds ratio as parameter of interest. In our study, this happens since the bias was large but the standard error was small. This will then make the confidence intervals narrow and true values to be outside of the intervals. Generally, the logit transformation outperformed the proposed methods in terms of coverage probability.

Regarding the vector valued bias and MSE, the proposed transformations have outperformed the standard logit transformation for smaller sample sizes. However, the logit transformation has better performance than the proposed methods in vector-valued bias and MSE for all the sensitivity & specificity pairs considered in this study as the sample size and number of studies gets larger.

We have also illustrated the methods discussed in this thesis on real data. The pattern of the point estimates from the methods mirrors our simulation results. The point estimates of sensitivity from the logit transformation is relatively smaller than the estimates of sensitivity from the proposed methods. Contrarily, the point estimates of the 1-specificity from the commonly used logit transformation is higher than those estimates from the proposed methods. Moreover, the proposed methods of transformation appear to have narrower confidence interval width than the usual logit transformation.

In general, we recommend either of the proposed transformations: the arcsine square root or the Freeman-Tukey double arcsine transformation when the sample

size is small, the sensitivity and specificity values are very large regardless of the between-study variances and correlation values. For moderate and large sample sizes, if the desire is accuracy, we recommend one to use the logit transformation, however, if precision is desired, we rather prefer the proposed arcsine-based transformations as the large bias might be traded-off for low variance, and, generally lower values of the precision measure, MSE, is desired.

We have evaluated the performance of a single diagnostic test using several estimation properties. However, in reality, the quantification of the accuracy of a single diagnostic test alone is not sufficient, as there can be more than one competing test to diagnose the same disease. Therefore, in the future, we aim at extending our simulation study to the case where there is more than one test and compare the performance of the proposed methods in terms of hypothesis testing optimality properties like power. We can also include other transformations like the Doebler et al. (2012) known as the 't_α family of transformations'. In this thesis, we have observed the consideration of the same true between-study variances for three of the transformations is unrealistic. Therefore, in the future, we aim to solve this issue by choosing more realistic true values of between-study variances based on assessment of real meta-analysis data sets from the medical literature. Moreover, we are interested in extending the assessment of our methods with the standard one by fitting the bivariate RE model of Chu et al. (2010)—the "exact likelihood" approach.

Bibliography

Arends L.R., Hamza T.H., van Houwelingen J.C., Heijenbrok-Kal M.H., Hunink M.G.M. and Stijnen T. (2008). "Bivariate Random Effects Meta-Analysis of ROC Curves". *Medical Decision Making*. **28**: 621–638.

Borenstein M, Hedges L.V., Higgins J.P.T. and Rothstein H.R. (2009). *Introduction to Meta-Analysis*. West Sussex: Wiley.

Burton A, Altman D. G., Royston P., and Holder R. L. (2006). "The design of simulation studies in medical statistics". *Statistics in Medicine*. **25**:4279 –4292.

Chu H. and Cole S.R. (2006). "Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach". *Journal of Clinical Epidemiology*. **59**:1331-1332.

Chu H., Guo H. and Zhou Y. (2010). " Bivariate Random Effects Meta-Analysis of Diagnostic Studies Using Generalized Linear Mixed Models". *Medical Decision Making*. **30**:499–508.

Doebler P. (2015). "mada: Meta-Analysis of Diagnostic Accuracy". *R package version 0.5.7*. http://CRAN.R-project.org/package=mada.

Doebler P., Holling H. and Bhning D. (2012). "A Mixed Model Approach to Meta-Analysis of Diagnostic Studies With Binary Test Outcome". *Psychological Methods*. http://www.personal.soton.ac.uk/dab1f10/psychomethods2012.pdf.

Doria A.S., Moineddin R., Kellenberger C.J., Epelman M., Beyene J., Schuh S., Babyn P.S. and Dick P.T. (2006). "US or CT for Diagnosis of Appendicitis in Children and Adults? A Meta-Analysis". *Radiology*. **241**(1):83–94.

Eusebi P., Reitsma J.B. and Vermunt J.K. (2014). "Latent class bivariate model for the meta-analysis of diagnostic test accuracy studies". *Medical Research Methodology*. **14**(88):1–9.

Fokom-Domgue J., Combescure C., Fokom-Defo V., Tebeu P.M., Vassilakos P., Kengne A.P. and Petignat P. (2015). "Performance of alternative strategies for primary cervical cancer screening in sub-Saharan Africa: systematic review and meta-analysis of diagnostic test accuracy studies". *the BMJ*. **351**.

Freeman M.F. and Tukey J.W. (1950). "Transformations Related to the Angular and the Square Root". *Annals of Mathematical Statistics*. **21**: 607–611.

Gasparrini A. (2015). "mvmeta: Multivariate and Univariate Meta-Analysis and Meta-Regression". *R package version 0.4.7*. http://www.ag-myresearch.com/package-mvmeta.

Hamza T.H., Reitsma J.B. and Stijnen T. (2008). "Meta-Analysis of Diagnostic Studies: A Comparison of Random Intercept, Normal-Normal, and Binomial-Normal Bivariate Summary ROC Approaches". *Medical Decision Making*. **28**:

639–649.

Hedges L.V. and Vevea J.L. (1998). "Fixed- and Random-Effects Models in Meta-Analysis". *Psychological Methods.* **3**(4): 486–504.

Jennrich R.I. and Schluchter M.D. (1986). "Unbalanced Repeated-Measures Models with Structured Covariance Matrices". *Biometrics.* **42**(4):805–820.

Kocken M., Uijterwaal M.H., de Vries A.L.M., Berkhof J., Ket J.C.F., Helmerhorst T.J.M. and Meijer C.J.L.M. (2012). "High-risk human papillomavirus testing versus cytology in predicting post-treatment disease in women treated for high-grade cervical disease: A systematic review and meta-analysis". *Gynecologic Oncology.* **125**: 500–507.

Kovalchik, S. (2013). " Tutorial On Meta-Analysis In R". *R useR.*

Kuss O., Hoyer A. and Solms A. (2013). "Meta-analysis for diagnostic accuracy studies: a new statistical model using beta-binomial distributions and bivariate copulas". *Statistics in Medicine.* **33**(1):17–30.

Moses L.E., Shapiro D. and Littenberg B. (1993). "Combining Independent Studies Of A Diagnostic Test Into A Summary ROC Curve: Data-Analytic Approaches And Some Additional Considerations". *Statistics in Medicine.* **12**: 1293–1316.

R Core Team (2015). "R: A language and environment for statistical computing." *R Foundation for Statistical Computing, Vienna, Austria.* http://www.R-project.org/.

Reitsma J.B., Glas A.S., Rutjes A.W.S., Scholten R.J.P.M., Bossuyt P.M. and Zwinderman A.H. (2005). "Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews". *Journal of Clinical Epidemiology.* **58**: 982–990.

Rutter C.M. and Gatsonis C.A. (2001). "A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations". *Statistics in Medicine.* **28**:2865–2884.

Schoenberg R. (2001). "Optimization with the Quasi-Newton Method".*Aptech Systems, Inc. Maple Valley, WA.*

Trikalinos T.A., Trow P. and Schmid C.H. (2013). "Simulation-Based Comparison of Methods for Meta-Analysis of Proportions and Rates. Methods Research Report. (Prepared by the Tufts Medical Center Evidence-based Practice Center under Contract No. 290-2007-10055-I.)". *AHRQ Publication No. 13(14)-EHC084-EF. Rockville, MD: Agency for Healthcare Research and Quality.* www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Venables W.N. and Ripley B.D. (2002). *Modern Applied Statistics with S, Fourth Edition.* Springer, New York. ISBN 0-387-95457-0.

Viechtbauer W. (2010). "Conducting Meta-Analyses in R with the metafor Package". *Journal of Statistical Software*, **36**(3). http://www.jstatsoft.org/v36/i03/paper.

Walter S.D. (2002). "Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data". *Statistics in Medicine.* **21**: 1237–1256.