GENERALIZED LINEAR MIXED MODELS FOR ANALYSIS OF CROSS-CORRELATED BINARY DATA IN MULTI-READER STUDIES OF DIAGNOSTIC IMAGING

by

Yuvika Paliwal

BSc in Mathematics, Statistics, Computer Science, Banasthali Vidyapith, India, 2004

MS in Statistics, University of Massachusetts, 2008

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Yuvika Paliwal

It was defended on

April 12, 2017

and approved by

Dissertation Advisor: Andriy I. Bandos, PhD
Assistant Professor, Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

Committee Members:

Abdus S. Wahed, PhD
Professor, Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

Dianxu Ren, PhD
Associate Professor, School of Nursing
University of Pittsburgh

Joyce Chang, PhD
Professor, Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

Ying Ding, PhD
Assistant Professor, Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

Andriy I. Bandos, PhD

GENERALIZED LINEAR MIXED MODELS FOR ANALYSIS OF CROSS-CORRELATED BINARY DATA IN MULTI-READER STUDIES OF DIAGNOSTIC IMAGING

Yuvika Paliwal, PhD

University of Pittsburgh, 2017

ABSTRACT

Cross-correlated data occur in multi-sample studies with a fully crossed design. An important type of binary cross-correlated data results from multi-reader diagnostic imaging studies where each of several readers independently evaluates the same sample of subjects for the presence or absence of a specific condition (e.g., disease).

The analysis of the fully crossed studies can be challenging because of the need to address both reader and subject variability and the related correlation structure. Generalized Linear Mixed Models (GLMM) are implemented in standard statistical software and offer a natural tool for the analysis of the cross-correlated data in the presence of covariates. However, performance of GLMMs for cross-correlated binary data from typical multi-reader studies is generally unknown and is questionable due to the specifics of the available estimation approaches.

In the first part of the dissertation we investigate the standard built-in GLMM methods for cross-correlated binary data with and without covariates and explore simple combinations of the built-in estimation techniques to overcome existing deficiencies. In the second part, we propose a half-marginal GLMM approach which offers a superior interpretation in the context of multi-reader studies of diagnostic accuracy. Our investigation of this model demonstrates good quality of statistical inferences in typical scenarios, but indicates possible large-sample problems stemming from the pseudo-likelihood estimation approach. In the third part of the dissertation we develop an explicit approach for estimating half-marginal model parameters without using pseudo-likelihood. The consistent fixed-effect estimator and its variance are evaluated in an extensive simulation study. The proposed approach can be implemented using the non-iterative combination of results from several robust Generalized Estimating Equation (GEE) models and, for simple scenarios, provides estimates that are equivalent to the empirical estimates.

Public Health Significance: Analyses of cross-correlated data from multi-reader studies are used to evaluate performance of medical diagnostic technologies at their development and regulatory approval stages. Enhanced methods of performance assessment help improve and accelerate optimal adaptation of diagnostic and screening technologies in clinical practice.

TABLE OF CONTENTS

LIST OF TABLES

PREFACE

I would like to thank my dissertation advisor, Dr. Andriy I. Bandos whose acceptance of me as a PhD student, guidance, patience and continued motivation during this entire journey has not only helped in earning this degree but also increased my understanding of statistical theory. I would also like to thank my committee members Dr. Joyce Chang, Dr. Ying Ding, Dr. Abdus S. Wahed and Dr. Dianxu Ren for serving on my dissertation committee and each providing their expert guidance in improving the quality of this dissertation. A special thanks to all the inspiring professors who taught me various statistical courses during the graduate program and shaped my basic foundation and ability to pursue my work as a PhD student. I feel very thankful to Dr. Abdus S. Wahed for his continued support and for lending an ear when needed. Thanks to Joanne Beer, for being a great friend and classmate, my experience at Pitt wouldn’t have been the same without you.

Finally, I am extremely grateful for my family – my husband, an extremely patient four year old son, my supportive parents and in-laws who always stood as pillars of strength, providing continuous emotional and mental support during this long but exciting journey. I dedicate this work to them.

1.0 INTRODUCTION

Cross-correlated data arise from experiments in many fields where the measures of interest depend on combination of factors from different populations. In medical imaging, cross-correlated data result from the “fully-crossed” multi-reader studies where readers evaluate the same set of subjects [2]. Both readers and subjects are sampled from the corresponding target populations (e.g., a sample of certified breast-imaging radiologists and a sample of patients undergoing evaluations of suspected lesions). The typical analytical goals are the evaluation of the accuracy characteristics of a diagnostic technology or/and comparisons across several diagnostic modalities in the targeted populations of readers and subjects. Usual multi-reader studies include around five to eight readers and fifty to hundreds of subjects with and without the condition of interest. Since human observers naturally have different abilities and experience, there is a natural variability in readers’ performance levels; the fully-crossed multi-readers studies provide an efficient way to make inferences about diagnostic accuracy accounting for variability due to readers and subjects [2]. To highlight the two important sources of variability, these multi-readers studies are frequently called Multi-Reader Multi-Case (MRMC) studies.

A variety of traditional approaches exist for analyzing data from MRMC studies. A number of methods address the so-called “fixed-reader” question, where readers are considered fixed factors and between-reader variability does not directly affects the overall variance [3, 4, 5, 6, 20]. Another category of methods address the so-called “random-reader” question, where readers are recognized as being sampled from a target population of readers and the between-reader variability is incorporated as a part of the overall variance [7, 8, 9, 10, 11, 12, 13, 20, 55, 56]. The random reader approaches extend inferences to the populations of subjects and readers, which is typically necessary for adequate evaluation of diagnostic technologies and practices. Some of these approaches [7, 8, 13] can model several different types of accuracy measures (e.g. AUC, sensitivity) and take into account the correlations and variability present in the data. These methods treat the outcome (e.g., the modeled summary index) as a continuous and asymptotically normal variable, but nevertheless perform quite well and are rather useful for covariate-free comparison of diagnostic modalities. However, majority of them do not allow incorporating continuous covariates (e.g., subjects' age, lesion size). Approaches that do allow handling various covariates include hierarchical Bayesian approach for MRMC ROC data [14, 15], tweaked original regression approach for ROC curves [16], and tweaked GEE approach for area under the ROC curves [17]. All these model-based approaches were developed for inferences about higher-level ROC indices, but can technically be applied to binary MRMC data, possibly after some adjustments. However, quality of inferences in such applications is not known and most importantly these methods require custom-made software for computing variance accounting for between-reader variability.

Outside of the traditional tool box, standard statistical packages provide flexible tools for MRMC analysis of binary cross-correlated data with covariates: namely, functions that solve Generalized Linear Mixed Model (GLMM) with non-linear link functions (e.g., PROC GLIMMIX in SAS v9.4, package "lme4" in R v3.3.2 [25]). Unlike the Generalized Estimating Equations (GEE) models which can only address nested levels of clustering, GLMMs can be used to account for the covariance in cross-correlated data by inclusion of random crossed factors (e.g., [40]).

Models with crossed random effects provide the most conventional approach to model cross-correlated data. For example, such models have been repeatedly considered for the classic Salamander data, which are binary cross-correlated data from a mating experiment [1]. However, estimating GLMM parameters for binary data is not straightforward because of the complexity of the marginal likelihood function. The standard estimation techniques can be broadly classified into two categories: (1) methods to approximate marginal likelihood and (2) methods to approximate the model. The most straightforward approach to approximate the marginal likelihood is based on Gauss-Hermite quadrature which is computationally infeasible for cross-correlated data with multiple factors and is excluded from available estimation techniques in standard packages (e.g., PROC GLIMMIX and PROC NLMIXED in SAS, package “lme4” in R). Another approach to approximate the marginal likelihood is based on the Laplace approximation [26] which can be implemented using PROC GLIMMIX in SAS, function “glmer” in the library “lme4” in R. The most well-known approach to approximating the model is the pseudo-likelihood approach [21], which can be implemented using PROC GLIMMIX, SAS and function “glmmPQL” in library “nlme” in R (however, function “glmmPQL” does not currently support crossed random effects).

The Laplace approximation (LA) provides reliable results for many scenarios, especially for large cluster sizes [26, 27]. However, scenarios when the dimension of the random effects increases with sample size, as happens in case of crossed random effects, the Laplace approach has been criticized for producing poor estimates especially for variance components [28, 37].

The Pseudo-Likelihood (PL) approach is a much less computer intensive and more versatile approach for estimating GLMMs [21, 29]. The underlying idea is to approximate the non-linear model with a specific linear model at each iteration step (initially described by Lindstrom and Bates, 1990 [30]). However the PL estimates for simple clustered binary data can lead to biased estimates of model parameters when there are only a few observations per cluster, or when random effects have large variances [31, 32, 33, 34, 35].

Much of the previous GLMM investigation has been focused on simple clustered data. Little is, however, known about properties of GLMM for binary cross-correlated data of type resulting from fully crossed multi-reader studies. Good properties of some linear methods for MRMC analysis [7, 8], as well as some non-linear GLM for fixed-reader inferences (e,g., Toledano and Gatsonis, 1996 [3]) are encouraging in regard to possible performance of GLMMs for these types of data.

A number of more computer intensive methods have been proposed in literature to possibly fix some of the issues with standard approaches. Some are based on modified Laplace approaches (e.g., [28, 37, 38]), some on model linearization (e.g., bias-corrected PL by Breslow and Lin, 1995 [31]; Lin and Breslow, 1996 [36]), and others are alternative estimation techniques [25, 35, 39]. Some of these methods are able to handle cross-correlated data within the GLMM framework. However, none of these have yet been developed and validated enough to be incorporated in standard statistical software, and have application-specific fitting problems. For example, one of the simplest approaches is a hybrid approach [25] combining Bayesian estimates of variance components with the Laplace-based estimates of the fixed (and random) effects. This specific approach requires programming by the user and needs to be modified every time in terms of the priors that one inputs for the prior distribution of variance components. Furthermore, from our experience, successful implementation of the Bayesian component of this technique relies heavily on providing reasonable prior values which seems to be data dependent and also sensitive to the choice of the algorithm to generate posterior samples. Hence, although the simulation results provided in the original paper look promising, the method does not yet allow straightforward implementation. As another example listed earlier, Rulis *et al.* (2016) [28] offers a modification to the standard Laplace approach which can be implemented using R package iLaplace. However, the package requires the user to code the log integrand along with it's first and second order derivatives which can be rather complicated for complex models.

Thus, currently a practicing statistician faced with the analysis of cross-correlated multi-reader data with covariates has a choice of either using the built-in techniques with potential reliability issues, or turning to Bayesian approaches based on Gibbs Sampling (e.g., [40]) which are computationally intensive, and has their own host of problems. In these settings use of built-in GLMM techniques is again rather tempting because of 1) availability and 2) possibility of good performance in fully crossed multi-reader studies.

1.1 NOTATIONS AND CONVENTIONS

In the simplest fully crossed multi-reader study each reader provides a binary response (e.g., Y = 0 for "test negative"; Y = 1 for "test positive") for every subject. For a study with n_r readers and n_c subjects the resulting data ($Y_{ij}, i = 1, \dots, n_c; j = 1, \dots, n_r$) can be arranged in a matrix with rows corresponding to observations for different subjects and columns corresponding to different readers. Observations in the same row (Y_{ij} and $Y_{ij'}$) or column (Y_{ij} and $Y_{i'j}$) are correlated due to "sharing" the same subject or reader. Such cross-correlated data does not allow defining independent clusters as is needed for the application of standard clustered data analysis (e.g., based on GEE, [18]).

In addition to the primary response Y, diagnostic imaging studies usually provide additional covariate information. The most typical covariate is the "true or reference status" of a subject (e.g., D = 1 for “diseased”; D = -1 for “non-diseased”), which is a subject-level covariate used to define sensitivity ($Pr(Y = 1|D = 1)$) and specificity ($Pr(Y = 0|D = -1)$). We shall use equivalent terms: the True Positive Fraction (TPF) equals sensitivity, and the False Positive fraction (FPF) equals 1-specificity. Another typical “assessment-level” covariate is the set of diagnostic conditions (“diagnostic modality”), which is of primary interest in studies comparing diagnostic technologies or practices [19]. Other covariates at subject or reader level can also be of interest (e.g., lesion size, years of radiologist’s experience).

1.2 OBJECTIVES

This dissertation is focused on investigating properties of GLMMs for analysis of cross-correlated binary data from multi-reader diagnostic imaging studies and developing simple approaches to correct existing deficiencies in standard GLMM tools for analyses of these studies. The overall goal of this work is to develop guidelines and necessary tools to enable straightforward covariate-based analysis of multi-reader studies of diagnostic accuracy in frequently encountered settings. The specific objectives are outlined below:

1.2.1 Crossed-random effect GLMM for analysis of binary data from fully crossed multi-reader studies of diagnostic accuracy

Analysis of cross-correlated data cannot be handled using a conventional GEE mechanism, but can be performed using GLMM methodology with crossed random effects (often called, “subject-specific” models). GLMMs offer an attractive alternative analytical technique due to 1) ability to handle any and multiple types of covariates, 2) ability to account for variability of crossed factors, 3) availability of built-in tools in standard software packages. In this chapter, we design a simulation study representing a wide range of possible analyses of multi-reader data. We investigate the properties of statistical inferences under the conventional GLMMs along with two ad-hoc approaches based on the same GLMMs for analyses of binary cross-correlated MRMC data. The results of this study provides guidelines on using conventional GLMM as well as the ad-hoc approaches for analyzing data typical for multi-reader diagnostic imaging studies.

1.2.2 Half-marginal GLMM for analysis of cross-correlated binary data in multi-reader studies of diagnostic accuracy

In multi-reader studies of diagnostic imaging, the frequent targets of interest are the quantities marginalized over the population of subjects (e.g., sensitivity and specificity). The marginal characteristics are only indirectly related to the parameters of the standard models based on crossed random effects, which we considered previously. In addition, it is common to make inferences for individual readers as well as for the average over all readers [2], which is even more complicated to achieve based on the results of the "subject-specific" models. In this chapter we propose a half-marginal model which, marginalizes over the population of subjects but not readers, and enables direct inferences about marginal characteristics for individual readers and overall. This makes the half-marginal model practically relevant for analysis of multi-reader studies of diagnostic imaging. Furthermore, because of the elimination of crossed random effects estimated by standard GLMMs, the half-marginal model can offer both statistical and computational advantages. Although the proposed model can be fit using built-in machinery, it is rarely recognized and little known. Development and assessment of half-marginal model extends the existing methodology for analysis of MRMC data by introducing a superior tool for analyzing cross-correlated multi-reader data in the presence of covariates.

1.2.3 An explicit approach for estimating half-marginal GLMM for analyzing cross-correlated binary data from multi-reader studies of diagnostic accuracy

While several estimation approaches exist for fitting crossed random effects GLMMs, the only built-in approach available for estimating half-marginal model is based on pseudo-likelihood under the linearized model [21]. The potential problems with the resulting estimates combined with non-explicit (in terms of probability distribution) nature of the half-marginal model can be discouraging. To alleviate the possible criticisms and to attempt to improve the half-marginal approach we develop an alternative approach for inferences under the half-marginal model. The proposed approach exploits the small number of readers typically available in the multi-reader studies and applies techniques similar to those used in within-cluster-resampling approach for model estimation [22], model averaging [23] and multiple imputation [24]. This part of research develops an explicit approach for obtaining consistent estimates for half-marginal models without using pseudo-likelihood. The proposed approach also lays a solid foundation for further improvement of analytical tools for analyzing cross-correlated multi-reader data.

2.0 CROSSED-RANDOM EFFECT GLMM FOR ANALYSIS OF BINARY DATA FROM FULLY CROSSED MULTIREADER STUDIES OF DIAGNOSTIC ACCURACY

Standard software packages provide a flexible tool to analyze cross-correlated binary data using the Generalized Linear Mixed Models (GLMM) which addresses heterogeneity in the samples of readers and subjects and allows accounting for covariates. However, reliability of GLMM estimates for this type of data is questionable because of possible bias in estimates noted in some applications. However, little is known about the severity and consequences of this bias for statistical inferences in data typically encountered in cross-correlated multi-reader studies. In this work we investigated the standard GLMM methods for cross-correlated binary data with and without covariates and explored a simple combination of built-in techniques that correct existing deficiencies. The primary focus of this investigation was on quality of fixed effect inferences provided by confidence intervals. In an extensive simulation study, we evaluated the coverage of confidence intervals for the fixed effects, as well as the bias and standard error of their estimates. We found that available built-in approaches fail in many practical scenarios, and that these deficiencies can be fixed by a simple combination of available techniques. Based on obtained results we provided guidelines for GLMM analysis in typical multi-reader studies.

2.1 RELEVANT GLMM BACKGROUND

Estimation in the GLMM models for cross-correlated binary data is not straightforward and can be computationally demanding because of the complexity of the marginal likelihood function. For example, consider a fully crossed multi-reader data set consisting of binary responses (e.g., Y = 0 - "test negative"; Y = 1 - "test positive") provided by a sample of readers ($j = 1, \dots, n_r$) for a sample of subjects ($i = 1, \dots, n_1$). Assuming all subjects have the condition of interest (i.e., D = 1 for "diseased"), we focus on statistical inferences of sensitivity which can be analyzed using the following GLMM:

$$logit(p_{ij}) = \mu + \alpha_i + \beta_j \, , \qquad \textbf{(2.1)}$$

where $Y_{ij}|\alpha_i, \beta_j \sim Bin(1, p_{ij})$ and μ is the logit of the conditional probability that the average reader correctly classifies the average subject, $Pr(Y_{ij} = 1|\alpha_i = 0, \beta_j = 0)$. Sharing of subject random effect $\alpha_i \sim N(0, \sigma_\alpha^2)$ induces correlation among observations from the i^{th}subject but different readers (Y_{ij}, $Y_{ij'}$). Similarly, sharing of reader random effect $\beta_j \sim N(0, \sigma_\beta^2)$ induces correlation among observations from the j^{th} reader but different subjects (Y_{ij}, $Y_{i'j}$). Variability of Y is determined by variance of subject and reader random effects as well as the binomial variability (which depends on p_{ij}).

The unconditional variance-covariance matrix for outcome Y i.e. $V(Y)$ can be shown as below (for simplicity assume $n_1 = 2$ and $n_r = 2$):

$$E[Y_{ij}|\alpha_i, \beta_j] = p_{ij} = logit^{-1}(\mu + \alpha_i + \beta_j)$$

$$V(Y) = V\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \end{pmatrix} \approx BZGZ'B + R \text{ using 1}^{\text{st}} \text{ order approximation}$$

The residual variance matrix, $R = A^{1/2}B^{1/2}PB^{1/2}A^{1/2} = B = diag\{p_{ij}(1 - p_{ij})\}$, where $A = I$ since outcome in bernoulli, correlation matrix $P = I$ since we don't have any residual-side effects.

$$V(Y) \approx BZGZ'B + A^{1/2}B^{1/2}PB^{1/2}A^{1/2} = BZGZ'B + B$$

$$B = \begin{bmatrix} p_{11}(1-p_{11}) & 0 & 0 & 0 \\ 0 & p_{12}(1-p_{12}) & 0 & 0 \\ 0 & 0 & p_{21}(1-p_{21}) & 0 \\ 0 & 0 & 0 & p_{22}(1-p_{22}) \end{bmatrix}$$

$$Z = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \text{ and } G = \begin{bmatrix} \sigma_\alpha^2 & 0 & 0 & 0 \\ 0 & \sigma_\alpha^2 & 0 & 0 \\ 0 & 0 & \sigma_\beta^2 & 0 \\ 0 & 0 & 0 & \sigma_\beta^2 \end{bmatrix}$$

$$V(Y) = \begin{bmatrix} g_1^2(\sigma_\alpha^2 + \sigma_\beta^2) & g_1 g_2 \sigma_\alpha^2 & g_1 g_3 \sigma_\beta^2 & 0 \\ g_1 g_2 \sigma_\alpha^2 & g_2^2(\sigma_\alpha^2 + \sigma_\beta^2) & 0 & g_2 g_4 \sigma_\beta^2 \\ g_1 g_3 \sigma_\beta^2 & 0 & g_3^2(\sigma_\alpha^2 + \sigma_\beta^2) & g_3 g_4 \sigma_\alpha^2 \\ 0 & g_2 g_4 \sigma_\beta^2 & g_3 g_4 \sigma_\alpha^2 & g_4^2(\sigma_\alpha^2 + \sigma_\beta^2) \end{bmatrix},$$

where $g_1 = p_{11}(1-p_{11})$; $g_2 = p_{12}(1-p_{12})$; $g_3 = p_{21}(1-p_{21})$; $g_4 = p_{22}(1-p_{22})$. This structure illustrates that even for simple models the estimation task is non-trivial.

The marginal likelihood of the above model involves averaging over subject and reader distributions and can be written as follows:

$$L(\psi) = \int \ldots \int \left\{ \prod_{j=1}^{n_r} \prod_{i=1}^{n_1} p_{ij}{}^{y_{ij}} \left(1 - p_{ij}\right)^{1-y_{ij}} f(\alpha_i) f(\beta_j) \right\} d\alpha_1 .. d\alpha_{n_1} d\beta_1 .. d\beta_{n_r}, \qquad (2.2)$$

where p_{ij} depends on μ, α_i and β_j according to equation (2.1) and $\psi = (\mu, \sigma_\alpha^2, \sigma_\beta^2)$ is a vector of unknown parameters in the model we wish to estimate. The key challenge in working with $L(\psi)$ is the inability to factor the integrand due to the cross-sharing of random effects. If the data were not correlated the ordering of integration of the product could have been interchanged leading to the structure addressable by Generalized Estimating Equations, or even by Maximum Likelihood (if the data were completely independent). However, in the presence of the cross-correlated data, it is not possible to simplify equation (2.2) and we must resort to approximations. Furthermore, the dependence of the integral dimension on the sample size makes implementation of the quadrature methods infeasible, necessitating the use of less straightforward approximations.

2.2 STANDARD ESTIMATION TECHNIQUES

2.2.1 Laplace Approximation

Laplace approximations (LA) are usually implemented to obtain a tractable expression for the parameter-dependent integral, with the goal to make numerical optimization tractable. LA is designed to approximate integrals of the following structure:

$$I = \int_{\mathbb{R}^d} e^{-Nh(x)} d\boldsymbol{x},$$

where N is the number of data points, $h(\boldsymbol{x})$ is a scalar function and $\boldsymbol{x}$ is a d-dimensional real vector. Applying the multi-variate Taylor expansion evaluated at $\tilde{\boldsymbol{x}}$ the following approximation is obtained [37]:

$$log[I] \approx -Nh(\tilde{x}) - \frac{1}{2} log[det(h''(\boldsymbol{x})|_{x=\tilde{x}})] + \frac{d}{2} log(2\pi) - \frac{d}{2} logN.$$

The h-function for the marginal likelihood in equation (2.2) is the sum of two parts: the first part is the log-likelihood conditional on the specific readers and subjects in the study,

$$\mu \sum_{ij} y_{ij} + \sum_j \beta_j y_{.j} + \sum_i \alpha_i y_{i.} - \sum_{ij} log\{1 + exp(\mu + \alpha_i + \beta_j)\},$$

$\left(y_{i.} = \sum_j y_{ij} \text{ and } y_{.j} = \sum_i y_{ij}\right)$ and the second part accounts for the randomness of the readers and subjects,

$$-n_1 log\sigma_\alpha - \sum_i \frac{\alpha_i^2}{2\sigma_\alpha^2} - n_r log\sigma_\beta - \sum_j \frac{\beta_j^2}{2\sigma_\beta^2}.$$

The resulting function is maximized with respect to both fixed effect and variance components. It can be implemented using METHOD=LAPLACE in PROC GLIMMIX, SAS.

2.2.2 Pseudo-Likelihood Method

The Pseudo-Likelihood (PL) is a model linearization technique based on Taylor series expansion and Gaussian approximation. To approximate the GLMM in equation (2.1), we take the 1st order Taylor's series expansion of $logit^{-1}(\mu + \alpha_i + \beta_j)$ about $\widehat{\mu}$, $\hat{\alpha}_i$ and $\hat{\beta}_j$:

$$logit^{-1}(\mu + \alpha_i + \beta_j) \ = logit^{-1}(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j) + (\widehat{D})_{ij}(\mu{+}\alpha_i + \beta_j - \widehat{\mu} - \hat{\alpha}_i - \hat{\beta}_j),$$

where $\widehat{D}$ is a diagonal matrix with elements consisting of first-order derivatives of $logit^{-1}(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j)$ evaluated at $\widehat{\mu}$, $\hat{\alpha}_i$ and $\hat{\beta}_j$ i.e. $diag[\hat{p}_{ij}(1 - \hat{p}_{ij})]$.

Rearranging terms, we have:

$$\widehat{D}_{ij}{}^{-1}[logit^{-1}(\mu{+}\alpha_i + \beta_j) - logit^{-1}(\widehat{\mu} + \hat{\alpha}_i + \hat{\beta}_j)] + (\widehat{\mu} + \hat{\alpha}_i + \hat{\beta}_j) = \mu + \alpha_i + \beta_j.$$

We call the left hand side of this equation $\boldsymbol{K}$; it is a vector of pseudo-variables of length $n_r \times n_c$ that we assume is Gaussian. The mean and covariance of $\boldsymbol{K}$ conditional on the true random effects are given by

$$E[K_{ij}|\ \alpha_i, \beta_j] = \ \mu + \alpha_i + \beta_j,$$

$$\text{Cov}[\boldsymbol{K}|\boldsymbol{\alpha}, \boldsymbol{\beta}] = \boldsymbol{D^{-1}}(\boldsymbol{Cov}((\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\beta}))\boldsymbol{D^{-1}} = \boldsymbol{D^{-1}},$$

since the conditional covariance of $\boldsymbol{Y}$ equals $\boldsymbol{D}$.

Thus, we define a new linear mixed model (LMM) as:

$$k_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \text{ where } Var(\boldsymbol{K}) = \boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{D^{-1}},$$

where $\boldsymbol{Z}$ is the design matrix for random effects $[(n_r * n_1) \times (n_r + n_1)]$ and $\boldsymbol{G}$ is the $[(n_r + n_1) \times (n_r + n_1)]$ variance-covariance matrix of random effects. We assume $\boldsymbol{G}$ is a diagonal matrix with σ_α^2 and σ_β^2 on the diagonal.

The PL estimation of a GLMM follows a doubly iterative algorithm. First, the LMM is established based on initial estimates of the fixed and random effects. Then using either Maximum-Likelihood (ML) or Restricted ML, the LMM is fit, yielding estimates of the elements of the $\boldsymbol{G}$ matrix: σ_α^2 and σ_β^2. Given the variance components, the mixed-model equations are then solved for the fixed effects and random effects: $\mu, \{\alpha_i\}_{i=1}^{n_1}$, and $\{\beta_j\}_{j=1}^{n_r}$. Solving for the fixed and random effects is an iterative process; it is the inner iteration. The solution of this inner iteration also includes variances of the estimates of the fixed and random effects. Given estimates of the fixed and random effects, the whole process is repeated until convergence.

2.2.3 Combination Approaches

Problems indicated in the literature for PL and LA estimation approaches led to several attempts to develop alternative combination remedies. These remedies are based on using the LA estimates of fixed effects, which are known to be rather accurate, while estimating the reference distribution or variance structure from different methods. Below we describe two such remedies that we will use.

Noting problems with statistical inferences, Stroup (2012) [44] suggested using the Satterthwaite degrees of freedom from the PL approach for the same model. In our work, we also consider this suggestion when trying to make statistical inferences using the LA estimation technique since it does not allow the Satterthwaite option.

As another example, Capanu *et al.* [25] suggested a hybrid technique in which Bayesian approaches are used to estimate the variance components of the random effects. These in turn are used to estimate the fixed-effects and corresponding standard errors (SE) based on fitting a model using LA approach. We explore a similar remedy. Based on preliminary studies we found that the PL approach had few convergence issues and produced accurate estimates of the overall variance of the fixed effects. These properties led us to combine the LA estimates of the fixed effects with the PL estimates of the variances in the instances when the LA approach resulted in a non-positive definite G matrix containing estimates of variance components. In instances when the G matrix was not positive definite for either approach, we picked the larger SE estimates to enable more stable statistical inferences.

In general, when we try to address the variability in the data through random effects, the problems of the inadmissible (negative values) estimates of variance and variance components becomes quite common during estimation. The chances become higher especially when variance components and sample sizes are small (Brown and Prescott [59]). These problems are typically handled by setting inadmissible estimates to ‘0’ (e.g.) which possibly induces bias in the results. However, due to specifics of estimation approaches the problems with inadmissible estimates of variance components are substantially more frequent for LA than for PL approach. The essence of the combination approach we propose is to use the PL estimates of variability in these instances.

To summarize, we implement two combination approaches:

1) Using LA estimates and borrowing Satterthwaite degrees of freedom from the same model fitted using PL.
2) Using LA fixed effect estimates and borrowing PL standard error estimates when estimated G-matrix using LA approach is not positive-definite. SAS default containment degrees of freedom are used in this approach.

2.3 TYPICAL MODELS FOR ANALYSING MULTI-READER STUDIES OF DIAGNOSTIC ACCURACY

In the simulation study we considered four models that address the frequent analytical questions in multi-reader studies of diagnostic imaging. While the models are not exhaustive of possible models in fully crossed multi-reader studies, they illustrate handling of the basic types of covariates. For each modeling setting we estimated the parameters of the models using the PL, LA and the combination approaches under various parameter configurations. It is also worthwhile to emphasize that the estimated parameters for each model are not "marginal" or "population-averaged" quantities but rather "subject-specific" quantities in the traditional sense. Marginal estimates and their corresponding confidence intervals, however, can be derived using the "subject-specific" model (Section 3.4). The structure of the GLMM models for all models are summarized in Table 1. Below we provide full specifications for each model:

2.3.1 Model A: Covariate free model (e.g., inferences on sensitivity or specificity for a single modality)

Performance of the covariate-free model is of interest and it provides a reference for more complicated models. The covariate-free GLMM can be used for estimating a single proportion in the fully crossed multi-reader study. This proportion can be sensitivity, specificity, or percent agreement. Here we focus on estimating sensitivity.

For inferences about the sensitivity under a given diagnostic modality it is natural to use the model defined using equation (2.1). This model takes into account variability between readers, variability between subjects, as well as the resulting correlations between observations by the same reader or from the same subject.

The estimation target for this model is μ (average logit), which is the log-odds of the probability that the average diseased subject is correctly diagnosed by the average reader; or alternatively, the corresponding reader-subject specific probability, i.e., $p = (1 + \exp(-\mu))^{-1}$. Exclusion of '0' from a 95% confidence interval (CI) for μ can be used for testing the difference of the reader-subject specific sensitivity from 0.5: $H_o: \mu = 0\ vs. H_1: \mu \neq 0$ is equivalent to $H_o: p = 0.5\ vs. H_1: p \neq 0.5$. Similar hypotheses can be set up for different values of reader-subject-specific sensitivities. More general inferences for this and other considered models can be based on integrating the confidence interval using the estimated fixed effect and variance components parameters (e.g., as described in Section 3.4).

2.3.2 Model B: Subject-level binary covariate (e.g. inferences on sensitivity and specificity combined)

The most common type of binary subject-level covariate in diagnostic imaging studies is the true status of a subject (e.g., truth="diseased" or "non-diseased"). In models with logit link this covariate is closely related to a "Diagnostic Likelihood Ratio" [42]. Naturally, inferences about the true status-related covariate requires modeling TPF and FPF simultaneously for which one can use the following GLMM:

$logit(p_{ij}) = \mu + \eta_D + \alpha_i + \beta_j + \gamma_{jD}$,

where $Y_{ij}|\alpha_i, \beta_j, \gamma_{jD} \sim Bin(1, p_{ij})$, $i = 1, \ldots, n_0 + n_1$ is the subject index, n_0= number of non-diseased subjects, n_1= number of diseased subjects, $j = 1, \ldots, n_r$ is the reader index, η_D is the fixed effect for true disease status ($D = 1{:}\, diseased, -1 = non - diseased$) such that D depends on index i, $\alpha_i \sim N(0, \sigma_\alpha^2)$ is the subject random effect, $\beta_j \sim N\left(0, \sigma_\beta^2\right)$ is the reader random effect, $\gamma_{jD} \sim N(0, \sigma_D^2)$ is the random effect of the interaction between true disease status and reader.

For simplicity, we consider similar distribution for the diseased and non-diseased subjects. However, a model with different variability for diseased and non-diseased subject effects can also be implemented. Apart from the usual variability and correlations modelled by introducing subject and reader random effects (as in model A), the random interaction effect between reader and true disease status $\left(\gamma_{jD}\right)$ models varying differences between true positive fraction and false positive fraction for each reader through σ_D^2.

The primary inferential target for this model is $\eta = \eta_1 - \eta_{-1} = logit(TPF) - logit(FPF) = ln\left(\frac{TPF/_{(1-TPF)}}{FPF/_{(1-FPF)}}\right)$, which is the log of Diagnostic Odds Ratio, or DOR, [2] for an average subject and reader. Exclusion of ‘0’, from the 95% confidence interval for η can be used for testing equality between TPF and FPF, or equivalently non-informativeness of the binary result for discriminating between diseased and non-diseased subjects. Ideally, we would like DOR>>1 since for a reasonable test sensitivity > 1-specificity.

2.3.3 Model C: Assessment-level binary covariate (e.g., comparison of sensitivity between two modalities)

In multi-reader studies, a typical covariate (“assessment-level”) that distinguishes evaluations of subject by a reader is the reading conditions, or “diagnostic modality”. Inference based on this covariate can be used to compare performance levels (e.g., TPF, FPF, DOR) under different modalities. Here we consider a task of comparing TPF levels of two diagnostic modalities which can be performed using the following GLMM:

$logit\left(p_{ijM}\right) = \mu + \delta_M + \alpha_i + \beta_j + \gamma_{jM},$

where $Y_{ijM}|\alpha_i, \beta_j, \gamma_{jM} \sim Bin\left(1, p_{ijM}\right)$, $i = 1, \dots, n_1$ is the index representing diseased subjects, $j = 1, \dots, n_r$ is the reader index, δ_M is the fixed effect of modality M $(M = -1{,}1)$, $\alpha_i \sim N(0, \sigma_\alpha^2)$ is the subject random effect, $\beta_j \sim N\left(0, \sigma_\beta^2\right)$ is the reader random effect, $\gamma_{jM} \sim N(0, \sigma_M^2)$ is the random effect of the interaction between modality and reader. This random interaction term models the variability in the difference between the sensitivity of the two modalities across readers via σ_M^2.

The primary inferential target for this model is the coefficient $\delta = \delta_1 - \delta_{-1} = logit(TPF_{M=1} = p_1) - logit(TPF_{M=-1} = p_2) = ln\left(\frac{TPF_{M=1}/_{(1-TPF_{M=1})}}{TPF_{M=-1}/_{(1-TPF_{M=-1})}}\right)$, which is the log of the odds ratio of the probability that an average diseased subject is correctly diagnosed by an average reader under two modalities. Exclusion of ‘0’, from the 95% confidence interval for δ can be used for testing difference in sensitivity levels between diagnostic modalities.

2.3.4 Model D: Subject-level continuous covariate (e.g., lesion size effect on sensitivity)

Similar to binary covariates, continuous factors can be related to subjects, readers, or “assessments”. In general, inferences on a continuous covariate in GLMM models are typically more stable than inferences on a binary covariate. However, while a single binary covariate can be easy to address in a model-free setting e.g., using contingency tables, non-parametric approaches [48], a continuous covariate requires modeling. Here we consider a model for a continuous subject-level covariate for “diseased” subjects.

For estimating the effect of a continuous subject-level covariate (e.g., lesion size) on TPF, we can use the following GLMM:

$logit\left(p_{ij}\right) = \mu + \tau * X_i + \alpha_i + \beta_j + \gamma_j * X_i,$

where $Y_{ij}|X_i, \alpha_i, \beta_j, \gamma_j \sim Bin\left(1, p_{ij}\right)$, $i = 1, \dots, n_1$ is the index representing diseased subjects, $j = 1, \dots, n_r$ is the reader index, τ is the fixed-effect for slope where X_i is the subject lesion size (mm), $\alpha_i \sim N(0, \sigma_\alpha^2)$ is the subject random effect, $\beta_j \sim N\left(0, \sigma_\beta^2\right)$ is the reader random effect, $\gamma_j \sim N(0, \sigma_X^2)$ is the random effect of the interaction between reader and lesion size which models variability between reader-specific slopes.

The parameter of interest is the slope $\tau = logit(TPF_{X=x+1}) - logit(TPF_{X=x}) = ln\left(\frac{TPF_{X=x+1}/(1-TPF_{X=x+1})}{TPF_{X=x}/(1-TPF_{X=x})}\right)$ which is the log of the odds ratio of "true positive" ($Y = 1$) by an average reader for an average diseased subject with a 1 mm increment in lesion size X_i. $TPF_{X=x}$ represents the TPF when $X = x$ whereas $TPF_{X=x+1}$ is the TPF when $X = x + 1$. Exclusion of '0' from the 95% confidence interval for τ can be used for testing the difference in the change in sensitivity as a function of lesion size.

Table 1 Subject-Specific Analysis Models

Model	**Scenario**	**Sampling Populations**	**Fixed Effects**	**Random Effects**	**Estimation Parameters**	**Primary Estimation Target**
A (Covariate free)	Estimating Sensitivity	- Diseased subjects - Readers	μ or $p = logit^{-1}(\mu)$	α_i, β_j	$\mu, \sigma_\alpha^2, \sigma_\beta^2$	$\mu = ln\left({}^{TPF}/_{(1-TPF)}\right)$
B (Subject level binary covariate)	Estimating Diagnostic Odds ratio	- Diseased subjects - Non-diseased subjects - Readers	μ, η_D	$\alpha_i, \beta_j, \gamma_{jD}$	$\mu, \eta_D, \sigma_\alpha^2, \sigma_\beta^2, \sigma_D^2$	$\eta = \eta_1 - \eta_{-1} = ln\left(\frac{{}^{TPF}/_{(1-TPF)}}{{}^{FPF}/_{(1-FPF)}}\right)$
C (Assessment level binary covariate)	Comparing Sensitivity levels for two diagnostic modalities	- Diseased subjects - Readers	μ, δ_M	$\alpha_i, \beta_j, \gamma_{jM}$	$\mu, \delta_M, \sigma_\alpha^2, \sigma_\beta^2, \sigma_M^2$	$\delta = \delta_1 - \delta_{-1}$ $= ln\left(\frac{{}^{TPF_{M=1}}/_{(1-TPF_{M=1})}}{{}^{TPF_{M=-1}}/_{(1-TPF_{M=-1})}}\right)$
D (Subject level continuous covariate)	Estimating effect of a continuous covariate (lesion size) on sensitivity	- Diseased subjects - Readers	μ, τ	$\alpha_i, \beta_j, \gamma_j$	$\mu, \tau, \sigma_\alpha^2, \sigma_\beta^2, \sigma_X^2$	$\tau = ln\left(\frac{{}^{TPF_{X=x+1}}/_{(1-TPF_{X=x+1})}}{{}^{TPF_{X=x}}/_{(1-TPF_{X=x})}}\right)$

2.4 SIMULATION STUDY

2.4.1 Simulation Study Parameters

For simulated datasets we considered sample sizes similar to those in real diagnostic imaging study by Slasky *et al.* [43] and other multi-reader studies of diagnostic imaging. This particular study included confidence ratings (on 0-4 scale) regarding the presence of lung nodules provided by 7 radiologists for conventional and digital images acquired from the same set of 175 examinations without and 55 examinations with known lung nodules (*www.roc.pitt.edu*). The original study was conducted under a fully-crossed design where each radiologist evaluated images of all subjects under all viewing modalities. For our investigations we dichotomized responses (i.e., "positive" if rating > 2, "negative" if rating < 2).

Each simulation scenario was determined by the numbers of subjects, readers, and specific true values for parameters of the considered subject-specific models. For each scenario we generated the cross-correlated data, starting with generating random effects (for subjects, readers, and considered interaction terms) and then generating binary responses with probabilities determined by the random effects and fixed parameters. The parameter configurations for models A, B and C were based on estimates obtained from GLMM models fit to real imaging study data [Table 2] with small alterations to make the scenarios more comparable across the considered model settings. For continuous covariate model D, we assigned true values for parameters based on our experience with prior analyses of detection accuracy for breast lesions in fully crossed multi-reader studies.

When planning and designing future studies or even simulation studies, there is often the need to understand and interpret the magnitude of the variance components (i.e., what is large, small, reasonable?). This specific task in the cross-correlated studies is not straightforward. To address this question and as an example, we developed a simple tool for interpreting magnitude of the variance components for the simplest no-covariate model A [0]. For example the between-reader variability can be interpreted in terms of the smallest and largest sensitivity levels that are likely to be observed. The between-subject variability can be interpreted in terms of the smallest and largest proportions of readers who label the subject “positive”. These desired quantities can be computed through numerical integration by using estimated GLMM parameters. Their correspondence can also be additionally checked by computing the same probabilities empirically using data from the multi-reader diagnostic imaging study [43].

We also used a similar tool to obtain reasonable values of variance components by fixing the fixed-effect values in the context of model D involving a continuous covariate [0].

These tools can help design new scenarios with practically reasonable structure of the variance components which can eventually help understand the generality of the phenomena and report findings across a range of values.

Table 2 Model estimates from the actual multi-reader study

Model	**Sample Size**	**Fixed effect (FE)**	**PL Estimation** *Variance Estimates*	**PL Estimation** *FE Estimate ± SE (95% t-based CI)*	**LA Estimation** *Variance Estimates*	**LA Estimation** *FE Estimate ± SE (95% t-based CI)*	**PL+LA Estimation** *FE Estimate ± SE (95% t-based CI)*
A (Modality=1)	n_1=55 n_r=7	$logit(TPF) = \mu$	$\hat{\sigma}_\alpha^2$= 2.30 $\hat{\sigma}_\beta^2$= 0.71	0.11 ± 0.40 (-0.86, 1.09)	$\hat{\sigma}_\alpha^2$ = 3.56 $\hat{\sigma}_\beta^2$ = 0.90	0.15 ± 0.46 (-0.98, 1.28)	0.15 ± 0.46 (-0.98, 1.28)
A (Modality=2)	n_1=55 n_r=7	$logit(TPF) = \mu$	$\hat{\sigma}_\alpha^2$ = 3.51 $\hat{\sigma}_\beta^2$ = 0.62	0.51 ± 0.41 (-0.50, 1.53)	$\hat{\sigma}_\alpha^2$ = 6.25 $\hat{\sigma}_\beta^2$ = 0.86	0.74 ± 0.52 (-0.54, 2.02)	0.74 ± 0.52 (-0.54, 2.02)
B (Modality=1)	n_0=175 n_1=55 n_r=7	$\eta = \eta_1 - \eta_{-1}$	$\hat{\sigma}_\alpha^2$ = 1.96 $\hat{\sigma}_\beta^2$= 0.54 $\hat{\sigma}_D^2$ = 0.14	3.33 ± 0.35 (2.45, 4.20)	$\hat{\sigma}_\alpha^2$ = 3.93 $\hat{\sigma}_\beta^2$ = 0.57 $\hat{\sigma}_D^2$= 0.11	4.53 ± 0.51 (3.27, 5.79)	4.53 ± 0.51 (3.27, 5.79)
		$logit(TPF)$		0.10 ± 0.38 (-0.83, 1.05)		0.16 ± 0.43 (-0.90, 1.24)	0.16 ± 0.43 (-0.90, 1.24)
		$logit(FPF)$		-3.22 ± 0.36 (-4.11, -2.32)		-4.36 ± 0.47 (-5.53, -3.19)	-4.36 ± 0.47 (-5.53, -3.19)
B (Modality=2)	n_0=175 n_1=55 n_r=7	$\eta = \eta_1 - \eta_{-1}$	$\hat{\sigma}_\alpha^2$ = 2.02 $\hat{\sigma}_\beta^2$ = 0.64 $\hat{\sigma}_D^2$ = 0.02	3.36 ± 0.29 (2.63, 4.10)	$\hat{\sigma}_\alpha^2$ = 3.78 $\hat{\sigma}_\beta^2$ = 0.69 $\hat{\sigma}_D^2$= 0	4.47 ± 0 (. , .)	4.47 ± 0.29 (3.74, 5.20)
		$logit(TPF)$		0.47 ± 0.38 (-0.47, 1.41)		0.65 ± 0.01 (0.63, 0.67)	0.65 ± 0.38 (-0.29, 1.59)
		$logit(FPF)$		-2.89 ± 0.35 (-3.75, -2.03)		-3.28 ± 0.01 (-3.84, -3.79)	-3.28 ± 0.35 (-4.14, -2.41)
C	n_1=55 n_r=7	$\delta = \delta_1 - \delta_{-1}$	$\hat{\sigma}_\alpha^2$= 3.71 $\hat{\sigma}_\beta^2$ = 0.86 $\hat{\sigma}_M^2$ = 0 (G-matrix not p.d.)	-0.42 ± 0.18 (-0.88, 0.03)	$\hat{\sigma}_\alpha^2$ = 5.05 $\hat{\sigma}_\beta^2$ = 0.95 $\hat{\sigma}_M^2$ = 0 (G-matrix not p.d.)	-0.47 ± 0.11 (-0.76, -0.18)	-0.47 ± 0.18 (-0.93, -0.01)
		$logit(TPF_{M=1})$		0.13 ± 0.45 (-0.99, 1.25)		0.18 ± 0.11 (-0.09, 0.47)	0.18 ± 0.45 (-0.93, 1.31)
		$logit(TPF_{M=-1})$		0.55 ± 0.46 (-0.56, 1.68)		0.66 ± 0 (. , .)	0.66 ± 0.46 (-0.46, 1.78)

1. p.d.= positive definite
2. All estimates are computed on logit scale
3. G-matrix is variance-covariance matrix of random effects
4. t-based CI use default containment degrees of freedom
5. PL+LA Estimation: Combination model with fixed effect estimates from Laplace technique; Standard Error (SE) estimates of LA model replaced by those of PL model only when G-matrix is not p.d.. In case G-matrix is p.d. for both models, the greater of the two SE is utilized

2.4.2 Simulation Study Details

As mentioned earlier, in the simulation studies, we considered the performance of all four models described in Section 2.3 using both the PL, LA estimation and combination techniques. Simulations were carried out using SAS using N=1,000 Monte Carlo independent simulated datasets for each parameter configuration. The models were fitted with PROC GLIMMIX, SAS using the specification statements provided in Appendix A.

Specifically, for each estimated model, we acquired estimates of the targeted fixed effects, their estimated standard errors, limits for the 95% confidence intervals, and the number of Monte Carlo simulations where the convergence was achieved (with and without positive definite estimates of the covariance matrix).

Simulation parameters were used as the reference for estimating coverage of the confidence intervals and bias of the fixed effect estimates of all models. We focused on evaluating the coverage of the 95% CI ("Coverage (%)") since it plays a central role in statistical inferences. It was estimated as the proportion of the times the true parameter (for concreteness denoted here as θ) of the corresponding model was contained within the estimated CI. The CI were based on the default t-based reference distribution with containment degrees of freedom unless stated otherwise.

We also estimated the following quantities which helped us gain insight in some of the observed trends and how they affected coverage rate:

a) To assess the performance of the variance estimators, we approximated the empirical standard deviation (MC SD or Monte Carlo Standard Deviation) calculated from 1,000 (sometimes 2000 to obtain a more precise estimate) MC trials per simulation configuration.

This is used as the gold standard or the supposedly true population standard deviation in this assessment. $MC\ SD = \widehat{SD}\left(\hat{\theta}\right) = \sqrt{\left[\frac{1}{N-1}\right]\sum_{n=1}^{N=100}\left(\hat{\theta}_n - \bar{\hat{\theta}}\right)^2}$.

b) Standardized bias of fixed effect estimate i.e. $SB = \frac{\frac{1}{N}\sum_{n=1}^{N=1000}(\hat{\theta}_n - \theta)}{\widehat{SD}\left(\hat{\theta}\right)}$ based on MC SD. This quantity tells us how big of a difference there is between the true parameter of interest and the average of the parameter estimates relative to the MC standard deviation of that parameter.

c) Relative bias of the standard error estimate (where bias is given by the difference between theoretical and empirical Monte Carlo standard deviation) i.e. $RBS = 100 * \left\{\left(\frac{1}{N}\sum_{n=1}^{N=1000}\widehat{SE}\left(\hat{\theta}_n\right)\right) - \widehat{SD}\left(\hat{\theta}\right)\right\}/\widehat{SD}\left(\hat{\theta}\right)$.

d) $Bias = \frac{\frac{1}{N}\sum_{n=1}^{N=1000}(\hat{\theta}_n - \theta)}{N}$ which is the bias of fixed-effect parameter estimate.

e) $SE = \frac{\sum_{n=1}^{N=1000}\widehat{SE}\{\hat{\theta}_n\}}{N}$ is the average estimated standard error for the fixed-effect parameter estimates.

f) $Est = \frac{\sum_{n=1}^{N=1000}\hat{\theta}_n}{N} = \bar{\hat{\theta}}$ which is the average of fixed-effect parameter estimates across simulations.

We also showed the number of Monte Carlo simulations ("Sim Used") used for computation of the quantities listed above. Specifically, we discarded all simulations where 1) standard error estimate was zero and 2) convergence was not achieved (based on PROC GLIMMIX "Convergence Status" OUTPUT). For the continuous covariate Laplace estimated model, we also discarded simulations where the conditional log likelihood was zero since that meant that the MLE did not exist (which could be seen in terms of huge estimates of fixed effect and corresponding SE).

In addition to considering the PL and LA estimation approaches on all convergent simulations with non-zero estimates of SE of fixed effect, we also separately considered instances with positive-definite estimates of the covariance matrix of random effects (G matrix). This was done because non-positive definite G-matrices often result in standard errors estimates of fixed effect which are either zero or non-reliable (perhaps smaller than usual).

Quality of confidence intervals, estimation of fixed effects and their standard errors are only indirectly but related to estimation of variance components. These quantities are also important for planning future studies. For illustrative purposes, we partially investigated the quality of variance component estimation for the simple no-covariate model A and compared them between PL and LA approaches [Appendix C]. This was done by computing the relative bias (%) of the estimated variance components by comparing against their true simulation values.

2.4.3 Simulation Study Results

For the covariate free setting (model A, Table 3, Table 4, Table 5), bias in fixed effect estimates was rather substantial for the PL (often larger than 1 standard error) and negligible for LA approach (mostly less than 0.06 standard errors). This observation was in concordance with the relative performance of these approaches in the simpler setting of binary clustered data [31]. For the PL approach, the bias worsened when the true sensitivity was far from 0.5 (e.g., TPF=0.1). The standard error of the fixed effect estimates was substantially underestimated under the LA method (by as much as -20%), but rather accurate for the PL approach. Interestingly, despite the substantial underestimation of the variance, the estimated 95% CI coverage rate was somewhat conservative for the LA estimated model in the considered scenarios. This agrees with previously reported properties of the LA estimates [25]. One possible explanation for this is that the distribution of the fixed-effect estimator looks like normal for most part but has a thin tail due to extreme observations in many settings. These extreme observations tend to drive up the true variance (or MC SD) leading to underestimation of the variance estimates. For the PL approach, the CI coverage was substantially lower than nominal level for scenarios with substantial bias in the fixed effect estimates (e.g., TPF=0.1). In simulated datasets resulting in positive definite G-matrix under the LA approach, the results were similar. Combination approaches based on the LA estimation (reference distribution, and variance borrowing summarized in Table 15) led to statistical inferences of the same quality.

Table 3 Simulation Results for Model A (small magnitude of variance components)

		PL Estimation*						LAPLACE Estimation*						LAPLACE Estimation**					
		p=0.1 μ=-2.20		p=0.5 μ=0		p=0.7 μ=0.85		p=0.1 μ=-2.20		p=0.5 μ=0		p=0.7 μ=0.85		p=0.1 μ=-2.20		p=0.5 μ=0		p=0.7 μ=0.85	
		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1	
	n_r	55	100	55	100	55	100	55	100	55	100	55	100	55	100	55	100	55	100
Coverage (%)	**5**	92	86	99	98	97	97	94	91	98	98	98	98	99	98	99	99	99	98
	10	84	77	97	97	96	94	89	90	96	96	96	97	97	97	97	97	97	97
Bias (SE)	**5**	0.257 (0.256)	0.265 (0.2)	0.004 (0.204)	0.005 (0.169)	-0.097 (0.212)	-0.101 (0.174)	0.002 (0.332)	0.007 (0.234)	0.006 (0.221)	0.007 (0.179)	-0.005 (0.23)	-0.007 (0.185)	-0.029 (0.369)	0.014 (0.266)	0.009 (0.232)	0.008 (0.188)	-0.005 (0.247)	-0.008 (0.197)
	10	0.215 (0.199)	0.22 (0.158)	0 (0.174)	0.003 (0.143)	-0.073 (0.177)	-0.073 (0.144)	-0.004 (0.227)	0.008 (0.177)	0.002 (0.187)	0.004 (0.153)	0.001 (0.192)	0.001 (0.155)	-0.006 (0.252)	0.008 (0.193)	0.003 (0.189)	0.003 (0.153)	0.001 (0.197)	0 (0.157)
SB (RBS (%))	**5**	1 (-1)	1.34 (1)	0.02 (-1)	0.03 (-3)	-0.45 (-3)	-0.56 (-3)	0.01 (-11)	0.03 (-14)	0.02 (-5)	0.04 (-8)	-0.02 (-9)	-0.04 (-10)	-0.08 (-1)	0.05 (-2)	0.04 (0)	0.04 (-3)	-0.02 (-2)	-0.04 (-4)
	10	1.03 (-4)	1.41 (1)	0 (0)	0.02 (0)	-0.42 (2)	-0.5 (0)	-0.01 (-14)	0.04 (-10)	0.01 (-1)	0.02 (-2)	0.01 (-1)	0.01 (-3)	-0.02 (-5)	0.04 (-1)	0.02 (0)	0.02 (-2)	0 (2)	0 (-2)
Sim Used	**5**	1000	1000	1000	1000	1000	1000	861	854	997	994	1000	1000	535	601	661	784	654	765
	10	1000	1000	1000	1000	1000	1000	885	907	992	997	999	1000	725	803	870	946	829	923

1. Simulation parameters: Subject Variance=1, Reader Variance=0.1
2. * : Sim Used=Used simulations: (1) Non-zero positive standard errors in convergent simulations (2) PROC GLIMMIX convergence criteria satisfied
3. ** : Sim Used=Used simulations: (1) PROC GLIMMIX convergence criteria satisfied (2) Positive-definite G matrix

Table 4 Simulation Results for Model A (medium magnitude of variance components)

		PL Estimation*						LAPLACE Estimation*						LAPLACE Estimation**					
		p=0.1		p=0.5		p=0.7		p=0.1		p=0.5		p=0.7		p=0.1		p=0.5		p=0.7	
		μ=-2.20		μ=0		μ=0.85		μ=-2.20		μ=0		μ=0.85		μ=-2.20		μ=0		μ=0.85	
		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1	
	n_r	55	100	55	100	55	100	55	100	55	100	55	100	55	100	55	100	55	100
Coverage (%)	5	89	87	97	97	96	95	95	95	97	96	97	96	97	97	98	96	98	96
	10	86	84	97	97	96	95	96	96	96	96	97	96	96	96	96	96	97	96
Bias	5	0.408	0.409	-0.01	0.003	-0.154	-0.149	-0.038	0.018	-0.009	0.004	-0.005	-0.008	-0.033	0.028	-0.008	0.002	-0.009	-0.007
(SE)		(0.382)	(0.344)	(0.357)	(0.33)	(0.356)	(0.333)	(0.493)	(0.398)	(0.396)	(0.358)	(0.4)	(0.361)	(0.512)	(0.406)	(0.402)	(0.361)	(0.407)	(0.366)
	10	0.331	0.325	0.002	0.007	-0.107	-0.099	0.004	0.02	0.003	0.008	0.005	0.007	0.005	0.02	0.002	0.008	0.005	0.007
		(0.305)	(0.275)	(0.301)	(0.273)	(0.3)	(0.272)	(0.366)	(0.312)	(0.33)	(0.294)	(0.333)	(0.296)	(0.368)	(0.312)	(0.33)	(0.294)	(0.333)	(0.296)
SB	5	1.07	1.17	-0.03	0.01	-0.43	-0.44	-0.06	0.04	-0.02	0.01	-0.01	-0.02	-0.05	0.06	-0.02	0	-0.02	-0.02
(RBS (%))		(0)	(-2)	(0)	(-3)	(-1)	(-2)	(-21)	(-16)	(-11)	(-14)	(-11)	(-14)	(-18)	(-14)	(-10)	(-13)	(-9)	(-13)
	10	1.12	1.24	0.01	0.03	-0.37	-0.37	0.01	0.06	0.01	0.03	0.01	0.02	0.01	0.06	0.01	0.03	0.01	0.02
		(3)	(5)	(1)	(3)	(4)	(3)	(-6)	(-9)	(-5)	(-7)	(-5)	(-9)	(-6)	(-9)	(-5)	(-7)	(-5)	(-9)
Sim Used	5	1000	1000	1000	1000	1000	1000	953	970	996	998	1000	1000	901	944	960	984	953	978
	10	1000	1000	1000	1000	1000	1000	998	1000	999	1000	1000	1000	989	1000	998	1000	998	999

1. Simulation parameters: Subject Variance=2, Reader Variance=0.7
2. * : Sim Used=Used simulations: (1) Non-zero positive standard errors in convergent simulations (2) PROC GLIMMIX convergence criteria satisfied
3. ** : Sim Used=Used simulations: (1) PROC GLIMMIX convergence criteria satisfied (2) Positive-definite G matrix

Table 5 Simulation Results for Model A (large magnitude of variance components)

		PL Estimation*						LAPLACE Estimation*						LAPLACE Estimation**					
		p=0.1 μ=-2.20		p=0.5 μ=0		p=0.7 μ=0.85		p=0.1 μ=-2.20		p=0.5 μ=0		p=0.7 μ=0.85		p=0.1 μ=-2.20		p=0.5 μ=0		p=0.7 μ=0.85	
		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1	
	n_r	55	100	55	100	55	100	55	100	55	100	55	100	55	100	55	100	55	100
Coverage (%)	**5**	91	90	97	97	96	96	96	97	96	97	96	96	97	97	96	97	96	96
	10	89	87	97	97	96	96	96	96	96	96	96	96	96	96	96	96	96	96
Bias (SE)	**5**	0.49 (0.489)	0.5 (0.457)	-0.018 (0.476)	-0.017 (0.45)	-0.199 (0.478)	-0.194 (0.451)	-0.11 (0.653)	0.027 (0.532)	-0.021 (0.547)	-0.023 (0.503)	-0.017 (0.555)	-0.023 (0.506)	-0.107 (0.663)	0.033 (0.537)	-0.024 (0.549)	-0.023 (0.504)	-0.018 (0.558)	-0.025 (0.508)
	10	0.374 (0.4)	0.367 (0.372)	-0.014 (0.4)	-0.01 (0.373)	-0.143 (0.401)	-0.137 (0.374)	-0.027 (0.484)	0 (0.425)	-0.016 (0.447)	-0.012 (0.409)	-0.011 (0.452)	-0.015 (0.412)	-0.027 (0.484)	0 (0.425)	-0.016 (0.447)	-0.012 (0.409)	-0.012 (0.452)	-0.015 (0.412)
SB (RBS (%))	**5**	0.99 (-1)	1.12 (2)	-0.04 (-4)	-0.04 (-2)	-0.4 (-3)	-0.42 (-2)	-0.12 (-30)	0.04 (-21)	-0.04 (-10)	-0.04 (-11)	-0.03 (-12)	-0.04 (-12)	-0.11 (-29)	0.05 (-21)	-0.04 (-9)	-0.04 (-11)	-0.03 (-12)	-0.04 (-12)
	10	0.98 (5)	1.07 (8)	-0.04 (2)	-0.03 (0)	-0.36 (0)	-0.37 (2)	-0.05 (-7)	0 (-7)	-0.04 (-4)	-0.03 (-6)	-0.02 (-4)	-0.03 (-5)	-0.05 (-7)	0 (-7)	-0.04 (-4)	-0.03 (-6)	-0.02 (-4)	-0.03 (-5)
Sim Used	**5**	1000	1000	1000	1000	1000	1000	986	994	999	1000	1000	1000	968	984	994	997	988	992
	10	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	999	1000

1. Simulation parameters: Subject Variance=3, Reader Variance=1.5
2. * : Sim Used=Used simulations: (1) Non-zero positive standard errors in convergent simulations (2) PROC GLIMMIX convergence criteria satisfied
3. ** : Sim Used=Used simulations: (1) PROC GLIMMIX convergence criteria satisfied (2) Positive-definite G matrix

Results for the setting with a subject level binary covariate (model B, Table 6, Table 7, Table 8) also indicated a frequently severe bias of the fixed effects estimates under the PL approach and a negligible bias under the LA method. The relative bias of the standard error estimate was again poor for the LA approach, especially for a small number of readers, while being relatively accurate for the PL approach. As before, the substantial bias in fixed effects of the PL model led to substantial under-coverage of CIs (as low as 57%). For this model, the LA-based CIs also demonstrated under-coverage in scenarios with few reader and extreme TPF or extreme FPF values. In contrast, in the subset of simulations with positive-definite estimates of the covariance matrix, the LA-based CIs were conservative, thereby enabling appropriate, albeit possibly under-powered statistical inferences.

The LA estimation approach very frequently had fitting problems leading to up to 34% of scenarios where the statistical inferences were not possible (mostly due to problems estimating the variance of the fixed effect). At the same time the PL approach had only a few convergence problems across all simulations in the considered scenarios.

Previously recommended combination approach based on borrowing Satterthwaite approximation from the PL approach for LA-based estimates did not lead to any improvement in coverage and even resulted in the decreased coverage in some scenarios. However, borrowing the PL estimates of variability in scenarios only where the LA estimates were not usable i.e. resulting from a non-positive definite G-matrix, resulted in substantial improvements in the coverage of confidence intervals, while enabling statistical inferences in almost all instances. Results are shown in Table 15.

Table 6 Simulation Results for Model B (small magnitude of variance components)

			PL Estimation*						**LAPLACE Estimation***						**LAPLACE Estimation****					
			TPF=0.5 FPF=0.03 η=3.48		**TPF=0.3 FPF=0.9 η=-3.04**		**TPF=0.6 FPF=0.6 η=0**		**TPF=0.5 FPF=0.03 η=3.48**		**TPF=0.3 FPF=0.9 η=-3.04**		**TPF=0.6 FPF=0.6 η=0**		**TPF=0.5 FPF=0.03 η=3.48**		**TPF=0.3 FPF=0.9 η=-3.04**		**TPF=0.6 FPF=0.6 η=0**	
			n_0		**n_0**		**n_0**		**n_0**		**n_0**		**n_0**		**n_0**		**n_0**		**n_0**	
	n_r	**n_1**	**100**	**175**	**100**	**175**	**100**	**175**	**100**	**175**	**100**	**175**	**100**	**175**	**100**	**175**	**100**	**175**	**100**	**175**
Coverage (%)	**5**	**55**	91	88	89	84	99	99	81	82	85	80	95	95	99	100	100	98	99	99
	5	**100**	90	87	84	80	98	99	81	80	85	83	94	95	99	98	99	98	99	99
	10	**55**	79	73	81	79	98	98	88	89	95	95	96	96	97	97	98	98	98	97
	10	**100**	77	67	76	73	98	98	89	89	95	94	96	97	98	98	98	97	98	98
Bias	**5**	**55**	-0.374	-0.386	0.367	0.374	0.003	0.004	0.028	0.007	0.016	0.05	0.005	0.014	0.01	0.003	0.012	0.054	0.007	0.014
(SE)			(0.323)	(0.281)	(0.275)	(0.253)	(0.253)	(0.239)	(0.321)	(0.278)	(0.294)	(0.247)	(0.271)	(0.249)	(0.406)	(0.355)	(0.344)	(0.307)	(0.285)	(0.269)
	5	**100**	-0.373	-0.381	0.374	0.374	0	0.007	0.014	0.008	0.021	0.052	0.002	0.006	-0.015	-0.025	0.037	0.047	0.003	0.014
			(0.304)	(0.255)	(0.249)	(0.224)	(0.226)	(0.21)	(0.284)	(0.244)	(0.254)	(0.225)	(0.255)	(0.218)	(0.368)	(0.306)	(0.301)	(0.265)	(0.254)	(0.23)
	10	**55**	-0.355	-0.359	0.301	0.299	0.007	0.003	0.012	0.004	0.023	0.013	0.006	0.004	0.007	0.002	0.02	0.011	0.008	0.002
			(0.256)	(0.225)	(0.227)	(0.21)	(0.217)	(0.205)	(0.276)	(0.253)	(0.254)	(0.232)	(0.235)	(0.219)	(0.307)	(0.268)	(0.264)	(0.24)	(0.237)	(0.223)
	10	**100**	-0.349	-0.353	0.296	0.299	0.006	0.005	0.015	0	0.012	0.019	0.008	0.003	0.01	0.001	0.015	0.017	0.008	0.003
			(0.237)	(0.203)	(0.204)	(0.185)	(0.19)	(0.176)	(0.251)	(0.213)	(0.226)	(0.2)	(0.203)	(0.187)	(0.276)	(0.234)	(0.233)	(0.207)	(0.206)	(0.189)
SB	**5**	**55**	-1.09	-1.37	1.3	1.4	0.01	0.02	0.07	0.02	0.04	0.16	0.02	0.05	0.02	0.01	0.03	0.17	0.02	0.05
(RBS (%))			(-6)	(0)	(-2)	(-5)	(-5)	(-3)	(-22)	(-24)	(-20)	(-22)	(-7)	(-7)	(-1)	(-2)	(-6)	(-3)	(-2)	(1)
	5	**100**	-1.18	-1.44	1.44	1.54	0	0.03	0.04	0.03	0.07	0.19	0.01	0.03	-0.04	-0.08	0.11	0.17	0.01	0.06
			(-4)	(-3)	(-4)	(-8)	(-6)	(-3)	(-25)	(-25)	(-20)	(-20)	(-7)	(-10)	(-3)	(-6)	(-6)	(-6)	(-7)	(-5)
	10	**55**	-1.45	-1.61	1.33	1.42	0.03	0.02	0.04	0.02	0.09	0.06	0.03	0.02	0.02	0.01	0.08	0.05	0.03	0.01
			(4)	(1)	(1)	(-1)	(1)	(2)	(-4)	(-6)	(-1)	(-3)	(1)	(0)	(7)	(-1)	(2)	(1)	(2)	(1)
	10	**100**	-1.57	-1.77	1.49	1.64	0.03	0.03	0.06	0	0.05	0.09	0.04	0.02	0.04	0	0.06	0.08	0.04	0.02
			(6)	(2)	(3)	(2)	(2)	(4)	(-1)	(-8)	(-2)	(-3)	(0)	(1)	(8)	(1)	(1)	(1)	(1)	(3)
Sim Used	**5**	**55**	1000	1000	1000	1000	1000	1000	627	597	486	550	779	781	314	335	407	431	562	625
	5	**100**	1000	1000	1000	1000	1000	1000	623	645	561	605	796	845	358	406	472	510	635	733
	10	**55**	1000	1000	1000	1000	1000	1000	778	778	731	787	908	914	598	649	704	760	807	845
	10	**100**	1000	1000	1000	1000	1000	1000	795	824	778	878	946	955	646	707	753	848	888	917

1. Simulation parameters: Subject Variance=1, Reader Variance=0.25, Reader*Truth Variance=0.07
2. * : Sim Used=Used simulations: (1) Non-zero positive standard errors in convergent simulations (2) PROC GLIMMIX convergence criteria satisfied
3. ** : Sim Used=Used simulations: (1) PROC GLIMMIX convergence criteria satisfied (2) Positive-definite G matrix

Table 7 Simulation Results for Model B (medium magnitude of variance components)

			PL Estimation*						LAPLACE Estimation*						LAPLACE Estimation**					
			TPF=0.5 FPF=0.03 η=3.48		TPF=0.3 FPF=0.9 η=-3.04		TPF=0.6 FPF=0.6 η=0		TPF=0.5 FPF=0.03 η=3.48		TPF=0.3 FPF=0.9 η=-3.04		TPF=0.6 FPF=0.6 η=0		TPF=0.5 FPF=0.03 η=3.48		TPF=0.3 FPF=0.9 η=-3.04		TPF=0.6 FPF=0.6 η=0	
			n_0		n_0		n_0		n_0		n_0		n_0		n_0		n_0		n_0	
	n_r	n_1	100	175	100	175	100	175	100	175	100	175	100	175	100	175	100	175	100	175
Coverage (%)	5	55	78	75	82	79	99	98	87	87	90	91	97	97	99	98	99	99	99	99
	5	100	73	70	75	72	98	98	85	85	91	93	96	96	98	97	98	99	98	98
	10	55	66	60	76	77	97	98	92	94	95	96	96	98	96	97	97	97	96	98
	10	100	64	57	72	69	97	98	93	95	94	96	96	98	97	97	95	96	96	98
Bias (SE)	5	55	-0.629 (0.352)	-0.621 (0.321)	0.55 (0.321)	0.531 (0.304)	0.004 (0.308)	0.003 (0.293)	0.003 (0.399)	0.002 (0.362)	0.059 (0.372)	0.03 (0.341)	-0.003 (0.353)	0.001 (0.329)	-0.011 (0.472)	-0.021 (0.415)	0.05 (0.41)	0.027 (0.374)	-0.005 (0.363)	0.004 (0.343)
	5	100	-0.626 (0.326)	-0.61 (0.295)	0.547 (0.294)	0.537 (0.274)	-0.006 (0.278)	0.002 (0.263)	-0.029 (0.348)	0 (0.317)	0.039 (0.336)	0.045 (0.304)	-0.017 (0.31)	-0.001 (0.291)	-0.059 (0.415)	-0.022 (0.365)	0.035 (0.362)	0.048 (0.324)	-0.022 (0.322)	-0.006 (0.299)
	10	55	-0.517 (0.295)	-0.527 (0.268)	0.407 (0.28)	0.397 (0.261)	-0.002 (0.275)	-0.01 (0.262)	0.015 (0.345)	-0.018 (0.318)	0.012 (0.326)	0.012 (0.3)	-0.002 (0.309)	-0.01 (0.291)	0.009 (0.366)	-0.016 (0.326)	0.012 (0.331)	0.011 (0.303)	-0.003 (0.306)	-0.01 (0.291)
	10	100	-0.511 (0.272)	-0.51 (0.244)	0.401 (0.252)	0.395 (0.231)	-0.005 (0.245)	-0.005 (0.229)	0.008 (0.312)	-0.008 (0.283)	0.008 (0.29)	0.003 (0.262)	-0.01 (0.271)	-0.004 (0.25)	0.006 (0.327)	-0.012 (0.287)	0.007 (0.292)	0.003 (0.264)	-0.009 (0.271)	-0.005 (0.251)
SB (RBS (%))	5	55	-1.77 (-1)	-1.89 (-3)	1.69 (-1)	1.76 (1)	0.01 (-4)	0.01 (-2)	0.01 (-19)	0 (-19)	0.14 (-12)	0.08 (-12)	-0.01 (-8)	0 (-6)	-0.02 (-4)	-0.05 (-8)	0.12 (-3)	0.07 (-3)	-0.01 (-5)	0.01 (-2)
	5	100	-1.89 (-1)	-2 (-3)	1.83 (-2)	1.94 (-1)	-0.02 (-5)	0.01 (-1)	-0.07 (-22)	0 (-23)	0.1 (-16)	0.13 (-13)	-0.05 (-11)	0 (-8)	-0.13 (-7)	-0.05 (-12)	0.09 (-9)	0.14 (-7)	-0.06 (-8)	-0.02 (-5)
	10	55	-1.72 (-2)	-1.95 (-1)	1.44 (-1)	1.53 (1)	-0.01 (-2)	-0.04 (1)	0.04 (-10)	-0.05 (-6)	0.04 (-4)	0.04 (-1)	-0.01 (-2)	-0.04 (-1)	0.02 (-5)	-0.05 (-3)	0.03 (-3)	0.04 (1)	-0.01 (-3)	-0.03 (-1)
	10	100	-1.91 (2)	-2.06 (-2)	1.58 (-1)	1.69 (-1)	-0.02 (-2)	-0.02 (1)	0.02 (-6)	-0.03 (-7)	0.02 (-6)	0.01 (-6)	-0.04 (-4)	-0.02 (-2)	0.02 (-2)	-0.04 (-6)	0.02 (-5)	0.01 (-5)	-0.03 (-4)	-0.02 (-2)
Sim Used	5	55	1000	1000	1000	1000	1000	1000	662	701	652	709	850	835	457	525	577	637	714	753
	5	100	1000	1000	1000	1000	1000	1000	670	722	726	775	881	897	494	569	660	723	791	830
	10	55	1000	1000	1000	1000	1000	1000	861	882	895	912	959	970	769	820	881	900	924	933
	10	100	1000	994	1000	1000	1000	1000	877	921	931	950	977	984	822	886	925	944	962	976

1. Simulation parameters: Subject Variance=1.9, Reader Variance=0.5, Reader*Truth Variance=0.14
2. * : Sim Used=Used simulations: (1) Non-zero positive standard errors in convergent simulations (2) PROC GLIMMIX convergence criteria satisfied
3. ** : Sim Used=Used simulations: (1) PROC GLIMMIX convergence criteria satisfied (2) Positive-definite G matrix

Table 8 Simulation Results for Model B (large magnitude of variance components)

			PL Estimation*						**LAPLACE Estimation***						**LAPLACE Estimation****					
			TPF=0.5 FPF=0.03 η=3.48		**TPF=0.3 FPF=0.9 η=-3.04**		**TPF=0.6 FPF=0.6 η=0**		**TPF=0.5 FPF=0.03 η=3.48**		**TPF=0.3 FPF=0.9 η=-3.04**		**TPF=0.6 FPF=0.6 η=0**		**TPF=0.5 FPF=0.03 η=3.48**		**TPF=0.3 FPF=0.9 η=-3.04**		**TPF=0.6 FPF=0.6 η=0**	
			n_0		**n_0**		**n_0**		**n_0**		**n_0**		**n_0**		**n_0**		**n_0**		**n_0**	
	n_r	**n_1**	**100**	**175**	**100**	**175**	**100**	**175**	**100**	**175**	**100**	**175**	**100**	**175**	**100**	**175**	**100**	**175**	**100**	**175**
Coverage (%)	**5**	**55**	76	76	82	80	97	98	91	93	95	94	95	96	97	97	97	97	97	97
	5	**100**	74	73	80	77	97	97	91	92	94	93	96	95	97	97	96	96	97	96
	10	**55**	75	69	81	80	98	97	96	97	97	96	98	96	97	97	97	96	98	97
	10	**100**	73	65	80	79	98	97	96	95	96	96	97	96	97	96	96	96	97	96
Bias	**5**	**55**	-0.803	-0.802	0.69	0.694	0	0	0	-0.012	0.058	0.075	-0.002	0.006	-0.028	-0.021	0.049	0.072	-0.007	0.007
(SE)			(0.451)	(0.428)	(0.439)	(0.424)	(0.431)	(0.418)	(0.555)	(0.507)	(0.53)	(0.49)	(0.497)	(0.478)	(0.583)	(0.531)	(0.544)	(0.503)	(0.508)	(0.484)
	5	**100**	-0.795	-0.788	0.689	0.689	0.005	0.008	-0.022	-0.025	0.066	0.067	0.008	0.016	-0.044	-0.037	0.06	0.07	0.008	0.018
			(0.433)	(0.406)	(0.415)	(0.399)	(0.408)	(0.395)	(0.499)	(0.449)	(0.483)	(0.444)	(0.465)	(0.442)	(0.531)	(0.476)	(0.493)	(0.457)	(0.469)	(0.447)
	10	**55**	-0.63	-0.632	0.503	0.495	0.007	0.004	0.027	-0.01	0.018	0.011	0.011	0.008	0.025	-0.01	0.017	0.012	0.009	0.007
			(0.394)	(0.37)	(0.387)	(0.371)	(0.389)	(0.375)	(0.475)	(0.438)	(0.452)	(0.425)	(0.438)	(0.418)	(0.478)	(0.438)	(0.453)	(0.426)	(0.437)	(0.419)
	10	**100**	-0.62	-0.62	0.495	0.492	0.006	0.009	0.016	-0.006	0.018	0.026	0.007	0.011	0.014	-0.006	0.018	0.025	0.007	0.01
			(0.371)	(0.346)	(0.361)	(0.343)	(0.36)	(0.343)	(0.43)	(0.395)	(0.413)	(0.384)	(0.399)	(0.378)	(0.435)	(0.397)	(0.413)	(0.385)	(0.4)	(0.379)
SB	**5**	**55**	-1.66	-1.78	1.5	1.59	0	0	0	-0.02	0.09	0.13	0	0.01	-0.04	-0.03	0.08	0.12	-0.01	0.01
(RBS (%))			(-7)	(-5)	(-5)	(-3)	(-7)	(-4)	(-21)	(-22)	(-16)	(-16)	(-14)	(-13)	(-17)	(-18)	(-14)	(-14)	(-12)	(-12)
	5	**100**	-1.74	-1.85	1.59	1.65	0.01	0.02	-0.03	-0.04	0.11	0.12	0.02	0.03	-0.07	-0.06	0.1	0.12	0.01	0.04
			(-5)	(-5)	(-4)	(-4)	(-6)	(-4)	(-21)	(-24)	(-17)	(-20)	(-14)	(-14)	(-16)	(-19)	(-16)	(-18)	(-13)	(-13)
	10	**55**	-1.69	-1.76	1.35	1.39	0.02	0.01	0.06	-0.02	0.04	0.03	0.02	0.02	0.05	-0.02	0.04	0.03	0.02	0.02
			(5)	(3)	(4)	(4)	(3)	(2)	(-2)	(-3)	(-2)	(-2)	(-1)	(-3)	(-2)	(-3)	(-2)	(-2)	(-1)	(-2)
	10	**100**	-1.77	-1.91	1.47	1.53	0.02	0.03	0.04	-0.01	0.04	0.07	0.02	0.03	0.03	-0.02	0.04	0.06	0.02	0.03
			(6)	(6)	(7)	(6)	(4)	(5)	(-4)	(-4)	(-2)	(-2)	(-1)	(0)	(-3)	(-4)	(-2)	(-2)	(-1)	(0)
Sim Used	**5**	**55**	1000	1000	1000	1000	1000	1000	837	840	834	865	903	913	724	752	802	830	831	863
	5	**100**	1000	999	1000	999	1000	1000	850	862	871	901	934	940	751	790	843	868	882	902
	10	**55**	1000	1000	1000	1000	1000	1000	971	972	969	986	982	990	956	957	963	983	976	982
	10	**100**	1000	996	1000	1000	1000	1000	973	982	976	982	983	990	959	972	975	980	975	983

1. Simulation parameters: Subject Variance=3, Reader Variance=1, Reader*Truth Variance=0.5
2. * : Sim Used=Used simulations: (1) Non-zero positive standard errors in convergent simulations (2) PROC GLIMMIX convergence criteria satisfied
3. ** : Sim Used=Used simulations: (1) PROC GLIMMIX convergence criteria satisfied (2) Positive-definite G matrix

For an assessment level covariate in model C, comparing the sensitivity between the two modalities (Table 9, Table 10, Table 11), we observed patterns similar to the previous setting. The PL approach yielded biased fixed effect estimates and inadequate confidence intervals when TPF were substantially different (e.g., $p_1 = 0.1$, $p_2 = 0.7$). The LA model resulted in negligible bias of the covariate's fixed effect across all scenarios, but lead to CIs with substantially lower than nominal coverage when the variance was substantially underestimated (more than by -20%). This tended to happen when the numbers of readers were small and TPF values were substantially different (e.g., $n_r = 5$, $p_1 = 0.1$, $p_2 = 0.7$). However, in the subset of simulations with positive-definite estimates of the covariance matrix, the LA based CIs had close to nominal coverage.

Again, the LA estimation approach very frequently had fitting problems leading up to 20% of scenarios where the statistical inferences were not possible (mostly due to problems estimating variance of the fixed effect). At the same time the PL approach had only a few convergence problems across all simulations in the considered scenarios.

The combination approach based on borrowing Satterthwaite approximation from the PL approach did not help to improve CI coverage in problematic scenarios. In contrast, borrowing the PL estimates of variability in scenarios where the LA estimates were not usable resulted in substantial improvements in the coverage of confidence intervals, while enabling statistical inferences in almost all instances. Results are shown in Table 15.

Table 9 Simulation Results for Model C (small magnitude of variance components)

		PL Estimation*						LAPLACE Estimation*						LAPLACE Estimation**					
		p_1=0.5 p_2=0.6 δ=-0.41		p_1=0.1 p_2=0.7 δ=-3.04		p_1=0.8 p_2=0.8 δ=0		p_1=0.5 p_2=0.6 δ=-0.41		p_1=0.1 p_2=0.7 δ=-3.04		p_1=0.8 p_2=0.8 δ=0		p_1=0.5 p_2=0.6 δ=-0.41		p_1=0.1 p_2=0.7 δ=-3.04		p_1=0.8 p_2=0.8 δ=0	
		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1	
	n_r	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**
Coverage (%)	5	98	96	91	90	99	98	91	92	74	82	97	93	99	97	99	99	99	98
	10	96	95	87	85	95	95	94	94	89	91	94	94	96	95	95	95	95	94
Bias (SE)	5	0.047	0.039	0.32	0.323	0.002	0	0.026	0.008	0.002	0.012	-0.003	-0.004	0.021	0.007	-0.022	0.021	-0.013	-0.009
		(0.276)	(0.243)	(0.316)	(0.268)	(0.293)	(0.256)	(0.276)	(0.238)	(0.283)	(0.254)	(0.332)	(0.259)	(0.305)	(0.256)	(0.381)	(0.306)	(0.324)	(0.274)
	10	0.025	0.02	0.209	0.201	-0.001	0	0.002	-0.004	-0.004	-0.001	0.004	-0.001	0	-0.005	0	-0.002	0.001	-0.003
		(0.204)	(0.182)	(0.233)	(0.201)	(0.214)	(0.187)	(0.211)	(0.182)	(0.242)	(0.206)	(0.225)	(0.19)	(0.214)	(0.185)	(0.26)	(0.214)	(0.226)	(0.191)
SB (RBS (%))	5	0.17	0.15	0.98	1.17	0.01	0	0.08	0.03	0	0.04	-0.01	-0.01	0.07	0.03	-0.06	0.07	-0.04	-0.03
		(-1)	(-7)	(-3)	(-3)	(-1)	(-5)	(-9)	(-16)	(-27)	(-21)	(1)	(-15)	(0)	(-10)	(-2)	(-5)	(-1)	(-10)
	10	0.12	0.11	0.86	0.95	0	0	0.01	-0.02	-0.01	0	0.02	-0.01	0	-0.03	0	-0.01	0	-0.01
		(-1)	(-1)	(-4)	(-4)	(-4)	(-4)	(-4)	(-7)	(-11)	(-11)	(-6)	(-8)	(-2)	(-5)	(-4)	(-8)	(-5)	(-7)
Sim Used	5	1000	1000	1000	1000	1000	1000	761	841	687	731	769	840	603	747	504	604	555	725
	10	1000	1000	1000	1000	1000	1000	904	963	833	927	908	959	850	941	773	890	807	932

1. Simulation parameters: Subject Variance=1.9, Reader Variance=0.5, Reader*Modality Variance=0.14
2. * : Sim Used=Used simulations: (1) Non-zero positive standard errors in convergent simulations (2) PROC GLIMMIX convergence criteria satisfied
3. ** : Sim Used=Used simulations: (1) PROC GLIMMIX convergence criteria satisfied (2) Positive-definite G matrix

Table 10 Simulation Results for Model C (medium magnitude of variance components)

		PL Estimation*						LAPLACE Estimation*						LAPLACE Estimation**					
		$p_1=0.5$		$p_1=0.1$		$p_1=0.8$		$p_1=0.5$		$p_1=0.1$		$p_1=0.8$		$p_1=0.5$		$p_1=0.1$		$p_1=0.8$	
		$p_2=0.6$		$p_2=0.7$		$p_2=0.8$		$p_2=0.6$		$p_2=0.7$		$p_2=0.8$		$p_2=0.6$		$p_2=0.7$		$p_2=0.8$	
		$\delta=-0.41$		$\delta=-3.04$		$\delta=0$		$\delta=-0.41$		$\delta=-3.04$		$\delta=0$		$\delta=-0.41$		$\delta=-3.04$		$\delta=0$	
		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1	
	n_r	55	100	55	100	55	100	55	100	55	100	55	100	55	100	55	100	55	100
Coverage (%)	5	95	96	89	88	97	96	93	93	86	89	96	94	96	95	97	96	98	96
	10	96	96	89	87	97	96	95	95	94	94	95	95	95	95	96	95	95	95
Bias	5	0.054	0.051	0.395	0.407	-0.002	-0.003	0	0	-0.032	0.003	-0.01	0.001	-0.008	0.003	-0.031	0.005	-0.007	-0.001
(SE)		(0.382)	(0.351)	(0.405)	(0.364)	(0.385)	(0.359)	(0.428)	(0.355)	(0.415)	(0.368)	(0.419)	(0.364)	(0.411)	(0.362)	(0.472)	(0.398)	(0.418)	(0.372)
	10	0.032	0.035	0.263	0.245	-0.001	0.002	0.002	0.007	0.003	0.011	-0.001	0.003	0	0.008	0.005	0.01	-0.001	0.003
		(0.292)	(0.278)	(0.313)	(0.289)	(0.298)	(0.279)	(0.3)	(0.282)	(0.332)	(0.298)	(0.312)	(0.285)	(0.3)	(0.283)	(0.339)	(0.301)	(0.307)	(0.283)
SB	5	0.13	0.14	0.93	1.06	-0.01	-0.01	0	0	-0.06	0.01	-0.02	0	-0.02	0.01	-0.06	0.01	-0.01	0
(RBS (%))		(-4)	(-7)	(-5)	(-5)	(-7)	(-7)	(-5)	(-18)	(-21)	(-20)	(-11)	(-17)	(-9)	(-16)	(-10)	(-13)	(-11)	(-15)
	10	0.11	0.13	0.82	0.85	0	0.01	0.01	0.03	0.01	0.03	0	0.01	0	0.03	0.01	0.03	0	0.01
		(-3)	(1)	(-2)	(0)	(2)	(0)	(-8)	(-5)	(-8)	(-8)	(-2)	(-6)	(-8)	(-5)	(-6)	(-7)	(-4)	(-6)
Sim Used	5	1000	1000	1000	1000	1000	1000	868	905	804	850	867	910	765	847	703	786	729	838
	10	1000	1000	1000	1000	1000	1000	977	991	963	981	979	992	957	988	944	971	959	983

1. Simulation parameters: Subject Variance=3.71, Reader Variance=0.8672, Reader*Modality Variance=0.4

2. * : Sim Used=Used simulations: (1) Non-zero positive standard errors in convergent simulations (2) PROC GLIMMIX convergence criteria satisfied

3. ** : Sim Used=Used simulations: (1) PROC GLIMMIX convergence criteria satisfied (2) Positive-definite G matrix

Table 11 Simulation Results for Model C (large magnitude of variance components)

		PL Estimation*						LAPLACE Estimation*						LAPLACE Estimation**					
		p_1=0.5		p_1=0.1		p_1=0.8		p_1=0.5		p_1=0.1		p_1=0.8		p_1=0.5		p_1=0.1		p_1=0.8	
		p_2=0.6		p_2=0.7		p_2=0.8		p_2=0.6		p_2=0.7		p_2=0.8		p_2=0.6		p_2=0.7		p_2=0.8	
		δ=-0.41		δ=-3.04		δ=0		δ=-0.41		δ=-3.04		δ=0		δ=-0.41		δ=-3.04		δ=0	
		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1	
	n_r	55	100	55	100	55	100	55	100	55	100	55	100	55	100	55	100	55	100
Coverage (%)	5	96	96	91	90	96	96	94	93	91	92	95	93	96	95	97	96	96	94
	10	96	95	91	92	96	95	95	93	95	94	94	95	95	93	95	95	95	95
Bias	5	0.069	0.057	0.47	0.468	-0.01	0.002	0.02	0.011	-0.001	0.012	-0.017	-0.002	0.021	0.008	-0.006	0.019	-0.019	0.002
(SE)		(0.524)	(0.507)	(0.538)	(0.512)	(0.527)	(0.507)	(0.552)	(0.52)	(0.572)	(0.53)	(0.585)	(0.525)	(0.56)	(0.526)	(0.61)	(0.55)	(0.568)	(0.526)
	10	0.031	0.035	0.296	0.276	-0.016	0.005	-0.008	0.001	-0.014	0.001	-0.018	0.004	-0.007	0	-0.014	0.002	-0.02	0.003
		(0.422)	(0.412)	(0.435)	(0.418)	(0.426)	(0.413)	(0.438)	(0.424)	(0.467)	(0.436)	(0.445)	(0.426)	(0.438)	(0.424)	(0.469)	(0.438)	(0.445)	(0.426)
SB	5	0.12	0.11	0.83	0.88	-0.02	0	0.03	0.02	0	0.02	-0.03	0	0.03	0.01	-0.01	0.03	-0.03	0
(RBS (%))		(-8)	(-5)	(-4)	(-4)	(-6)	(-6)	(-17)	(-17)	(-19)	(-18)	(-10)	(-17)	(-16)	(-16)	(-14)	(-15)	(-14)	(-16)
	10	0.07	0.08	0.68	0.66	-0.04	0.01	-0.02	0	-0.03	0	-0.04	0.01	-0.02	0	-0.03	0	-0.04	0.01
		(1)	(-1)	(0)	(0)	(1)	(-1)	(-5)	(-7)	(-6)	(-7)	(-5)	(-7)	(-5)	(-7)	(-6)	(-7)	(-5)	(-7)
Sim Used	5	1000	1000	1000	1000	1000	1000	915	930	886	927	921	940	859	892	830	893	844	894
	10	1000	1000	1000	1000	1000	1000	991	996	986	990	994	992	983	990	982	986	985	988

1. Simulation parameters: Subject Variance=5, Reader Variance=2, Reader*Modality Variance=1
2. * : Sim Used=Used simulations: (1) Non-zero positive standard errors in convergent simulations (2) PROC GLIMMIX convergence criteria satisfied
3. ** : Sim Used=Used simulations: (1) PROC GLIMMIX convergence criteria satisfied (2) Positive-definite G matrix

For the setting with a continuous covariate (model D, Table 12, Table 13, Table 14), the PL estimated model resulted in adequate coverage for smaller values of slope but started to decline for higher values (slope = 0.1). Like other models, the standardized bias using the LA approach was negligible. The coverage for the LA approach was conservative despite the serious under-estimation of standard error especially with fewer readers. Similar observations prevailed in the subset of simulations with positive-definite estimates of the covariance matrix. Combination approaches (Table 15) based on the Satterthwaite approximation led to somewhat better coverage rate, while borrowing PL variance estimator enabled use of the model in larger number of instances, while making CIs slightly more conservative.

Table 12 Simulation Results for Model D (small magnitude of variance components)

		PL Estimation*								**LAPLACE Estimation***								**LAPLACE Estimation****							
		τ=0 **μ=0**		**τ=.03** **μ=0**		**τ=.06** **μ=0**		**τ=0.1** **μ=0**		**τ=0** **μ=0**		**τ=.03** **μ=0**		**τ=.06** **μ=0**		**τ=0.1** **μ=0**		**τ=0** **μ=0**		**τ=.03** **μ=0**		**τ=.06** **μ=0**		**τ=0.1** **μ=0**	
		n_1		**n_1**		**n_1**		**n_1**		**n_1**		**n_1**		**n_1**		**n_1**		**n_1**		**n_1**		**n_1**		**n_1**	
	n_r	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**
Coverage (%)	**5**	98	98	98	96	96	92	93	88	98	96	98	96	97	96	98	97	98	97	99	96	97	97	99	97
	10	97	97	96	95	90	89	82	77	96	97	96	97	96	96	97	97	96	97	96	97	96	96	97	97
Bias (SE)	**5**	0 (0.01)	0 (0.009)	-0.004 (0.01)	-0.004 (0.009)	-0.007 (0.011)	-0.008 (0.01)	-0.013 (0.015)	-0.013 (0.012)	0 (0.011)	0 (0.009)	0 (0.011)	-0.001 (0.009)	0 (0.012)	0 (0.01)	0.003 (0.018)	0.001 (0.013)	0 (0.01)	0 (0.009)	-0.001 (0.011)	-0.001 (0.009)	0 (0.013)	-0.001 (0.01)	0.003 (0.019)	0 (0.014)
	10	0 (0.008)	0 (0.007)	-0.003 (0.008)	-0.003 (0.007)	-0.007 (0.008)	-0.007 (0.007)	-0.012 (0.011)	-0.012 (0.009)	0 (0.008)	0 (0.007)	0 (0.009)	0 (0.008)	0 (0.01)	0 (0.008)	0.001 (0.013)	0 (0.01)	0 (0.008)	0 (0.007)	0 (0.009)	0 (0.007)	0 (0.01)	0 (0.008)	0.001 (0.013)	0 (0.01)
SB (RBS (%))	**5**	0 (4)	-0.03 (-3)	-0.36 (3)	-0.44 (-1)	-0.66 (3)	-0.79 (-1)	-0.94 (12)	-1.05 (3)	0.03 (4)	0.01 (-10)	-0.02 (-4)	-0.05 (-12)	0.01 (-8)	-0.02 (-14)	0.16 (-5)	0.05 (-11)	0.01 (-3)	-0.05 (-10)	-0.07 (-2)	-0.07 (-10)	-0.03 (-4)	-0.11 (-11)	0.16 (1)	0 (-6)
	10	0.03 (3)	0.01 (0)	-0.37 (2)	-0.44 (1)	-0.8 (1)	-0.92 (1)	-1.17 (3)	-1.36 (2)	0.05 (0)	0.02 (-2)	0.02 (-3)	-0.02 (-3)	0 (-2)	-0.04 (-4)	0.05 (-3)	-0.02 (-4)	0.02 (-2)	0.01 (-3)	0.01 (-3)	-0.03 (-4)	-0.02 (-3)	-0.05 (-4)	0.06 (-2)	-0.01 (-3)
Sim Used	**5**	907	956	908	948	879	919	603	721	978	973	988	988	993	994	997	998	569	737	542	713	449	627	306	459
	10	984	993	986	998	966	982	728	845	975	988	980	998	989	996	975	997	831	932	805	928	738	897	571	764

1. Simulation parameters: Subject Variance=1, Reader Variance=0.1, Reader*LesionSize Variance=0.0004
2. Continuous variable (LesionSize)~Unif(1,100) and has been centered at 50.5 mm during simulation and fitting process
3. * : Sim Used=Used simulations: (1) Non-zero positive standard errors in convergent simulations (2) PROC GLIMMIX convergence criteria satisfied
4. ** : Sim Used=Used simulations: (1) PROC GLIMMIX convergence criteria satisfied (2) Positive-definite G matrix

Table 13 Simulation Results for Model D (medium magnitude of variance components)

		PL Estimation*								LAPLACE Estimation*								LAPLACE Estimation**							
		τ=0		τ=.03		τ=.06		τ=0.1		τ=0		τ=.03		τ=.06		τ=0.1		τ=0		τ=.03		τ=.06		τ=0.1	
		μ=0		μ=0		μ=0		μ=0		μ=0		μ=0		μ=0		μ=0		μ=0		μ=0		μ=0		μ=0	
		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1	
	n_r	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**
Coverage (%)	**5**	99	98	98	96	94	91	86	84	97	96	98	96	98	96	97	97	98	96	98	97	98	97	97	97
	10	97	97	96	95	91	90	74	75	96	96	96	96	96	96	97	96	96	97	96	96	97	96	97	96
Bias (SE)	**5**	0	0	-0.006	-0.006	-0.012	-0.011	-0.02	-0.019	0.001	0	0	0	0	0	0.002	0	0	0	-0.001	-0.001	-0.001	-0.001	0.001	-0.001
		(0.014)	(0.013)	(0.014)	(0.013)	(0.014)	(0.013)	(0.017)	(0.015)	(0.015)	(0.014)	(0.015)	(0.014)	(0.017)	(0.014)	(0.022)	(0.017)	(0.015)	(0.014)	(0.015)	(0.014)	(0.017)	(0.014)	(0.022)	(0.017)
	10	0	0	-0.004	-0.004	-0.009	-0.009	-0.018	-0.016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		(0.011)	(0.01)	(0.011)	(0.01)	(0.011)	(0.01)	(0.013)	(0.011)	(0.012)	(0.011)	(0.012)	(0.011)	(0.013)	(0.012)	(0.016)	(0.013)	(0.012)	(0.011)	(0.012)	(0.011)	(0.013)	(0.012)	(0.016)	(0.013)
SB (RBS (%))	**5**	0	-0.02	-0.42	-0.43	-0.83	-0.87	-1.25	-1.38	0.03	0.01	-0.01	-0.01	0	-0.02	0.06	0.01	0.01	-0.01	-0.06	-0.06	-0.07	-0.06	0.06	-0.03
		(3)	(-3)	(4)	(-4)	(4)	(0)	(8)	(5)	(-4)	(-11)	(-6)	(-15)	(-6)	(-15)	(-12)	(-14)	(-7)	(-12)	(-5)	(-14)	(-6)	(-14)	(-8)	(-13)
	10	0.02	0.01	-0.38	-0.41	-0.87	-0.9	-1.46	-1.56	0.04	0.01	0.03	0	0.02	-0.02	0.03	-0.02	0.01	0	0.03	-0.01	0.02	-0.03	0.02	-0.03
		(2)	(1)	(2)	(2)	(6)	(4)	(5)	(4)	(-2)	(-4)	(-3)	(-4)	(-1)	(-4)	(-4)	(-6)	(-3)	(-4)	(-4)	(-4)	(-2)	(-4)	(-4)	(-6)
Sim Used	**5**	939	980	940	970	921	953	794	892	968	975	985	994	991	994	984	997	758	855	728	851	690	822	537	725
	10	996	997	992	999	971	984	891	950	979	997	991	997	991	998	986	997	935	977	926	976	897	972	821	942

1. Simulation parameters: Subject Variance=2, Reader Variance=0.2, Reader*LesionSize Variance=0.001
2. Continuous variable (LesionSize)~Unif(1,100) and has been centered at 50.5 mm during simulation and fitting process
3. * : Sim Used=Used simulations: (1) Non-zero positive standard errors in convergent simulations (2) PROC GLIMMIX convergence criteria satisfied
4. ** : Sim Used=Used simulations: (1) PROC GLIMMIX convergence criteria satisfied (2) Positive-definite G matrix

Table 14 Simulation Results for Model D (large magnitude of variance components)

		PL Estimation*								LAPLACE Estimation*								LAPLACE Estimation**							
		τ=0 μ=0		τ=.03 μ=0		τ=.06 μ=0		τ=0.1 μ=0		τ=0 μ=0		τ=.03 μ=0		τ=.06 μ=0		τ=0.1 μ=0		τ=0 μ=0		τ=.03 μ=0		τ=.06 μ=0		τ=0.1 μ=0	
		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1		n_1	
	n_r	55	100	55	100	55	100	55	100	55	100	55	100	55	100	55	100	55	100	55	100	55	100	55	100
Coverage (%)	**5**	97	96	96	96	96	96	92	93	96	95	96	95	97	95	96	96	96	95	95	95	97	96	96	96
	10	96	96	96	95	94	95	90	91	96	95	95	95	96	95	95	95	96	95	95	95	96	95	95	95
Bias (SE)	**5**	0 (0.034)	0 (0.034)	-0.007 (0.034)	-0.007 (0.034)	-0.016 (0.035)	-0.014 (0.034)	-0.027 (0.035)	-0.025 (0.035)	0.001 (0.041)	0 (0.039)	0.001 (0.041)	-0.001 (0.039)	0 (0.043)	0 (0.04)	0.002 (0.046)	-0.001 (0.041)	0 (0.041)	-0.001 (0.039)	-0.001 (0.041)	-0.001 (0.039)	-0.001 (0.043)	-0.001 (0.04)	0.001 (0.046)	-0.002 (0.041)
	10	0 (0.027)	0 (0.027)	-0.006 (0.027)	-0.005 (0.027)	-0.013 (0.027)	-0.011 (0.027)	-0.023 (0.027)	-0.02 (0.027)	0 (0.032)	0 (0.03)	0 (0.032)	0 (0.031)	0 (0.032)	0 (0.031)	0.001 (0.034)	-0.001 (0.031)	0 (0.032)	0 (0.03)	0 (0.032)	0 (0.031)	0 (0.032)	0 (0.031)	0.001 (0.034)	-0.001 (0.031)
SB (RBS (%))	**5**	0 (0)	0 (-3)	-0.22 (1)	-0.2 (-4)	-0.46 (3)	-0.4 (-3)	-0.81 (5)	-0.72 (1)	0.03 (-12)	0 (-15)	0.01 (-12)	-0.01 (-16)	0.01 (-12)	0 (-17)	0.04 (-25)	-0.03 (-15)	-0.01 (-12)	-0.01 (-15)	-0.02 (-13)	-0.03 (-16)	-0.03 (-13)	-0.02 (-17)	0.01 (-24)	-0.05 (-15)
	10	0.02 (2)	0.01 (0)	-0.24 (1)	-0.2 (0)	-0.5 (2)	-0.42 (1)	-0.91 (5)	-0.78 (4)	0.01 (-4)	0.01 (-6)	0.01 (-6)	0 (-6)	0.01 (-5)	0 (-6)	0.02 (-5)	-0.03 (-6)	0.01 (-5)	0.01 (-6)	0.01 (-6)	0 (-6)	0.01 (-5)	0 (-6)	0.02 (-6)	-0.03 (-6)
Sim Used	**5**	994	994	990	997	983	995	968	978	971	991	992	995	996	1000	998	998	915	968	911	967	918	969	890	951
	10	989	998	990	996	983	997	967	989	999	1000	999	1000	998	1000	996	1000	994	1000	993	999	993	999	990	1000

1. Simulation parameters: Subject Variance=3, Reader Variance=1, Reader*LesionSize Variance=0.01
2. Continuous variable (LesionSize)~Unif(1,100) and has been centered at 50.5 mm during simulation and fitting process
3. * : Sim Used=Used simulations: (1) Non-zero positive standard errors in convergent simulations (2) PROC GLIMMIX convergence criteria satisfied
4. ** : Sim Used=Used simulations: (1) PROC GLIMMIX convergence criteria satisfied (2) Positive-definite G matrix

Table 15 Coverage of 95% CI for Models A, B, C and D using Combination Approaches

			LA + Satterthwaite d.f. from PL [44]				PL+LA Estimation (containment d.f.)			
			Coverage (%)		Sim Used		Coverage (%)		Sim Used	
			n_r		n_r		n_r		n_r	
Model	Fixed Effect Parameters	Sample Size	5	10	5	10	5	10	5	10
A	p=0.1, μ=-2.20	n_1=55	92	94	953	998	95	96	999	1000
		n_1=100	94	95	970	1000	95	96	1000	1000
	p=0.5, μ=0	n_1=55	94	95	996	999	97	96	1000	1000
		n_1=100	93	95	998	1000	95	96	1000	1000
	p=0.7, μ=0.85	n_1=55	95	96	1000	1000	97	97	1000	1000
		n_1=100	93	95	1000	1000	96	96	1000	1000
B	TPF=0.3, FPF=0.9, η=-3.04	n_1=55, n_0=100	88	93	652	895	96	96	1000	999
		n_1=55, n_0=175	88	94	709	912	97	96	1000	999
		n_1=100, n_0=100	88	93	726	931	96	95	1000	1000
		n_1=100, n_0=175	90	93	775	950	96	95	1000	1000
	TPF=0.6, FPF=0.6, η=0	n_1=55, n_0=100	93	94	850	959	97	96	999	1000
		n_1=55, n_0=175	93	97	835	970	97	97	1000	1000
		n_1=100, n_0=100	92	94	881	977	97	96	998	1000
		n_1=100, n_0=175	94	96	897	984	97	98	1000	999
	TPF=0.5, FPF=0.03, η=3.48	n_1=55, n_0=100	83	90	662	861	96	95	999	996
		n_1=55, n_0=175	84	92	701	882	95	97	1000	997
		n_1=100, n_0=100	81	91	670	877	95	95	999	999
		n_1=100, n_0=175	83	94	722	921	94	96	1000	998
C	p_1=0.5, p_2=0.6, δ=-0.41	n_1=55	92	94	868	977	94	95	1000	999
		n_1=100	92	95	905	991	94	95	1000	1000
	p_1=0.1, p_2=0.7, δ=-3.04	n_1=55	85	94	804	963	95	96	1000	1000
		n_1=100	89	94	850	981	94	95	1000	1000
	p_1=0.8, p_2=0.8, δ=0	n_1=55	94	95	867	979	95	95	1000	999
		n_1=100	93	95	910	992	93	95	999	1000
D	μ=0, τ=0	n_1=55	95	95	968	979	97	96	993	1000
		n_1=100	94	95	975	997	96	97	998	1000
	μ=0, τ=0.03	n_1=55	96	96	985	991	98	96	996	994
		n_1=100	94	95	994	997	96	96	997	999
	μ=0, τ=0.06	n_1=55	97	96	991	991	98	96	997	997
		n_1=100	94	95	994	998	96	96	998	1000
	μ=0, τ=0.1	n_1=55	96	96	984	986	97	97	988	991
		n_1=100	95	95	997	997	97	96	997	999

1. LA + Satterthwaite d.f. from PL: Laplace model with sattherthwaite degrees of freedom borrowed from PL model
2. PL+LA Estimation: Combination model with fixed effect estimates from Laplace technique; Standard Error (SE) estimates of LA model replaced by those of PL model only when G-matrix is not positive definite (p.d.). In case G-matrix is p.d. for both models, the greater of the two SE is utilized

2.4.4 Overall Summary of Simulation Results

Overall, we observed that both the default PL and LA approaches had problems of different nature. The PL approach led to bias estimating the fixed effect and poor CI coverage across all considered models, except for scenarios when the true probabilities were close to 0.5 or when the slope was small with bigger values of variance components. The LA approach performed well when the reader sample size was large and the probabilities were closer to 0.5. However, in other scenarios, the LA approach, while producing nearly unbiased estimates of the fixed effects, occasionally led to substantial underestimation of variability, non-positive definite covariance matrix and the related inadequately low CI coverage. In general, we encountered more convergence problems with this approach. We also observed that in instances when the LA approach led to positive definite covariance matrix, fixed effect estimates remain accurate and the confidence intervals become somewhat conservative. However, the chances of model converging and having positive definite covariance matrix were not high.

Previously recommended use of the Satterthwaite approximation had little effect on coverage of the LA-based confidence intervals, and by design, could not remedy the instances where statistical inferences with the LA estimates were not possible. However, a simple borrowing of the PL variance estimate in instances when the LA estimates of the G matrix was not positive definite, enabled adequate statistical inferences in virtually all instances (Table 15).

To cover a broader range of variance parameters, we also provided results of simulations for models A, B, C and D for scenarios with smaller and larger variance structures. Across all four models we observed that for smaller choice of variance components, there were more issues with the convergence and the confidence interval coverage. These specific cases would especially benefit from the combination strategy we considered. In cases when the variance components were large, there were fewer issues with the CI coverage, but still a number of scenarios where simulations did not converge or the estimated SE was zero. In these scenarios the combination approach also enabled adequate statistical inferences.

One of the remaining questions is the relative performance of the built-in approaches as compared with more complicated methods based on the modified LA, MCMC or other techniques. Our simulation results indicate that application of these extensions is not necessary for inferences about fixed effect estimates since the drawbacks of the built-in LA approach, with a possible combination with PL estimates, are negligible, if present, for these purposes. However, application of more advanced techniques might be necessary for variance components estimation, where both PL and LA approaches have substantial deficiencies. Appendix F illustrates possible differences between the results of various approaches, by comparing estimates for fixed effects and variance components obtained from the well-known Salamander mating dataset [1]. Assuming that the relationship between these estimates are representative of a general scenario (i.e., LA estimates are virtually unbiased for fixed effects, but biased downward for variance components), these results indicate that while MCMC approaches [40, 60] may lead to variance components estimates without negative bias, they might provide upwardly biased estimates of the fixed effects.

2.5 SUMMARY AND DISCUSSION

Our investigation demonstrated that for typical data from fully crossed multi-reader studies, the available built-in GLMMs for cross-correlated data while being easy to implement, frequently require adjustments to enable adequate statistical inferences. Our results indicated that the problems of Pseudo-Likelihood (PL) estimation approach stemmed from the large bias of fixed effect estimates, while the problems of the Laplace (LA) approach resulted from frequent convergence issues and substantial underestimation of variance. At the same time we demonstrated that when the LA approach leads to positive definite estimate of covariance matrix, the coverage of its confidence intervals is never less than nominal, across all scenarios. Unfortunately, positive definite estimates are not frequent, and in some scenarios are obtained only in 55% of instances, which make the restricted application of the LA approach impractical.

In line with previously proposed combination approaches [25, 44], we noted that the complementary nature of problems in the PL and LA estimation of the same model suggest a simple combination strategy. Namely, in instances when the LA approach leads to non-positive definite estimates of covariance matrix one can enable adequate statistical inferences by borrowing variance from PL approach. Due to the virtual absence of convergence issues with the PL approach in the considered models, this simple combination approach enables statistical analysis with adequate statistical properties in the absolute majority of cases. A potentially better but cumbersome approach could be to only replace certain variance components instead of borrowing the estimate of SE of fixed effect (similar to Capanu *et al.* [25]). Even though it is known that the PL estimates of variance components are more biased than the corresponding LA estimates, borrowing them only in situations when the latter are degenerate (zero or negative) might have little detrimental effect on the CI coverage, if any.

These results also indicate that for inferences about fixed effects it is not necessary to use more advanced estimation techniques since the built-in LA approach, with a possible combination with PL estimates, provides adequate results for these purposes. At the same time, our results confirm the expectation that more advanced techniques are necessary if the goal is to obtain accurate estimates of the variance components (which are often used for planning future studies). Once of the commonly recommended approaches is the Markov Chain Monte Carlo methods. However, these methods can also lead to substantially different estimates of the variance components (e.g., [60], [40]). For simpler models, e.g., models A-C, the probability-scale variance components can also be estimated using non-parametric methods (Gallas *et al.* 2007 [52]).

A possible reason for smaller bias in fixed effect estimates of the LA vs. PL approach could be the fact that the former is a single-iterative approach where fixed effect and variance components are estimated simultaneously, leading to less severe dependence on the estimated variance parameters. However, this approach also seems to lead to more frequent problems in obtained positive-definite estimates of covariance matrix. For the PL estimation of a GLMM with a non-linear link function, the pseudo-deviates are estimated at every iteration step. Estimation of the pseudo-deviates depends on the variance components estimated based on the pseudo-likelihood. As a result, the PL estimation approach imposes substantial dependence between the fixed effect estimates and the estimated variance components [29]. This property, on the other hand, seems to reduce the problems with inadmissible variance estimates enabling a practical approach based on combination of LA and PL estimates.

The models we considered in the simulation study represent typical analyses of multi-reader diagnostic imaging studies. These models are complex enough to identify some essential advantages and drawbacks of standard GLMM in multi-reader studies. However, our assessment was focused on scenarios where the structure of the estimated GLMM is perfectly correct (i.e., when the simulations and analytical models are exactly the same). Robustness of the considered approaches to model misspecification can be evaluated using alternative simulation models for multi-reader data [58].

Overall, the GLMM approaches currently available in standard software packages offer simple and flexible tools for handling categorical and continuous covariates in non-linear models for fully crossed data. However, deficiencies of the LA and PL estimation techniques individually, require awareness of possible problems. It is a good practice to carefully examine both types of estimates for a given dataset. When the covariance matrix is not positive definite for LA, but not for PL approach, borrowing the PL estimate of variance could enable conservative inferences. Thus, when used in combination, the Laplace and Pseudo-likelihood estimates of GLMMs enable straightforward and adequate, albeit somewhat conservative, statistical inferences for analyses of cross-correlated binary data from typical multi-reader studies of diagnostic imaging.

3.0 HALF-MARGINAL GLMM FOR ANALYSIS OF CROSS-CORRELATED BINARY DATA IN MULTI-READER STUDIES OF DIAGNOSTIC ACCURACY

A standard approach for analyzing cross-correlated data is based on the Generalized Linear Mixed Models (GLMM) with crossed random effects. For models with non-linear link this implies a "subject-specific" interpretation of the estimated coefficients. For typical multi-reader studies of diagnostic imaging, the corresponding coefficients are not the primary targets of interests and have a rather artificial interpretation which is difficult to illustrate with data.

In this section, we propose a half-marginal GLMM which offers a more natural parametrization for modeling cross-correlated data from a multi-reader study. We illustrate that the model can easily be implemented using the built-in machinery and that for simple models the resulting Pseudo-Likelihood estimates are close to the simple empirical estimates.

Investigations of statistical properties of half-marginal models are complicated by the difficulty to fully specify the probability distribution of the data in terms of model's parameters. To circumvent this problem we considered half-marginal models induced by a range of the subject-specific GG models. The half-marginal model coefficients were derived for standard models for multi-reader data and estimated numerically for the considered range of parameter configurations. Simulation results indicate that the model performs very well across the number of scenarios typical for multi-reader studies. However, the patterns observed for bias of fixed effect estimates indicate potential large-sample problems at least in some modeling scenarios.

3.1 MECHANISMS FOR USING SUBJECT-SPECIFIC (G) AND MARGINAL (R) STRUCTURES IN GLMM

Because of the non-linear link function in GLMM, the interpretation of a model's parameter changes from individuals to entire population depending on whether we address the correlation through G side (via. introduction of random effects) or R side (directly modeling correlation) or a mixture of both. In standard mixed model notation, G is the variance-covariance matrix of random effects whereas R is the variance-covariance matrix of residuals. To illustrate that coefficients have a different interpretation in GLMMs, observe that for a linear mixed model with identity link:

$$E\left(Y_{ij}|X_{ij}, b_i\right) = X'_{ij}\beta + Z'_{ij}b_i,$$

where X is the design matrix for fixed-effects and Z is the design matrix for random effects.

Also, $E\left(Y_{ij}|\, X_{ij}\right) = X'_{ij}\beta$ when averaged over distribution of random effects b_i i.e. fixed effects (β) in model for conditional means also have same interpretation in terms of population means. However, for a GLMM model with logit link (in case of binary data):

$$logit^{-1}\left(E\left(Y_{ij}|X_{ij}, b_i\right)\right) = X'_{ij}\beta + Z'_{ij}b_i,$$

i.e. $E\left(Y_{ij}|X_{ij}, b_i\right)$ is a non-linear function of β, b and $logit^{-1}\left(E\left(Y_{ij}|X_{ij}\right)\right) \neq X'_{ij}\beta$ for any β.

In the context of multi-reader studies where we have a representative sample of both readers and subjects, when using a GG/"subject-specific" model as in Chapter 2.0 (where readers and subjects are both specified as crossed-random effects), the regression parameters have a "subject-specific" interpretation (or more precisely "reader-subject specific") e.g., the effect on probability of calling an average subject "positive" by an average reader. However, in such studies of diagnostic imaging, the frequent target of interest is for e.g., sensitivity or specificity, which are marginal quantities (averaged over the population of subjects). Hence, it is natural to account for subject-related correlation i.e. between Y_{ij} and $Y_{ij'}$ with R side effects. In contrast, correlation and heterogeneity due to readers is natural to address with the G side mechanism since it is common to make inferences for individual readers as well as for the average over all readers in the study [2]. The corresponding half-marginal model (which we call RG model, with R side effects for subjects and G side effects for readers) allows inferring about reader-specific, subject-marginal characteristics, e.g., effect on sensitivity for an average reader. Considering the most basic scenario, the half-marginal GLMM model can be written as follows:

$$logit(\tilde{p}_j) = \tilde{\mu} + \tilde{\beta}_j, \tag{3.1}$$

where $Y_{ij}|\tilde{\beta}_j \sim Bernoulli(1, \tilde{p}_j)$, $i = 1, \dots, n_1$ is the index representing diseased subjects, $j = 1, \dots, n_r$ is the reader index, $\tilde{\beta}_j \sim N(0, \tilde{\sigma}_\beta^2)$ is the reader random effect. $Corr(Y_{ij}, Y_{ij'}|\tilde{\beta}_j, \tilde{\beta}_{j'}) = \rho$ is conditional correlation (i.e., compound symmetry structure), hence the correlation matrix $P \neq I$ for this model (unlike the GG model). The residual matrix should be $R = A^{1/2}B^{1/2}PB^{1/2}A^{1/2} = B^{1/2}PB^{1/2}$ such that $B = (diag\{\tilde{p}_j(1 - \tilde{p}_j)\})$ and $A = I$.

Here, $logit^{-1}(\tilde{\mu} + \tilde{\beta}_j)$ is the sensitivity for reader j and $logit^{-1}(\tilde{\mu})$ is the reader-averaged sensitivity.

The estimation target for this RG model is $\tilde{\mu}$, which is the log-odds of a probability for a population of diseased subjects being correctly diagnosed by an average reader for which $\tilde{\beta}_j = 0$ (or alternatively the sensitivity itself: $\tilde{p} = (1 + exp(-\tilde{\mu}))^{-1}$).

Ability to address both reader-specific and overall parameters of diagnostic accuracy makes a half-marginal (RG) model rather relevant for analysis of multi-readers studies of diagnostic imaging. However, properties of this type of model are little known. Moreover, in contrast with the GG model, the RG model, although not directly comparable, does not require estimation of multiple subject-related random effects and therefore could lead to more stable estimates with decreased computation problems.

3.2 ESTIMATION TECHNIQUE FOR RG MODEL

Currently, the only standard built-in estimation technique to estimate parameters of the RG model is the PL technique by Wolfinger and O'Connell (1993) [21] as described in Section 2.2.2. The difference from the GG model lies in the specification of the correlation matrix P which will not be identity and will assume a different working correlation structure. For the RG models that we consider, P has a compound symmetry structure which mimics the default variance-covariance structure implied by GG models. Laplace approximation in most software packages cannot currently handle R side random effects and hence is not a valid option to fit RG models.

Since the RG model is implied from the GG model by averaging out the subject random effects, it is possible to obtain the estimates of the fixed-effect RG parameter and corresponding confidence limits using the GG model. Numerical integration can be performed conditioning on the GG estimates of fixed effect or CI along with estimates of variance components. Similar approaches have been described in Section 3.4 (illustrate computation of fully-marginal estimates from RG model) and Section 3.5.2 (derivation of true RG fixed-effect parameters from GG model). Also, as seen earlier in Chapter 2.0 , the GG model can be fit using either the PL or LA approaches. However, since both the approaches have been criticized for producing biased variance component estimates [25, Appendix C], possible bias may creep in the RG estimates of fixed-effect or CI estimates which are computed using these variance components.

3.3 HALF-MARGINAL MODELS FOR TYPICAL ANALYSES OF MULTI-READER DIAGNOSTIC ACCURACY STUDIES

To investigate the quality of statistical inferences based on fitting RG model using PL approach, we performed simulation studies and again considered modeling scenarios similar to ones considered in Section 2.3. The model structure this time is different due to the inclusion of both R and G side effects instead of only G side effects. In effect, we still consider the important sources of variability and correlations possible in a particular multi-reader data but through a combination of R and G side mechanism. The SAS code to implement these models is provided in Appendix B. Below is a brief description of the RG analysis models. The indexes hold the same meaning as in Section 2.3.

3.3.1 RG Model A: Covariate free model (e.g. inferences on sensitivity or specificity for a single modality)

$logit(\tilde{p}_j) = \tilde{\mu} + \tilde{\beta}_j.$

This model has been fully specified using equation (3.1). Here, $logit^{-1}(\tilde{\mu} + \tilde{\beta}_j)$ is the sensitivity for reader j.

3.3.2 RG Model B: Subject-level binary covariate (e.g. inferences on sensitivity and specificity combined)

$logit(\tilde{p}_{jD}) = \tilde{\mu} + \tilde{\eta}_D + \tilde{\beta}_j + \tilde{\gamma}_{jD},$

where $Y_{ij}|\tilde{\beta}_j, \tilde{\gamma}_{jD} \sim Bin(1, \tilde{p}_{jD})$, $i = 1, \dots, n_0 + n_1$ is the subject index, n_0= number of non-diseased subjects, n_1= number of diseased subjects, $j = 1, \dots, n_r$ is the reader index, $\tilde{\eta}_D$ is the fixed effect for true disease status $(D = 1{:}\, diseased, -1{:}\, non - diseased)$, $\tilde{\beta}_j \sim N\left(0, \tilde{\sigma}_\beta^2\right)$ is the random reader effect, $\tilde{\gamma}_{jD} \sim N(0, \tilde{\sigma}_D^2)$ is the random interaction term between reader and true disease status.

$Corr\left(Y_{ijD}, Y_{ij'D} | \tilde{\beta}_j, \tilde{\beta}_{j'}, \tilde{\gamma}_{jD}, \tilde{\gamma}_{j'D}\right) = \rho$ is the conditional correlation structure for any given subject specified using compound symmetry and assumed to be same across both disease statuses.

The primary inferential target for this model is coefficient $\tilde{\eta} = \tilde{\eta}_1 - \tilde{\eta}_{-1} = ln\left[\frac{\widetilde{TPF}/(1-\widetilde{TPF})}{\widetilde{FPF}/(1-\widetilde{FPF})}\right]$, which is the natural log of Diagnostic Odds Ratio (DOR) [2] for an average reader for a population of diseased and non-diseased subjects. Additional targets which could be of interest are an average reader's $\widetilde{TPF}$ i.e. $logit^{-1}(\tilde{\mu} + \tilde{\eta}_1)$ and $\widetilde{FPF}$ i.e. $logit^{-1}(\tilde{\mu} + \tilde{\eta}_{-1})$.

3.3.3 RG Model C: Assessment-level binary covariate (e.g., comparisons of sensitivity between two modalities)

$logit\left(\tilde{p}_{jM}\right) = \tilde{\mu} + \tilde{\delta}_M + \tilde{\beta}_j + \tilde{\gamma}_{jM}$,

where $Y_{ijM} | \tilde{\beta}_j, \tilde{\gamma}_{jM} \sim Bin\left(1, \tilde{p}_{jM}\right)$, $i = 1, \dots, n_1$ is the subject index representing only diseased subjects, $j = 1, \dots, n_r$ is the reader index, $\tilde{\delta}_M$ is the fixed effect of modality M $(M = 1, -1)$, $\tilde{\beta}_j \sim N\left(0, \tilde{\sigma}_\beta^2\right)$ is the random reader effect, $\tilde{\gamma}_{jM} \sim N(0, \tilde{\sigma}_M^2)$ is the random interaction term between reader and modality.

$Corr\left(Y_{ijM}, Y_{ij'M} | \tilde{\beta}_j, \tilde{\beta}_{j'}, \tilde{\gamma}_{jM}, \tilde{\gamma}_{j'M}\right) = Corr\left(Y_{ijM}, Y_{ijM'} | \tilde{\beta}_j, \tilde{\gamma}_{jM}, \tilde{\gamma}_{jM'}\right) =$ $Corr\left(Y_{ijM}, Y_{ij'M'} | \tilde{\beta}_j, \tilde{\beta}_{j'}, \tilde{\gamma}_{jM}, \tilde{\gamma}_{j'M'}\right) = \rho$ is assumed to be compound symmetry for any given subject

The primary estimation target for this RG model is coefficient $\tilde{\delta} = \tilde{\delta}_1 - \tilde{\delta}_{-1} = ln\left[\frac{\widetilde{TPF}_{M=1} / (1-\widetilde{TPF}_{M=1})}{\widetilde{TPF}_{M=-1} / (1-\widetilde{TPF}_{M=-1})}\right]$ which is the log of odds ratio for comparing sensitivity levels of two modalities for an average reader. Additional targets of interest can be an average reader's $\widetilde{TPF}_{M=1}$ i.e. $logit^{-1}(\tilde{\mu} + \tilde{\delta}_1)$ and $\widetilde{TPF}_{M=-1}$ i.e. $logit^{-1}(\tilde{\mu} + \tilde{\delta}_{-1})$.

3.3.4 RG Model D: Subject-level continuous covariate (e.g., lesion size effect on sensitivity)

$logit\left(\tilde{p}_{jX}\right) = \tilde{\mu} + \tilde{\tau} * X + \tilde{\beta}_j + \tilde{\gamma}_j * X,$

where $Y_{ij}|X, \tilde{\beta}_j, \tilde{\gamma}_j \sim Bin\left(1, \tilde{p}_{jX}\right)$, $i = 1, \dots, n_1$ is the subject index, $j = 1, \dots, n_r$ is the reader index, $\tilde{\tau}$ is the fixed effect of slope, $\tilde{\beta}_j \sim N\left(0, \tilde{\sigma}_{\beta}^2\right)$ is the random reader effect, $\tilde{\gamma}_j \sim N(0, \tilde{\sigma}_X^2)$ is the random interaction term between reader and lesion size (X). $Corr\left(Y_{ij}, Y_{ij'} | \tilde{\beta}_j, \tilde{\beta}_{j'}, \tilde{\gamma}_j, \tilde{\gamma}_{j'}\right) = \rho$ is the working correlation structure for any given subject.

The estimation target for this RG model is coefficient $\tilde{\tau} = ln\left(\frac{\widetilde{TPF}_{X=x+1} / (1-\widetilde{TPF}_{X=x+1})}{\widetilde{TPF}_{X=x} / (1-\widetilde{TPF}_{X=x})}\right)$ which is the slope for an average reader.

3.4 COMPUTATION OF FULLY-MARGINAL ESTIMATES FROM RESULTS OF HALF-MARGINAL MODELS

Given that we primarily focus on GG and RG models in this dissertation, it may also be of interest to fit a fully marginal (RR) model since the parameters of that model are directly relevant in practice e.g., overall sensitivity of a modality across a population of readers and subjects. However, standard software like PROC GLIMMIX in SAS does not allow fitting such models due to the inability to fit two residual side effects in the same model. Regardless, one can compute the estimate and the corresponding confidence interval for the marginal parameter by conditioning on the estimated variance components and integrating the RG fixed-effect estimate or individual CI over the distribution of reader. In the simplest scenario this leads to the CI for the marginal probability (e.g., sensitivity), which can then be compared to the CI based on the non-parametric variance estimate for U-statistic [48].

For example, consider using the simple no-covariate RG model A:

$$logit\left(P\left(Y_{ij}=1|\tilde{\beta}_j\right)\right)=\tilde{\mu}+\tilde{\beta}_j \text{ such that } \tilde{\beta}_j \sim N(0,\tilde{\sigma}_\beta^2):$$

Let $\hat{\tilde{\mu}}$ be the RG estimate of interest with CI limits as $\left(\widehat{\tilde{\mu}_l},\widehat{\tilde{\mu}_u}\right)$. The marginal sensitivity can be derived through numerical integration as follows:

$$\hat{p}^{RR}=\int_{-\infty}^{\infty} logit^{-1}\left(\hat{\tilde{\mu}}+\tilde{\beta}\right) f\left(\tilde{\beta}|\widehat{\tilde{\sigma}_\beta^2}\right)d\tilde{\beta} \text{ such that } \hat{\mu}^{RR}=logit(\hat{p}^{RR})$$

A naïve way to calculate this integral is by means of taking a simple average across readers:

$$\hat{p}^{RR}=\frac{1}{n_r}\sum_{j=1}^{n_r} logit^{-1}\left(\hat{\tilde{\mu}}+\widehat{\tilde{\beta}_J}\right)$$

For the 95% fully-marginal confidence limits:

$$\hat{p}_l{}^{RR}=\int logit^{-1}\left(\widehat{\tilde{\mu}_l}+\tilde{\beta}\right) f\left(\tilde{\beta}|\widehat{\tilde{\sigma}_\beta^2}\right)d\tilde{\beta} \text{ such that } \hat{\mu}_l{}^{RR}=logit\left(\hat{p}_l{}^{RR}\right) \text{ is the lower limit}$$

$\hat{p}_u{}^{RR} = \int logit^{-1}\left(\widehat{\tilde{\mu}_u} + \tilde{\beta}\right) f\left(\tilde{\beta}|\widehat{\tilde{\sigma}_\beta^2}\right) d\tilde{\beta}$ such that $\hat{\mu}_u{}^{RR} = logit\left(\hat{p}_u{}^{RR}\right)$ is the upper limit

Similar procedure can be used to estimate fully marginal parameters and their CI from GG model by performing integration over both the distribution of reader and subject as below:

$$\hat{p}^{RR} = \int\int logit^{-1}(\hat{\mu} + \alpha + \beta)\, f(\alpha|\hat{\sigma}_\alpha^2)\, f\left(\beta|\hat{\sigma}_\beta^2\right) d\alpha\, d\beta$$

$$\hat{p}_l{}^{RR} = \int\int logit^{-1}(\hat{\mu}_l + \alpha + \beta)\, f(\alpha|\hat{\sigma}_\alpha^2)\, f\left(\beta|\hat{\sigma}_\beta^2\right) d\alpha\, d\beta$$

$$\hat{p}_u{}^{RR} = \int\int logit^{-1}(\hat{\mu}_u + \alpha + \beta)\, f(\alpha|\hat{\sigma}_\alpha^2)\, f\left(\beta|\hat{\sigma}_\beta^2\right) d\alpha\, d\beta$$

We illustrate the above method using an actual multi-reader study [43] where we obtain the fully-marginal estimates using the RG model. Results are summarized in Table 16. We observed that the marginal estimate of fixed-effect $\tilde{\mu}$ obtained using direct integration i.e. 0.0644 was closer, although a bit smaller to the empirical marginal estimate i.e. 0.0675. The corresponding 95% CI was also comparable. Similar behavior was seen for estimates and CI on probability scale.

Table 16 Compute marginal estimates of fixed-effect and corresponding CI from RG estimates based on real dataset

				Logit Scale		Probability Scale	
Model	**Sample Size**	**Parameter**	**PL-RG model variance estimate (logit scale)**	**PL-RG model (Estimate ± SE) 95% t-based CI**	**Empirical estimates (Estimate ± SE) 95% z-based CI**	**PL-RG model (Estimate ± SE) 95% t-based CI**	**Empirical estimates (Estimate ± SE) 95% z-based CI**
A (modality=1)	$n_r = 7$ $n_1 = 55$	$\tilde{\mu}$ (average reader i.e. $\tilde{\beta} = 0$)	$\tilde{\sigma}_{\beta}^{2} = 0.4035$	0.0705 ± 0.3044 (-0.67, 0.81)	0.0763^{RG} 0.0675^{M} $(-0.53, 0.67)^{M}$	0.5176 ± 0.0759 (0.33, 0.69)	0.5190^{RG} $0.5168^{M} \pm 0.0745^{D}$ $(0.37, 0.66)^{M}$
		$Reader\ 1: logit(TPF)$		0.4338 ± 0.2674 (-0.22, 1.08)	0.4818 ± 0.2775^{B} (-0.06, 1.02)	0.6067 ± 0.0638 (0.44, 0.74)	0.6181 ± 0.0655^{B} (0.48, 0.73)
		$Reader\ 2: logit(TPF)$		-0.1589 ± 0.2626 (-0.80, 0.48)	-0.1823 ± 0.2708^{B} (-0.71, 0.34)	0.4603 ± 0.0652 (0.30, 0.61)	0.4545 ± 0.0671^{B} (0.33, 0.58)
		$Reader\ 3: logit(TPF)$		0.5717 ± 0.2715 (-0.09, 1.23)	0.6391 ± 0.2836^{B} (0.08, 1.19)	0.6391 ± 0.0626 (0.47, 0.77)	0.6545 ± 0.0641^{B} (0.52, 0.76)
		$Reader\ 4: logit(TPF)$		-0.7781 ± 0.2798 (-1.46, -0.09)	-0.8910 ± 0.2969^{B} (-1.47, -0.31)	0.3147 ± 0.0603 (0.18, 0.47)	0.2909 ± 0.0612^{B} (0.18, 0.42)
		$Reader\ 5: logit(TPF)$		0.9415 ± 0.2881 (0.23, 1.64)	1.0745 ± 0.3095^{B} (0.46, 1.68)	0.7194 ± 0.0581 (0.55, 0.83)	0.7454 ± 0.0587^{B} (0.61, 0.84)
		$Reader\ 6: logit(TPF)$		-0.3575 ± 0.2656 (-1.00, 0.29)	-0.4055 ± 0.2752^{B} (-0.94, 0.13)	0.4115 ± 0.0643 (0.26, 0.57)	0.4 ± 0.066^{B} (0.28, 0.53)
		$Reader\ 7: logit(TPF)$		-0.1589 ± 0.2626 (-0.80, 0.48)	-0.1823 ± 0.2708^{B} (-0.71, 0.34)	0.4603 ± 0.0652 (0.30, 0.61)	0.4545 ± 0.0671^{B} (0.33, 0.58)
Averaging technique (approximation to direct integration)				**Marginal FE estimate: 0.064 **Marginal CI: (-0.55, 0.68)	0.0675^{M} $(-0.53, 0.67)^{M}$ Length of CI=1.2	*Marginal FE estimate: 0.5160 *Marginal 95% CI: (0.36, 0.66)	$0.5168^{M} \pm 0.0745^{D}$ $(0.37, 0.66)^{M}$ Length of CI=0.29
Direct Integration				**Marginal FE estimate: 0.0644 **Marginal 95% CI: (-0.62 0.74) Length of CI: 1.36		*Marginal FE estimate: 0.5161 *Marginal 95% CI: (0.35, 0.67) Length of CI=0.32	

1. Marginal quantities on probability scale (*): have been derived by averaging the fixed effects, lower limit, upper limit across all readers and using direct integration as well
2. Marginal quantities on logit scale (**): have been computed taking the logit of the corresponding marginal quantities (*)
3. SE=Standard Error, M=Marginal, FE=Fixed-Effect, RG=Half-Marginal, CI=Confidence Interval
4. D=Variance calculated using Delong's Method (to account for variability between subjects and readers) [48]
5. B=binomial variance for each reader(observations within reader are independent) obtained using simple logistic models
6. CI are based on t-distribution with containment degrees of freedom
7. RG empirical estimates on logit scale for target parameter are obtained by averaging reader-specific logit scale estimates
8. RG empirical estimates for each reader can be computed in a model-free way using the data or by simple logistic regression. The resulting estimates will be same.

3.5 SIMULATION STUDY

3.5.1 Simulation Study Details

For evaluating performance of the RG models we use the same data as was generated for evaluation of GG models. However, the true fixed effects values for RG models as well as the variance components are different from the GG model parameters. For each considered simulation scenario, these true RG values under the RG model were computed from GG parameters using the formulations derived in Section 3.5.2. The true RG values were used to compute bias of the fixed effect and coverage rates.

To evaluate statistical properties of RG estimates, we used the same summary indices as in the investigation of GG models (Section 2.4.2). For each simulation, 95% confidence intervals were constructed using the default t-distribution with the containment degrees of freedom ($n_r - 1$) as well as the Satterthwaite degrees of freedom, which is the frequently recommended option.

3.5.2 Derivation of true RG fixed-effect parameters

In this section we outline derivation of the true RG fixed effect parameters for all four models (A, B, C and D). Since the resulting integrals don’t have closed form solutions, we need to numerically approximate the true RG values.

The true RG parameters should be smaller in magnitude than the original GG parameters, shrinking towards 0 in logit scale, or 0.5 in probability scale as demonstrated by Nehuas *et al.* (1991) [45].

3.5.2.1 Model A: Derived RG parameter $\tilde{\mu}$

RG model A is defined as follows:

$logit(\tilde{p}_j) = \tilde{\mu} + \tilde{\beta}_j$, where $\tilde{\mu} = E[logit(\tilde{p}_j)]$ **(3.2)**

$\tilde{p}_j$ is the reader-specific probability, which can be derived from the GG model A (namely $p_{ij} = logit^{-1}(\mu + \alpha_i + \beta_j))$ as follows:

$\tilde{p}_j = E[p_{ij}|\beta_j] = \int logit^{-1} \left(\mu + \alpha_i + \beta_j\right) f(\alpha_i) d\alpha_i$ **(3.3)**

By combining equations (3.2) and (3.3), we obtain the following expressions for $\tilde{\mu}$:

$$\tilde{\mu} = \int logit \left\{ \int logit^{-1} \left(\mu + \alpha_i + \beta_j\right) f(\alpha_i) d\alpha_i \right\} f(\beta_j)\, d\beta_j$$

Correspondingly, $\tilde{p} = logit^{-1}(\tilde{\mu})$.

3.5.2.2 Model B: Derived RG parameter $\tilde{\eta}$

RG model B is defined as follows:

$$logit(\tilde{p}_{jD}) = \tilde{\mu} + \tilde{\eta}_D + \tilde{\beta}_j + \tilde{\gamma}_{jD}$$

$\tilde{p}_{j(1)} = logit^{-1}(\tilde{\mu} + \tilde{\eta}_1 + \tilde{\beta}_j + \tilde{\gamma}_{j(1)})$, hence $\tilde{\eta}_1 = E[\, logit(\tilde{p}_{j(1)}\,)] - \tilde{\mu}$

$\tilde{p}_{j(-1)} = logit^{-1}(\tilde{\mu} + \tilde{\eta}_{-1} + \tilde{\beta}_j + \tilde{\gamma}_{j(-1)})$, hence $\tilde{\eta}_{-1} = E[logit(\tilde{p}_{j(-1)}\,)] - \tilde{\mu}$

Subtracting the terms, we get:

$\tilde{\eta} = E[logit(\tilde{p}_{j(1)}\,)] - E[\, logit(\tilde{p}_{j(-1)}\,)]$ **(3.4)**

$\tilde{p}_{jD}$ is the reader and truth specific probability, which can be derived from the GG model B as follows:

$$\tilde{p}_{jD} = E\left[p_{ij}\middle|\beta_j, \gamma_{jD}\right] = \int logit^{-1}\left(\mu + \eta_D + \alpha_i + \beta_j + \gamma_{jD}\right) f(\alpha_i)\, d\alpha_i \qquad (3.5)$$

By combining equations (3.4) and (3.5) we obtain the following expression for $\tilde{\eta}$:

$$\tilde{\eta} = \left[\iint logit\left\{\int logit^{-1}\left(\mu + \eta_1 + \alpha_i + \beta_j + \gamma_{j(1)}\right) f(\alpha_i)\, d\alpha_i\right\} f\left(\beta_j\right) f\left(\gamma_{j(1)}\right) d\beta_j\, d(\gamma_{j(1)})\right]$$

$$- \left[\iint logit\left\{\int logit^{-1}\left(\mu + \eta_{-1} + \alpha_i + \beta_j + \gamma_{j(-1)}\right) f(\alpha_i)\, d\alpha_i\right\} f\left(\beta_j\right) f\left(\gamma_{j(-1)}\right) d\beta_j\, d(\gamma_{j(-1)})\right]$$

3.5.2.3 Model C: Derived RG parameter $\tilde{\delta}$

RG model C is defined as follows:

$$logit\left(\tilde{p}_{jM}\right) = \tilde{\mu} + \tilde{\delta}_M + \tilde{\beta}_j + \tilde{\gamma}_{jM}$$

$\tilde{p}_{j(1)} = logit^{-1}(\tilde{\mu} + \tilde{\delta}_1 + \tilde{\beta}_j + \tilde{\gamma}_{j(1)})$, hence $\tilde{\delta}_1 = E\left[\, logit\left(\tilde{p}_{j(1)}\right)\right] - \tilde{\mu}$

$\tilde{p}_{j(-1)} = logit^{-1}(\tilde{\mu} + \tilde{\delta}_{(-1)} + \tilde{\beta}_j + \tilde{\gamma}_{j(-1)})$, hence $\tilde{\delta}_{(-1)} = E\left[\, logit\left(\tilde{p}_{j(-1)}\right)\right] - \tilde{\mu}$

Subtracting the terms, we get:

$$\tilde{\delta} = E\left[\, logit\left(\tilde{p}_{j(1)}\right)\right] - E\left[\, logit\left(\tilde{p}_{j(-1)}\right)\right] \qquad (3.6)$$

$\tilde{p}_{jM}$ is the reader and modality specific probability, which can be derived from the GG model C as follows:

$$\tilde{p}_{jM} = E\left[p_{ijM}\middle|\beta_j, \gamma_{jM}\right] = \int logit^{-1}\left(\mu + \delta_M + \alpha_i + \beta_j + \gamma_{jM}\right) f(\alpha_i)\, d\alpha_i \qquad (3.7)$$

Combining equations (3.6) and (3.7) we obtain the following expression for $\tilde{\delta}$:

$$\tilde{\delta} = \left[\iint logit\left\{\int logit^{-1}\left(\mu + \delta_1 + \alpha_i + \beta_j + \gamma_{j(1)}\right) f(\alpha_i)\, d\alpha_i\right\} f\left(\beta_j\right) f\left(\gamma_{j(1)}\right) d\beta_j\, d\gamma_{j(1)}\right]] -$$

$$\left[\iint logit\left\{\int logit^{-1}\left(\mu + \delta_{-1} + \alpha_i + \beta_j + \gamma_{j(-1)}\right) f(\alpha_i)\, d\alpha_i\right\} f\left(\beta_j\right) f\left(\gamma_{j(-1)}\right) d\beta_j\, d\gamma_{j(-1)}\right]$$

3.5.2.4 Model D: Derived RG parameter $\tilde{\tau}$

RG model D is defined as follows:

$$logit(\tilde{p}_{jX}) = \tilde{\mu} + \tilde{\tau} * X + \tilde{\beta}_j + \tilde{\gamma}_j * X$$

$logit(\tilde{p}_{j(x)}) = \tilde{\mu} + \tilde{\tau} * x + \tilde{\beta}_j + \tilde{\gamma}_j * x$, hence $\tilde{\mu} + \tilde{\tau} * x = E[logit(\tilde{p}_{j(x)})]$

$logit(\tilde{p}_{j(x+1)}) = \tilde{\mu} + \tilde{\tau} * (x+1) + \tilde{\beta}_j + \tilde{\gamma}_j * (x+1)$, hence $\tilde{\mu} + \tilde{\tau} * x + \tilde{\tau} = E[logit(\tilde{p}_{j(x+1)})]$

Subtracting the terms, we get:

$$\tilde{\tau} = E[logit(\tilde{p}_{j(x+1)})] - E[logit(\tilde{p}_{j(x})] \qquad (3.8)$$

Based on the GG model D, the reader-specific probabilities at a particular value of covariate can be expressed as follows:

$$\tilde{p}_{j(x)} = E[p_{ij}|\beta_j, \gamma_j, X = x] = \int logit^{-1}(\mu + \tau * x + \alpha_i + \beta_j + \gamma_j * x)\, f(\alpha_i) d\alpha_i \qquad (3.9)$$

$$\tilde{p}_{j(x+1)} = E[p_{ij}|\beta_j, \gamma_j, X = x+1] = \int logit^{-1}(\mu + \tau * (x+1) + \alpha_i + \beta_j + \gamma_j * (x + 1))\, f(\alpha_i) d\alpha_i \qquad (3.10)$$

Combining equations (3.9) and (3.10), we obtain the following expression for the estimation target in RG model D:

$$\tilde{\tau} = \iint logit\left\{\int logit^{-1}(\mu + \tau * (x+1) + \alpha_i + \beta_j + \gamma_j * (x+1))\, f(\alpha_i) d\alpha_i\right\} f(\beta_j) f(\gamma_j) d(\beta_j) d(\gamma_j)$$
$$- \iint logit\left\{\int logit^{-1}(\mu + \tau * (x) + \alpha_i + \beta_j + \gamma_j * (x))\, f(\alpha_i) d\alpha_i\right\} f(\beta_j) f(\gamma_j) d(\beta_j) d(\gamma_j)$$

For the simulation study the targeted value of the parameter was estimated at X = 0 and X = 1 (i.e. unit increment).

3.5.3 Simulation Study Results

Model A:

For the covariate free setting (model A, Table 17, Table 18, Table 19), PL-RG model led to rather accurate estimates of the intercept and its standard error across the considered scenarios. The standardized bias of the fixed effect estimate was well below 1 and tended to be larger for larger variance components (Table 18, Table 19). However, the fixed effect estimate illustrated a bothersome trend to increasingly underestimate the true value for larger sample sizes, which may be an indication of the inconsistency of the pseudo-likelihood estimates for the considered models. Nevertheless, the magnitude of the bias was small enough to be inconsequential for the coverage of confidence intervals. The confidence intervals tended to be overly conservative when the default containment degrees of freedom was used, but had nearly ideal coverage with the Satterthwaite approximation. In this simple model we observed very few convergence problems.

Model B:

Results for model B [Table 20, Table 21, Table 22] showed a similar behavior as in model A, with conservative containment-based confidence intervals which were substantially improved using the Satterthwaite degrees of freedom. The standardized bias of fixed effect estimator was well below 1, but both raw and standardized bias demonstrated a bothersome downward trend with increasing sample size. Nevertheless, this did not have any substantial effect on the coverage of the confidence intervals. The relative bias of SE estimate was relatively small across the considered scenarios. Convergence problems were more frequent with this model than with model A, but were still within 3%.

Model C:

When considering RG model C, i.e. comparing sensitivity levels across the two modalities [Table 23, Table 24, Table 25], we observed accurate fixed effect estimates but noticeably underestimated variance. The standardized bias was within 1 across all settings. The relative bias of estimated SE, however, was ranged up to 14% for scenarios when the true probabilities were small, especially in presence of small variance components. The underestimation of variance appeared to have a slight, but noticeable effect on the coverage of confidence interval, especially those based on the Satterthwaite degrees of freedom. Very few convergence problems were noted for this model.

Model D:

For model D [Table 26, Table 27, Table 28], the 95% CIs had coverage above the nominal level and were less conservative under the Satterthwaite approximation across all parameter combinations. The standardized bias was small but increased for large value of true slope, especially when variance components were small. The bias of the SE estimates was also relatively small. With this model we also observed more frequent convergence issues, especially in presence of larger variance components.

Table 17 Simulation Results for RG Model A (small magnitude of variance components)

		$p^{GG}=0.1$ **$\mu^{GG}=-2.20$** **$\mu^{RG}=-1.87$**		**$p^{GG}=0.5$** **$\mu^{GG}=0$** **$\mu^{RG}=0$**		**$p^{GG}=0.7$** **$\mu^{GG}=0.85$** **$\mu^{RG}=0.70$**	
		n_1		**n_1**		**n_1**	
	n_r	**55**	**100**	**55**	**100**	**55**	**100**
Coverage (%) (c/s)	**5**	99/96	98/94	99/96	99/95	99/95	98/96
	10	97/95	96/94	97/96	97/95	97/95	97/95
Bias	**5**	-0.02	0	-0.002	-0.004	-0.003	-0.006
(SE)		(0.254)	(0.201)	(0.194)	(0.162)	(0.204)	(0.168)
	10	0.007	0.009	-0.003	-0.003	-0.011	-0.01
		(0.195)	(0.154)	(0.159)	(0.13)	(0.163)	(0.133)
SB	**5**	-0.08	0	-0.01	-0.03	-0.02	-0.03
(RBS (%))		(-2)	(-1)	(0)	(4)	(1)	(1)
	10	0.03	0.05	-0.02	-0.02	-0.07	-0.07
		(-4)	(-5)	(1)	(-1)	(-2)	(-2)
Sim Used	**5**	1000	1000	1000	1000	1000	1000
	10	999	999	1000	1000	1000	1000

1. Simulation parameters: Subject Variance=1, Reader Variance=0.1
2. GG="Subject-Specific" parameter, RG="Half-Marginal" parameter
3. Coverage (%)=Estimated 95% coverage of t-based confidence interval; c=containment df; s=sattherthwaite df
4. Bias=Bias of fixed effect parameter estimate; SE=Average Estimated Standard Error for fixed effect parameter estimate
5. SB=Standardized bias of fixed effect parameter estimate (using MC Standard Deviation); RBS (%)=Relative bias of estimated standard error for fixed-effect parameter estimate (using MC Standard Deviation)
6. Sim Used=Used simulations: (1) Non-zero positive standard errors (2) PROC GLIMMIX convergence criteria satisfied

Table 18 Simulation Results for RG Model A (medium magnitude of variance components)

		p^{GG}=0.1 **μ^{GG}=-2.20** **μ^{RG}=-1.66**		**p^{GG}=0.5** **μ^{GG}=0** **μ^{RG}=0**		**p^{GG}=0.7** **μ^{GG}=0.85** **μ^{RG}=0.63**	
		n_1		**n_1**		**n_1**	
	n_r	**55**	**100**	**55**	**100**	**55**	**100**
Coverage (%) (c/s)	**5**	98/94	97/95	98/95	97/95	98/95	98/95
	10	96/93	95/94	97/96	96/95	97/96	96/95
Bias	**5**	-0.001	0.009	-0.003	-0.007	-0.009	-0.009
(SE)		(0.364)	(0.328)	(0.313)	(0.292)	(0.323)	(0.297)
	10	0.059	0.036	-0.004	-0.004	-0.029	-0.018
		(0.275)	(0.247)	(0.249)	(0.224)	(0.252)	(0.228)
SB	**5**	0	0.03	-0.01	-0.02	-0.03	-0.03
(RBS (%))		(-5)	(-3)	(-2)	(-1)	(-3)	(-1)
	10	0.2	0.14	-0.01	-0.02	-0.11	-0.08
		(-4)	(-4)	(2)	(0)	(0)	(-1)
Sim Used	**5**	1000	999	1000	1000	1000	1000
	10	1000	1000	1000	999	1000	1000

1. Simulation parameters: Subject Variance=2, Reader Variance=0.7
2. GG="Subject-Specific" parameter, RG="Half-Marginal" parameter
3. Coverage (%)=Estimated 95% coverage of t-based confidence interval; c=containment df; s=sattherthwaite df
4. Bias=Bias of fixed effect parameter estimate; SE=Average Estimated Standard Error for fixed effect parameter estimate
5. SB=Standardized bias of fixed effect parameter estimate (using MC Standard Deviation); RBS (%)=Relative bias of estimated standard error for fixed-effect parameter estimate (using MC Standard Deviation)
6. Sim Used=Used simulations: (1) Non-zero positive standard errors (2) PROC GLIMMIX convergence criteria satisfied

Table 19 Simulation Results for RG Model A (large magnitude of variance components)

		p^{GG}=0.1 μ^{GG}=-2.20 μ^{RG}=-1.52		p^{GG}=0.5 μ^{GG}=0 μ^{RG}=-0		p^{GG}=0.7 μ^{GG}=0.85 μ^{RG}=0.57	
		n_1		n_1		n_1	
	n_r	**55**	**100**	**55**	**100**	**55**	**100**
Coverage (%) (c/s)	**5**	97/95	96/94	97/95	97/95	97/95	97/95
	10	96/94	94/94	97/96	96/95	97/96	96/95
Bias (SE)	**5**	0.01 (0.449)	0.01 (0.415)	0 (0.398)	-0.007 (0.376)	-0.013 (0.405)	-0.014 (0.382)
	10	0.072 (0.332)	0.049 (0.308)	-0.006 (0.31)	-0.004 (0.286)	-0.035 (0.314)	-0.019 (0.289)
SB (RBS (%))	**5**	0.02 (-5)	0.02 (-4)	0 (-2)	-0.02 (-2)	-0.03 (-2)	-0.03 (-2)
	10	0.21 (-4)	0.15 (-3)	-0.02 (3)	-0.01 (0)	-0.11 (2)	-0.07 (-2)
Sim Used	**5**	1000	999	1000	1000	1000	1000
	10	1000	999	1000	999	1000	999

1. Simulation parameters: Subject Variance=3, Reader Variance=1.5
2. GG="Subject-Specific" parameter, RG="Half-Marginal" parameter
3. Coverage (%)=Estimated 95% coverage of t-based confidence interval; c=containment df; s=sattherthwaite df
4. Bias=Bias of fixed effect parameter estimate; SE=Average Estimated Standard Error for fixed effect parameter estimate
5. SB=Standardized bias of fixed effect parameter estimate (using MC Standard Deviation); RBS (%)=Relative bias of estimated standard error for fixed-effect parameter estimate (using MC Standard Deviation)
6. Sim Used=Used simulations: (1) Non-zero positive standard errors (2) PROC GLIMMIX convergence criteria satisfied

Table 20 Simulation Results for RG Model B (small magnitude of variance components)

			TPF^{GG}=0.5, FPF^{GG}=0.03 η^{GG}=3.48 η^{RG}=3.05 n_0		TPF^{GG}=0.3, FPF^{GG}=0.9 η^{GG}=-3.04 η^{RG}=-2.58 n_0		TPF^{GG}=0.6, FPF^{GG}=0.6 η^{GG}=0 η^{RG}=0 n_0	
	n_r	n_1	100	175	100	175	100	175
Coverage (%) (c/s)	5	55	99/95	98/94	99/96	98/95	99/96	99/95
	5	100	99/95	98/94	98/95	98/95	99/96	98/94
	10	55	97/95	96/94	97/94	97/94	97/94	97/95
	10	100	97/95	97/95	97/95	96/94	98/97	97/95
Bias	5	55	-0.026	-0.023	0.021	0.02	-0.006	0.004
(SE)			(0.321)	(0.272)	(0.269)	(0.245)	(0.243)	(0.228)
	5	100	-0.025	-0.023	0.018	0.009	0.002	-0.001
			(0.31)	(0.258)	(0.248)	(0.22)	(0.215)	(0.199)
	10	55	-0.063	-0.04	0.048	0.033	-0.006	-0.001
			(0.25)	(0.21)	(0.215)	(0.195)	(0.197)	(0.187)
	10	100	-0.062	-0.049	0.046	0.029	-0.001	-0.008
			(0.245)	(0.201)	(0.198)	(0.175)	(0.174)	(0.162)
SB	5	55	-0.08	-0.08	0.08	0.08	-0.02	0.02
(RBS (%))			(1)	(-5)	(3)	(-2)	(1)	(1)
	5	100	-0.08	-0.08	0.07	0.04	0.01	0
			(2)	(-5)	(1)	(-3)	(-2)	(-3)
	10	55	-0.26	-0.19	0.22	0.17	-0.03	-0.01
			(2)	(-2)	(0)	(-2)	(0)	(1)
	10	100	-0.27	-0.25	0.24	0.16	-0.01	-0.05
			(6)	(4)	(3)	(-2)	(3)	(3)
Sim Used	5	55	991	996	999	999	1000	999
	5	100	988	994	999	998	1000	999
	10	55	977	987	996	996	999	999
	10	100	981	995	997	998	998	996

1. Simulation parameters: Subject Variance=1, Reader Variance=0.25, Reader*Truth Variance=0.07
2. GG="Subject-Specific" parameter, RG="Half-Marginal" parameter
3. Coverage (%)=Estimated 95% coverage of t-based confidence interval; c=containment df; s=sattherthwaite df
4. Bias=Bias of fixed effect parameter estimate; SE=Average Estimated Standard Error for fixed effect parameter estimate
5. SB=Standardized bias of fixed effect parameter estimate (using MC Standard Deviation); RBS (%)=Relative bias of estimated standard error for fixed-effect parameter estimate (using MC Standard Deviation)
6. Sim Used=Used simulations: (1) Non-zero positive standard errors (2) PROC GLIMMIX convergence criteria satisfied

Table 21 Simulation Results for RG Model B (medium magnitude of variance components)

			TPF^{GG}=0.5, FPF^{GG}=0.03 **η^{GG}=3.48** **η^{RG}=2.75** **n_0**		**TPF^{GG}=0.3, FPF^{GG}=0.9** **η^{GG}=-3.04** **η^{RG}=-2.31** **n_0**		**TPF^{GG}=0.6, FPF^{GG}=0.6** **η^{GG}=0** **η^{RG}=0** **n_0**	
	n_r	**n_1**	**100**	**175**	**100**	**175**	**100**	**175**
Coverage (%) (c/s)	**5**	**55**	99/95	98/93	99/96	99/95	99/96	99/95
	5	**100**	98/95	98/94	98/94	98/94	98/95	98/94
	10	**55**	96/94	96/93	96/94	97/95	97/95	97/95
	10	**100**	96/94	96/95	97/94	97/95	98/96	98/96
Bias	**5**	**55**	-0.041	-0.035	0.03	0.023	-0.003	0.004
(SE)			(0.334)	(0.298)	(0.303)	(0.279)	(0.278)	(0.264)
	5	**100**	-0.039	-0.037	0.025	0.018	0.002	0.001
			(0.322)	(0.278)	(0.278)	(0.251)	(0.25)	(0.235)
	10	**55**	-0.099	-0.067	0.064	0.043	-0.005	-0.001
			(0.27)	(0.238)	(0.247)	(0.229)	(0.233)	(0.221)
	10	**100**	-0.1	-0.079	0.062	0.037	0.001	-0.008
			(0.26)	(0.225)	(0.228)	(0.206)	(0.206)	(0.194)
SB	**5**	**55**	-0.12	-0.11	0.1	0.08	-0.01	0.01
(RBS (%))			(-1)	(-7)	(0)	(-2)	(0)	(0)
	5	**100**	-0.12	-0.13	0.09	0.07	0.01	0
			(0)	(-7)	(-1)	(-3)	(-1)	(-3)
	10	**55**	-0.37	-0.27	0.26	0.19	-0.02	-0.01
			(0)	(-4)	(0)	(0)	(2)	(2)
	10	**100**	-0.4	-0.35	0.28	0.18	0.01	-0.04
			(3)	(1)	(3)	(0)	(4)	(4)
Sim Used	**5**	**55**	994	996	999	1000	1000	1000
	5	**100**	997	994	1000	996	1000	1000
	10	**55**	976	983	997	997	999	999
	10	**100**	995	984	995	997	999	996

1. Simulation parameters: Subject Variance=1.9, Reader Variance=0.5, Reader*Truth Variance=0.14
2. GG="Subject-Specific" parameter, RG="Half-Marginal" parameter
3. Coverage (%)=Estimated 95% coverage of t-based confidence interval; c=containment df; s=satherthwaite df
4. Bias=Bias of fixed effect parameter estimate; SE=Average Estimated Standard Error for fixed effect parameter estimate
5. SB=Standardized bias of fixed effect parameter estimate (using MC Standard Deviation); RBS (%)=Relative bias of estimated standard error for fixed-effect parameter estimate (using MC Standard Deviation)
6. Sim Used=Used simulations: (1) Non-zero positive standard errors (2) PROC GLIMMIX convergence criteria satisfied

Table 22 Simulation Results for RG Model B (large magnitude of variance components)

			TPF^{GG}=0.5, FPF^{GG}=0.03		TPF^{GG}=0.3, FPF^{GG}=0.9		TPF^{GG}=0.6, FPF^{GG}=0.6	
			η^{GG}=3.48		η^{GG}=-3.04		η^{GG}=0	
			η^{RG}=2.48		η^{RG}=-2.09		η^{RG}=0	
			n_0		n_0		n_0	
	n_r	n_1	100	175	100	175	100	175
Coverage (%) (c/s)	**5**	**55**	98/94	97/94	98/95	98/94	98/96	98/94
	5	**100**	97/94	97/94	97/94	96/94	98/95	96/94
	10	**55**	95/93	95/93	96/94	96/95	98/96	98/96
	10	**100**	96/94	95/94	97/95	97/96	98/96	98/96
Bias	**5**	**55**	-0.069	-0.046	0.035	0.033	-0.003	0.003
(SE)			(0.412)	(0.393)	(0.391)	(0.372)	(0.37)	(0.36)
	5	**100**	-0.07	-0.054	0.037	0.024	0.001	0
			(0.398)	(0.373)	(0.373)	(0.349)	(0.347)	(0.336)
	10	**55**	-0.128	-0.085	0.084	0.063	-0.006	-0.003
			(0.33)	(0.308)	(0.315)	(0.301)	(0.304)	(0.292)
	10	**100**	-0.134	-0.098	0.081	0.052	0.003	-0.007
			(0.317)	(0.293)	(0.296)	(0.279)	(0.28)	(0.267)
SB	**5**	**55**	-0.16	-0.11	0.09	0.08	-0.01	0.01
(RBS (%))			(-3)	(-4)	(-2)	(-5)	(-2)	(-1)
	5	**100**	-0.17	-0.14	0.1	0.07	0	0
			(-3)	(-5)	(-2)	(-6)	(-3)	(-4)
	10	**55**	-0.39	-0.27	0.27	0.22	-0.02	-0.01
			(1)	(-1)	(1)	(2)	(3)	(4)
	10	**100**	-0.45	-0.34	0.28	0.19	0.01	-0.03
			(5)	(0)	(3)	(5)	(4)	(6)
Sim Used	**5**	**55**	996	989	999	998	1000	998
	5	**100**	989	987	999	996	1000	999
	10	**55**	982	974	999	993	1000	998
	10	**100**	971	980	995	994	1000	998

1. Simulation parameters: Subject Variance=3, Reader Variance=1, Reader*Truth Variance=0.5
2. GG="Subject-Specific" parameter, RG="Half-Marginal" parameter
3. Coverage (%)=Estimated 95% coverage of t-based confidence interval; c=containment df; s=sattherthwaite df
4. Bias=Bias of fixed effect parameter estimate; SE=Average Estimated Standard Error for fixed effect parameter estimate
5. SB=Standardized bias of fixed effect parameter estimate (using MC Standard Deviation); RBS (%)=Relative bias of estimated standard error for fixed-effect parameter estimate (using MC Standard Deviation)
6. Sim Used=Used simulations: (1) Non-zero positive standard errors (2) PROC GLIMMIX convergence criteria satisfied

Table 23 Simulation Results for RG Model C (small magnitude of variance components)

		p_1^{GG}=0.5, p_2^{GG}=0.6 **δ^{GG}=-0.41** **δ^{RG}=-0.30** **n_1**		**p_1^{GG}=0.1, p_2^{GG}=0.7** **δ^{GG}=-3.04** **δ^{RG}=-2.31** **n_1**		**p_1^{GG}=0.8, p_2^{GG}=0.8** **δ^{GG}=0** **δ^{RG}=0** **n_1**	
	n_r	**55**	**100**	**55**	**100**	**55**	**100**
Coverage (%) (c/s)	**5**	98/96	96/94	97/93	96/93	97/94	97/94
	10	95/95	96/96	93/92	93/93	96/95	95/95
Bias	**5**	-0.007	-0.008	0.001	0.005	-0.007	-0.009
(SE)		(0.229)	(0.201)	(0.267)	(0.223)	(0.243)	(0.211)
	10	0	-0.002	0.017	0.013	-0.005	-0.002
		(0.161)	(0.146)	(0.187)	(0.162)	(0.173)	(0.153)
SB	**5**	-0.03	-0.04	0	0.02	-0.03	-0.04
(RBS (%))		(3)	(-4)	(-7)	(-14)	(-1)	(-7)
	10	0	-0.01	0.08	0.07	-0.03	-0.01
		(-3)	(0)	(-13)	(-11)	(-1)	(-1)
Sim Used	**5**	1000	1000	1000	1000	1000	1000
	10	999	999	999	997	1000	998

1. Simulation parameters: Subject Variance=1.9, Reader Variance=0.5, Reader*Modality Variance=0.14
2. GG="Subject-Specific" parameter, RG="Half-Marginal" parameter
3. Coverage (%)=Estimated 95% coverage of t-based confidence interval; c=containment df; s=sattherthwaite df
4. Bias=Bias of fixed effect parameter estimate; SE=Average Estimated Standard Error for fixed effect parameter estimate
5. SB=Standardized bias of fixed effect parameter estimate (using MC Standard Deviation); RBS (%)=Relative bias of estimated standard error for fixed-effect parameter estimate (using MC Standard Deviation)
6. Sim Used=Used simulations: (1) Non-zero positive standard errors (2) PROC GLIMMIX convergence criteria satisfied

Table 24 Simulation Results for RG Model C (medium magnitude of variance components)

		p_1^{GG}=0.5, p_2^{GG}=0.6 δ^{GG}=-0.41 δ^{RG}=-0.26 n_1		p_1^{GG}=0.1, p_2^{GG}=0.7 δ^{GG}=-3.04 δ^{RG}=-1.96 n_1		p_1^{GG}=0.8, p_2^{GG}=0.8 δ^{GG}=0 δ^{RG}=0 n_1	
	n_r	55	100	55	100	55	100
Coverage (%) (c/s)	**5**	97/96	95/95	96/94	95/94	96/94	96/95
	10	96/95	96/96	92/92	93/93	95/94	96/96
Bias	**5**	-0.01	-0.008	0.002	-0.001	-0.003	-0.006
(SE)		(0.278)	(0.259)	(0.302)	(0.273)	(0.283)	(0.267)
	10	-0.001	-0.001	0.019	0.013	-0.003	0.001
		(0.199)	(0.19)	(0.216)	(0.203)	(0.205)	(0.194)
SB	**5**	-0.04	-0.03	0.01	0	-0.01	-0.02
(RBS (%))		(-1)	(-5)	(-10)	(-13)	(-7)	(-7)
	10	-0.01	-0.01	0.08	0.06	-0.01	0
		(-1)	(1)	(-15)	(-11)	(-4)	(-2)
Sim Used	**5**	1000	998	998	999	1000	1000
	10	1000	1000	1000	999	1000	994

1. Simulation parameters: Subject Variance=3.71, Reader Variance=0.8672, Reader*Modality Variance=0.4
2. GG="Subject-Specific" parameter, RG="Half-Marginal" parameter
3. Coverage (%)=Estimated 95% coverage of t-based confidence interval; c=containment df; s=sattherthwaite df
4. Bias=Bias of fixed effect parameter estimate; SE=Average Estimated Standard Error for fixed effect parameter estimate
5. SB=Standardized bias of fixed effect parameter estimate (using MC Standard Deviation); RBS (%)=Relative bias of estimated standard error for fixed-effect parameter estimate (using MC Standard Deviation)
6. Sim Used=Used simulations: (1) Non-zero positive standard errors (2) PROC GLIMMIX convergence criteria satisfied

Table 25 Simulation Results for RG Model C (large magnitude of variance components)

		p_1^{GG}=0.5, p_2^{GG}=0.6 δ^{GG}=-0.41 δ^{RG}=-0.24 n_1		p_1^{GG}=0.1, p_2^{GG}=0.7 δ^{GG}=-3.04 δ^{RG}=-1.81 n_1		p_1^{GG}=0.8, p_2^{GG}=0.8 δ^{GG}=0 δ^{RG}=0 n_1	
	n_r	55	100	55	100	55	100
Coverage (%) (c/s)	5	97/96	96/96	96/95	94/94	96/95	96/96
	10	96/96	97/97	94/94	95/95	95/95	96/96
Bias	5	-0.01	-0.009	0.001	0.005	-0.011	-0.005
(SE)		(0.377)	(0.363)	(0.394)	(0.373)	(0.381)	(0.372)
	10	-0.003	0.001	0.023	0.015	-0.004	-0.002
		(0.274)	(0.267)	(0.289)	(0.278)	(0.279)	(0.27)
SB	5	-0.03	-0.02	0	0.01	-0.03	-0.01
(RBS (%))		(-5)	(-5)	(-10)	(-10)	(-8)	(-6)
	10	-0.01	0	0.07	0.05	-0.01	-0.01
		(-1)	(0)	(-7)	(-6)	(-4)	(-2)
Sim Used	5	1000	1000	1000	1000	1000	1000
	10	999	999	995	997	998	997

1. Simulation parameters: Subject Variance=5, Reader Variance=2, Reader*Modality Variance=1
2. GG="Subject-Specific" parameter, RG="Half-Marginal" parameter
3. Coverage (%)=Estimated 95% coverage of t-based confidence interval; c=containment df; s=sattherthwaite df
4. Bias=Bias of fixed effect parameter estimate; SE=Average Estimated Standard Error for fixed effect parameter estimate
5. SB=Standardized bias of fixed effect parameter estimate (using MC Standard Deviation); RBS (%)=Relative bias of estimated standard error for fixed-effect parameter estimate (using MC Standard Deviation)
6. Sim Used=Used simulations: (1) Non-zero positive standard errors (2) PROC GLIMMIX convergence criteria satisfied

Table 26 Simulation Results for RG Model D (small magnitude of variance components)

		$\mu^{GG}=0$		$\mu^{GG}=0$		$\mu^{GG}=0$		$\mu^{GG}=0$	
		$\tau^{GG}=0$		$\tau^{GG}=0.03$		$\tau^{GG}=0.06$		$\tau^{GG}=0.10$	
		$\tau^{RG}=0$		$\tau^{RG}=0.025$		$\tau^{RG}=0.050$		$\tau^{RG}=0.083$	
		n_1		n_1		n_1		n_1	
	n_r	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**
Coverage (%) (c/s)	**5**	98/96	98/95	99/97	98/95	98/96	98/95	99/96	98/95
	10	98/96	97/96	97/96	97/96	97/95	97/96	96/95	97/95
Bias	**5**	0	0	0	0	0.001	0.001	0.005	0.003
(SE)		(0.009)	(0.008)	(0.01)	(0.009)	(0.011)	(0.009)	(0.014)	(0.012)
	10	0	0	-0.001	-0.001	-0.001	-0.001	0.001	0.001
		(0.007)	(0.006)	(0.007)	(0.007)	(0.008)	(0.007)	(0.01)	(0.008)
SB	**5**	0	-0.02	-0.02	-0.02	0.07	0.06	0.3	0.27
(RBS (%))		(4)	(-3)	(2)	(-2)	(0)	(-4)	(-6)	(-6)
	10	0.03	0	-0.1	-0.1	-0.12	-0.08	0.12	0.13
		(6)	(3)	(5)	(2)	(1)	(1)	(-7)	(-3)
Sim Used	**5**	999	1000	997	998	993	993	938	972
	10	996	997	1000	999	982	993	917	930

1. Simulation parameters: Subject Variance=1, Reader Variance=0.1, Reader*LesionSize Variance=0.0004
2. Continuous variable (LesionSize)~Unif(1,100) and has been centered at 50.5 mm during simulation and fitting process
3. GG="Subject-Specific" parameter, RG="Half-Marginal" parameter
4. Coverage (%)=Estimated 95% coverage of t-based confidence interval; c=containment df; s=sattherthwaite df
5. Bias=Bias of fixed effect parameter estimate; SE=Average Estimated Standard Error for fixed effect parameter estimate
6. SB=Standardized bias of fixed effect parameter estimate (using MC Standard Deviation); RBS (%)=Relative bias of estimated standard error for fixed-effect parameter estimate (using MC Standard Deviation)
7. Sim Used=Used simulations: (1) Non-zero positive standard errors (2) PROC GLIMMIX convergence criteria satisfied

Table 27 Simulation Results for RG Model D (medium magnitude of variance components)

		$\mu^{GG}=0$ $\tau^{GG}=0$ $\tau^{RG}=0$		$\mu^{GG}=0$ $\tau^{GG}=0.03$ $\tau^{RG}=0.022$		$\mu^{GG}=0$ $\tau^{GG}=0.06$ $\tau^{RG}=0.044$		$\mu^{GG}=0$ $\tau^{GG}=0.10$ $\tau^{RG}=0.073$	
		n_1		n_1		n_1		n_1	
	n_r	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**
Coverage (%) (c/s)	**5**	98/96	98/95	99/97	97/95	98/96	97/95	99/95	98/95
	10	98/97	97/97	97/97	97/96	97/96	97/96	95/94	97/95
Bias	**5**	0	0	0	0	0	0	0.003	0.002
(SE)		(0.012)	(0.011)	(0.012)	(0.011)	(0.013)	(0.012)	(0.016)	(0.014)
	10	0	0	-0.001	-0.001	-0.002	-0.001	-0.001	0
		(0.009)	(0.009)	(0.009)	(0.009)	(0.01)	(0.009)	(0.011)	(0.01)
SB	**5**	0	-0.01	-0.01	-0.01	0.02	0.04	0.21	0.17
(RBS (%))		(4)	(-2)	(4)	(-3)	(0)	(-5)	(-7)	(-4)
	10	0.02	0.01	-0.12	-0.08	-0.21	-0.13	-0.07	-0.01
		(10)	(6)	(7)	(4)	(5)	(1)	(-9)	(-4)
Sim Used	**5**	999	1000	995	998	989	994	943	972
	10	981	995	982	993	972	979	900	918

1. Simulation parameters: Subject Variance=2, Reader Variance=0.2, Reader*LesionSize Variance=0.001
2. Continuous variable (LesionSize)~Unif(1,100) and has been centered at 50.5 mm during simulation and fitting process
3. GG="Subject-Specific" parameter, RG="Half-Marginal" parameter
4. Coverage (%)=Estimated 95% coverage of t-based confidence interval; c=containment df; s=sattherthwaite df
5. Bias=Bias of fixed effect parameter estimate; SE=Average Estimated Standard Error for fixed effect parameter estimate
6. SB=Standardized bias of fixed effect parameter estimate (using MC Standard Deviation); RBS (%)=Relative bias of estimated standard error for fixed-effect parameter estimate (using MC Standard Deviation)
7. Sim Used=Used simulations: (1) Non-zero positive standard errors (2) PROC GLIMMIX convergence criteria satisfied

Table 28 Simulation Results for RG Model D (large magnitude of variance components)

		μ^{GG}=0 τ^{GG}=0 τ^{RG}=0		μ^{GG}=0 τ^{GG}=0.03 τ^{RG}=0.020		μ^{GG}=0 τ^{GG}=0.06 τ^{RG}=0.040		μ^{GG}=0 τ^{GG}=0.10 τ^{RG}=0.067	
		n_1		n_1		n_1		n_1	
	n_r	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**
Coverage (%) (c/s)	**5**	97/96	97/96	97/96	96/96	97/96	97/96	97/96	97/95
	10	97/97	97/97	98/97	96/96	97/96	96/96	96/96	96/96
Bias	**5**	0	0	0	0	0.001	0.001	0.002	0.002
(SE)		(0.032)	(0.031)	(0.032)	(0.031)	(0.032)	(0.031)	(0.033)	(0.032)
	10	0	0	-0.001	0	-0.001	-0.001	-0.002	-0.001
		(0.023)	(0.022)	(0.023)	(0.022)	(0.023)	(0.023)	(0.023)	(0.023)
SB	**5**	0	0.01	0	0	0.02	0.03	0.05	0.05
(RBS (%))		(1)	(-1)	(1)	(-3)	(0)	(-5)	(-3)	(-6)
	10	0	0.01	-0.02	0	-0.06	-0.03	-0.09	-0.05
		(9)	(4)	(7)	(4)	(3)	(2)	(-2)	(-2)
Sim Used	**5**	984	981	979	975	969	975	961	954
	10	919	935	905	946	913	939	916	922

1. Simulation parameters: Subject Variance=3, Reader Variance=1, Reader*LesionSize Variance=0.01
2. Continuous variable (LesionSize)~Unif(1,100) and has been centered at 50.5 mm during simulation and fitting process
3. GG="Subject-Specific" parameter, RG="Half-Marginal" parameter
4. Coverage (%)=Estimated 95% coverage of t-based confidence interval; c=containment df; s=sattherthwaite df
5. Bias=Bias of fixed effect parameter estimate; SE=Average Estimated Standard Error for fixed effect parameter estimate
6. SB=Standardized bias of fixed effect parameter estimate (using MC Standard Deviation); RBS (%)=Relative bias of estimated standard error for fixed-effect parameter estimate (using MC Standard Deviation)
7. Sim Used=Used simulations: (1) Non-zero positive standard errors (2) PROC GLIMMIX convergence criteria satisfied

3.6 SUMMARY AND DISCUSSION

The RG model estimated used the PL approach demonstrated the ability to perform valid statistical inferences in all modeling settings and across a range of parameter configurations. The usage of Satterthwaite containment degrees of freedom led to slight improvement in the coverage rates based on using the t-reference distribution as compared to using the default containment degrees of freedom. The coverage was either adequate, although often conservative for both approaches. Using data from an actual multi-reader study of diagnostic imaging, we illustrated the ability of the half-marginal model to provide estimates of parameters that are relevant in multi-reader studies that agree with empirical estimates. In addition, the RG model is easy to implement (e.g., using PROC GLIMMIX, SAS) and leads to substantial reduction in computation time compared to more standard GG models.

As other modeling techniques, the proposed RG model has limitations. First, the RG model can be criticized for its non-probabilistic nature. Second, RG model necessitates the use of the pseudo-likelihood (PL) estimation approach. Although our results indicate very good performance of PL-RG estimates in typical multi-reader settings, there could still be a lingering concern about possible large-sample reliability problems based on trends we observed in simulation study and the results discussed in other articles for simple binary clustered data.

Overall, for data with structure and size typical for multi-reader studies, half-marginal models provides a convenient and flexible framework for analyzing multi-reader studies. Being GLMMs, they allow incorporating both continuous and categorical types of covariates into the analysis. Usefulness of the half-marginal modeling approach can be further increased by developing robust alternatives to the pseudo-likelihood estimation technique.

4.0 AN EXPLICIT APPROACH FOR ESTIMATING HALF-MARGINAL GLMM FOR ANALYZING CROSS-CORRELATED BINARY DATA FROM MULTI-READER STUDIES OF DIAGNOSTIC ACCURACY

Previously, we proposed a half-marginal model that is particularly suited for binary data from fully-crossed multi-reader studies. The half-marginal model offers parameters that are more interpretable in practice, enables estimation of both overall and reader-specific estimates, and provides estimates that are more in agreement with non-parametric estimates obtainable from raw data. However, a half-marginal model is often faced with the criticism of being artificial due to its non-probabilistic nature and of relying solely on pseudo-likelihood techniques, whose validity for binary data is questionable. To address these criticisms we have developed a new semi-parametric approach for estimating half-marginal model which offers straightforward estimation based on probability models for analyzing cross-correlated data from multi-reader diagnostic imaging studies.

4.1 INTRODUCTION

One of the essential components of the proposed work is focused on developing the variance estimator for the estimated model parameters based on cross-correlated data. This task is usually not straightforward [16]. The general difficulty stems from the need to reflect the two sources of variability (in our application: subjects and readers) and account for correlation between the observations. A frequent remedy is to use bootstrap resampling [54, 55] to obtain variance estimates. However, resampling approaches for iteratively fit models can be very computer intensive.

One of the basic examples of the variance estimator for cross-correlated data is the variance estimator of the two-sample U-statistic. Based on the theory of U-statistic, variability can be conveniently partitioned into two components reflecting variability due to each of the involved populations [46]. The components permit simple estimates leading to an accurate estimator of the overall variance [47, 48, 49].

Extensions of these approaches have also been used to derive variance for quantities of interest in cross-correlated multi-reader data in ROC analysis [6]. In such a setting, the overall variance estimator includes additional components that reflects between-reader variability. The estimator of between-reader variance is based on a small number of readers. However, the resulting estimator of the overall variance is adequate for typical multi-readers studies of diagnostic accuracy.

Despite good overall accuracy of variance estimators for cross-correlated data, they are typically biased (often upwardly). A number of successful efforts have been employed to reduce and sometimes even eliminate the bias of variance estimates for simple U-statistics [50, 51] and for statistics in multi-reader data [52].

4.2 PROPOSED APPROACH

As we discussed previously, the only standard approach for estimating parameters of the half-marginal models is based on Pseudo-Likelihood for linearized model, which suffers from it’s own criticisms and potentially sub-optimal performance for binary data in complex studies. We develop a semi-parametric estimation approach for parameters of the half-marginal model introduced earlier in Section 3.1 which involves a two-step model fitting process:

(1) A separate model is fit for each reader based on the corresponding subset of the cross-correlated dataset. We note that the binary observations are still correlated between the sub-datasets since all the readers evaluate the same set of subjects. But, within the reader-specific sub-dataset, the observations are independent in some simple models (e.g., model A, B and D) or have clustered correlation structure in other models (e.g., model C). Each of the reader-specific model is a probability model (e.g., GLM model with logit link fitted using GEE), with reliable and easy-to-obtain estimates.

(2) Another model is fit to the entire data as a whole with readers as fixed effect and accounting for correlation arising from observations from the same subject. This model is fit using GEE estimation approach which is known to produce consistent estimates of regression parameters and their estimated standard errors (when using robust estimator) even if the dependency is mis-specified, as long as the model for the marginal mean is correctly specified. The efficiency of the parameter estimates increases if the chosen working correlation structure is closer to the true correlation structure. Hence, instead of using the model-based estimates of SE, and given that we have a moderate number of subjects/clusters in our balanced datasets, we opt to using the robust estimates of standard error which are consistent regardless of the true form of the correlation structure.

Variability of the RG parameter estimate can then be constructed based on the variance estimates within readers (obtained from the individual reader-specific models), estimation of variability across the models (obtained based on estimates of the fixed-effects) and variance estimate of average coefficient conditional on readers (obtained from model fitted to overall data with readers as fixed-effects). Similar approach has been utilized previously in different applications [16, 22, 53] but no analogs currently exist for half-marginal models of cross-correlated binary data. We note that because of the relatively small number of readers in multi-reader studies, the proposed approach is not computationally intensive.

We now lay down formal derivation of the RG fixed-effect estimator and it's corresponding variance estimator. For these derivations, we assume a general notation i.e. n_c representing number of subjects and n_r representing number of readers.

4.2.1 Derivation of fixed-effect estimator

In order to estimate an RG fixed-effect parameter (e.g., $\tilde{\theta} = E\left[\tilde{\theta}_j\right]$) which in essence is a reader-averaged quantity, we can use individual reader-specific logistic models and estimate $\tilde{\theta}_j$ $(j = 1, .., n_r)$ for a sample of independent readers in a given study. For logistic models, $\tilde{\theta}_j$ is the logit of some probability $\tilde{p}_j$ and $\tilde{\theta}$ could represent e.g. the expectation of log odds of probability $\tilde{p}_j$ or the difference in the expectation of log odds of probabilities $\tilde{p}_{1j}$and $\tilde{p}_{2j}$ respectively for reader j.

Next, we can define the following fixed-effect estimator for generic $\tilde{\theta}$:

$$\bar{\hat{\theta}}_{\cdot} = \frac{\sum_{j=1}^{n_r} \hat{\theta}_j}{n_r}, \qquad (4.1)$$

which is a simple average of the estimates across individual readers. $\hat{\boldsymbol{\theta}}_j$'s are still dependent since the readers share overlapping set of subjects. However, this dependency does not affect this proposed estimator and is taken into account when constructing the variance estimator.

Note that in situations when we can assume independence among observations from the same reader (e.g., RG models A without covariate, B with subject level binary covariate and D with subject level continuous covariate), we can simply fit reader-specific logistic models using GEE technique with independence correlation structure (same as fitting a simple logistic model using maximum likelihood approach). The GEE estimators are known to be consistent and asymptotically normally distributed and hence the estimates resulting from these individual models are consistent. Averaging them also yields a consistent estimate.

However, when we need to assume dependence (e.g., RG model C with assessment level binary covariate) due to the structure of the data, we can again fit reader-specific logistic models using GEE approach but with a different correlation structure other than independence. Again since we have consistent estimates, when averaging across readers, we still get a consistent estimate of the reader-averaged effect of interest.

Finally, a positive aspect of this approach is that the estimates of the average effect that we obtain for models A, B and C are exactly similar to the empirical estimates.

4.2.2 Derivation of variance estimator

In this section we will exploit the simple structure of the proposed estimator of the fixed effect to develop the variance estimator accounting for both subject and reader-related variability sources.

Note, that the estimator will have the same form for all models considered. We now derive $V\left(\bar{\hat{\theta}}_{.}\right)$ i.e. $V\left(\frac{\sum_{j=1}^{n_r}\hat{\theta}_j}{n_r}\right)$:

By conditioning on reader-related random effects, including all possible interactions, for brevity all denoted as $(\boldsymbol{\beta})$, we can decompose the overall variance as follows:

$$V\left(\bar{\hat{\theta}}_{.}\right) = V\left[\mathrm{E}\left(\bar{\hat{\theta}}_{.}|\boldsymbol{\beta}\right)\right] + \mathrm{E}\left[V\left(\bar{\hat{\theta}}_{.}|\boldsymbol{\beta}\right)\right]$$

$V\left(\bar{\hat{\theta}}_{.}\right)$ = reader-related component + expected variance of reader-averaged index

Using the structure of the proposed estimator of the fixed-effect, we obtain

$$V\left(\bar{\bar{\theta}}_{.}\right) = V\left[\frac{1}{n_r}\sum_j E\left(\hat{\theta}_j|\beta_j\right)\right] + E\left[V\left(\frac{1}{n_r}\sum_j\left(\hat{\theta}_j|\beta_j\right)\right)\right] = (A) + (B) \qquad \mathbf{(\ 4.2\ )}$$

In equation (4.2) above, term (A) is the variance of the sum of independent terms (due to independence of the reader-effects and marginalizing over the subjects). In contrast, variance within term (B) is over $\hat{\theta}_j's$ which are still correlated due to sharing of same subjects. Thus, using the identical distribution of random effects, we further partition the variance as follows:

$$V\left(\bar{\bar{\theta}}_{.}\right) = \frac{1}{n_r^2} * n_r * V\left[E\left(\hat{\theta}_j|\beta_j\right)\right] + E\left[\frac{1}{n_r^2}\left\{\sum_j V\left(\hat{\theta}_j|\beta_j\right) + \sum_{j\neq k} 2 * Cov\left(\hat{\theta}_j,\hat{\theta}_k|\beta_j,\beta_k\right)\right\}\right] \qquad \mathbf{(\ 4.3\ )}$$

Finally, we can use the following plug-in estimators for equation (4.3):

$$\hat{V}\left(\bar{\bar{\theta}}_{.}\right)^{original} = \hat{V}\left(\bar{\bar{\theta}}_{.}\right) = \frac{1}{n_r} * \frac{1}{n_r-1}\sum_{j=1}^{n_r}\left(\hat{\theta}_j - \bar{\bar{\theta}}_{.}\right)^2 + \hat{V}\left(\bar{\bar{\theta}}_{.}|\boldsymbol{\beta}\right) = \frac{1}{n_r} * \boldsymbol{B1} + \boldsymbol{B2}\ , \qquad \mathbf{(\ 4.4\ )}$$

where $\boldsymbol{B1}$ represents the sample variance of reader-specific coefficient estimates which is a consistent estimator and $\boldsymbol{B2}$ represents the variance estimator of average regression coefficient which is also consistent since we obtain it from fixed-reader model fitted using GEE.

4.2.3 Derivation of bias-corrected variance estimator

$\hat{V}\left(\bar{\bar{\theta}}_{.}\right)^{original}$ is generally upward-biased i.e. it tends to overestimate variance for finite sample sizes, primarily due to the bias of the estimate of the first term. Repeating the formulation of this original estimator:

$$\hat{V}\left(\bar{\bar{\theta}}_{.}\right)^{original} = \hat{V}\left(\bar{\bar{\theta}}_{.}\right) = \frac{1}{n_r} * \frac{1}{n_r-1}\sum_{j=1}^{n_r}\left(\hat{\theta}_j - \bar{\bar{\theta}}_{.}\right)^2 + \hat{V}\left(\bar{\bar{\theta}}_{.}|\boldsymbol{\beta}\right) = \frac{1}{n_r} * \boldsymbol{B1} + \boldsymbol{B2}\ ,$$

the bias of the first term i.e. $\frac{1}{n_r} * \boldsymbol{B1}$ can be showed through the following steps noting that these derivations are for fixed-readers:

$$E\left[\frac{1}{n_r}*\frac{1}{n_r-1}\sum_{j=1}^{n_r}\left(\hat{\theta}_j-\bar{\hat{\theta}}_.\right)^2\right]=\frac{1}{n_r}*\frac{1}{n_r-1}E\left\{\sum_{j=1}^{n_r}\left(\hat{\theta}_j{}^2+\bar{\hat{\theta}}_.{}^2-2\hat{\theta}_j\bar{\hat{\theta}}_.\right)\right\}$$

$$=\frac{1}{n_r}*\frac{1}{n_r-1}*E\left\{\sum_{j=1}^{n_r}\hat{\theta}_j{}^2-\ n_r\bar{\hat{\theta}}_.{}^2\right\}$$

$$=\frac{1}{n_r}*\frac{1}{n_r-1}*\left[\sum_{j=1}^{n_r}\left\{V\left(\hat{\theta}_j\right)+E\left(\hat{\theta}_j\right)^2\right\}-n_r\left\{V\left(\bar{\hat{\theta}}_.\right)+E\left(\bar{\hat{\theta}}_.\right)^2\right\}\right]$$

$$=\frac{1}{n_r}*\frac{1}{n_r-1}*\left[\sum_{j=1}^{n_r}E\left(\hat{\theta}_j\right)^2-n_rE\left(\bar{\hat{\theta}}_.\right)^2\right]+\frac{1}{n_r}*\frac{1}{n_r-1}*\sum_{j=1}^{n_r}V\left(\hat{\theta}_j\right)-\frac{1}{n_r-1}$$

$$*V\left(\bar{\hat{\theta}}_.\right)$$

Rearranging terms, we have:

$$E\left[\frac{1}{n_r}*\frac{1}{n_r-1}\sum_{j=1}^{n_r}\left(\hat{\theta}_j-\bar{\hat{\theta}}_.\right)^2\right]=\frac{1}{n_r}*\frac{1}{n_r-1}\sum_{j=1}^{n_r}\left[E\left(\hat{\theta}_j\right)-E\left(\bar{\hat{\theta}}_.\right)\right]^2+\frac{1}{n_r-1}\sum_{j=1}^{n_r}\frac{V\left(\hat{\theta}_j\right)}{n_r}-\frac{1}{n_r-1}*V\left(\bar{\hat{\theta}}_.\right)$$

Thus, the bias of the first term $\left(\frac{1}{n_r}*B1\right)$ i.e. $\frac{1}{n_r-1}\sum_{j=1}^{n_r}\frac{V(\hat{\theta}_j)}{n_r}-\frac{1}{n_r-1}*V\left(\bar{\hat{\theta}}_.\right)$ depends on other variance terms of order $\frac{1}{n_r-1}$ as $n_r, n_c\to\infty$, which can be noticeable due to typically small number of readers in the multi-reader studies. We estimate this bias using the following plug-in estimator: $\frac{1}{n_r-1}\sum_j\frac{\hat{V}\left(\hat{\theta}_j|\beta_j\right)}{n_r}-\frac{1}{n_r-1}*\hat{V}\left(\bar{\hat{\theta}}_.|\boldsymbol{\beta}\right)$ based on combination of consistent estimates of variance.

The bias of the second term $(B2)$ is typically of order $\frac{1}{n_c}$, which is much less noticeable, due to relatively large number of subjects in multi-reader studies.

Thus, we propose the following bias-corrected (bc) variance estimator:

$$\hat{V}\left(\ \bar{\hat{\theta}}_.\right)^{bc}=\hat{V}\left(\ \bar{\hat{\theta}}_.\right)^{original}\ -\widehat{bias}$$

$$\hat{V}\left(\ \bar{\hat{\theta}}_.\right)^{bc}=\frac{1}{n_r}*\frac{1}{n_r-1}\sum_{j=1}^{n_r}\left(\hat{\theta}_j-\bar{\hat{\theta}}_.\right)^2+\hat{V}\left(\bar{\hat{\theta}}_.|\boldsymbol{\beta}\right)-\left[\frac{1}{n_r-1}\sum_j\frac{\hat{V}\left(\hat{\theta}_j|\beta_j\right)}{n_r}-\frac{1}{n_r-1}*\hat{V}\left(\bar{\hat{\theta}}_.|\boldsymbol{\beta}\right)\right]$$

$$\hat{V}\left(\bar{\bar{\theta}}_{.}\right)^{bc} = \frac{1}{n_r} * \frac{1}{n_r - 1} \sum_{j=1}^{n_r} \left(\hat{\theta}_j - \bar{\bar{\theta}}_{.}\right)^2 + \frac{n_r}{n_r - 1} * \hat{V}\left(\bar{\bar{\theta}}_{.} | \boldsymbol{\beta}\right) - \frac{1}{n_r - 1} \sum_{j=1}^{n_r} \frac{\hat{V}\left(\hat{\theta}_j | \beta_j\right)}{n_r}$$

$$\hat{V}\left(\bar{\bar{\theta}}_{.}\right)^{bc} = \frac{1}{n_r} * \boldsymbol{B1} + \frac{n_r}{n_r - 1} * \boldsymbol{B2} - \frac{1}{n_r - 1} * \boldsymbol{B3} = \boldsymbol{C1} + \boldsymbol{C2} - \boldsymbol{C3}$$

Usually, $\boldsymbol{C2}$ is greater than $\boldsymbol{C3}$ in absolute value such that the overall variance (over both readers and subjects) is greater than $\boldsymbol{C1}$ (sample variability of reader-specific summary index). When this is not true, which sometimes happens in practice, the overall variance becomes less than $\boldsymbol{C1}$ (even attains a negative value sometimes). In this case, we may use $\boldsymbol{C1}$ i.e. $\left(\frac{1}{n_r} * \frac{1}{n_r-1} \sum_{j=1}^{n_r} \left(\hat{\theta}_j - \bar{\bar{\theta}}_{.}\right)^2\right)$ as our overall random-reader variance estimate (as suggested by the original DBM approach [7] or [13]). This constraint on the variance corrects instances of invalid estimates of specific variance components, thereby helping with attaining coverage closer to the nominal value.

Overall, both the original and bias-corrected variance estimators for the RG fixed-effect estimator have a simple structure which is linear combination of different variance terms that are simple to compute.

4.2.4 Asymptotic properties of fixed-effect estimator

Next, we would like to demonstrate that our fixed-effect estimator satisfies the much desirable MSE consistency property which implies consistency (convergence in probability). For this purpose, we show 1) it is asymptotically unbiased and 2) it's variance approaches "0" as sample size increases.

First, the fixed-effects estimator can be shown to be asymptotically unbiased with respect to $\tilde{\theta}$ since a) each $\widehat{p_J} \xrightarrow{p} p_j$ by Law of Large Numbers and by continuity theorem (CMT), $\widehat{\theta_J} \xrightarrow{p} \theta_j$ as subject sample size (denoted as n_c here) increases, b) convergence in probability implies convergence in distribution i.e. $\widehat{\theta_J} \xrightarrow{d} \theta_j$, hence $\lim_{n_c\to\infty} E\left(\widehat{\theta_J}\right) = E\left(\theta_j\right)$, c) for a random sample of readers $E\left(\bar{\hat{\theta}}_{\cdot}\right) = E\left(\theta_j|\boldsymbol{\alpha}\right)$, and $\lim_{n_c\to\infty} \mathrm{E}\left(\bar{\hat{\theta}}_{\cdot}\right) = \lim_{n_c\to\infty} \mathrm{E}\left(\frac{\sum_{j=1}^{n_r}\hat{\theta}_j}{n_r}\right) = E\left(\theta_j\right) = \tilde{\theta}$. Hence, $\bar{\hat{\theta}}_{\cdot}$ is an asymptotically unbiased estimator for $\tilde{\theta}$ as subject and reader sample sizes increase.

Second, by examining the variance of the fixed-effect estimator i.e.:

$$V\left(\bar{\hat{\theta}}_{\cdot}\right) = \frac{1}{n_r^2} V\left[\sum_j E\left(\hat{\theta}_j|\beta_j\right)\right] + \mathrm{E}\left[V\left(\bar{\hat{\theta}}_{\cdot}|\boldsymbol{\beta}\right)\right] = (A) + (B),$$

we note that term (A) is the variance of the average of independent conditional expectations which diminishes with increasing number of readers, and (B) is an expectation of the fixed reader variance of the average fixed effect which converges to 0 with increasing number of subjects. Thus, the Mean Square Error (MSE) of the proposed fixed effect estimator diminishes with increasing number of subjects and readers, ensuring the consistency of the proposed estimator.

It is also worth noting that, due to the balanced structure of the fully-crossed data, the proposed estimator (simple average of the estimates from the reader specific models) is equivalent to the GEE estimator with the independent working correlation matrix. Then, due to the consistency and asymptotic normality of the GEE estimates for clustered data regardless of the correctness of working correlation matrix, the same properties can be claimed for proposed estimator conditional on the readers' effects.

4.2.5 Algorithm for computing fixed-effect estimate and its variance

As mentioned earlier, the first step of the proposed approach involves fitting GLM models with logit link for individual readers $j = 1, \dots, n_r$ (all other notations remain the same as defined in previous chapters). For each modeling scenario, we lay down the model specification and the different estimators that are used to compute estimates of fixed effect and different components of the fixed-effect variance:

Table 29: Computing estimates from logistic models

Modeling Scenario	**Fitted Reader-Specific logistic model using GEE**	**Fixed-Effect Estimator**	**Component B1**	**Component B3**
Model A	$logit(p_j) = \mu_j$	$\bar{\hat{\mu}}_{.} = \frac{\sum_{j=1}^{n_r} \hat{\mu}_j}{n_r}$	$\frac{1}{n_r - 1} \sum_{j=1}^{n_r} \left(\hat{\mu}_j - \bar{\hat{\mu}}_{.}\right)^2$	$\frac{1}{n_r} \sum_{j=1}^{n_r} \hat{V}\left(\hat{\mu}_j \mid \beta_j\right)$
Model B	$logit(p_{jD}) = \mu_j + \eta_j * D$	$\bar{\hat{\eta}}_{.} = \frac{\sum_{j=1}^{n_r} \hat{\eta}_j}{n_r}$	$\frac{1}{n_r - 1} \sum_{j=1}^{n_r} \left(\hat{\eta}_j - \bar{\hat{\eta}}_{.}\right)^2$	$\frac{1}{n_r} \sum_{j=1}^{n_r} \hat{V}\left(\hat{\eta}_j \mid \beta_j\right)$
Model C	$logit(p_{jM}) = \mu_j + \delta_j * M$ with $Corr(Y_{jiM}, Y_{jiM'}) = \rho$ for given single subject i since each subject was examined under two modalities	$\bar{\hat{\delta}}_{.} = \frac{\sum_{j=1}^{n_r} \hat{\delta}_j}{n_r}$	$\frac{1}{n_r - 1} \sum_{j=1}^{n_r} \left(\hat{\delta}_j - \bar{\hat{\delta}}_{.}\right)^2$	$\frac{1}{n_r} \sum_{j=1}^{n_r} \hat{V}\left(\hat{\delta}_j \mid \beta_j\right)$
Model D	$logit(p_{jX}) = \mu_j + \tau_j * X$	$\bar{\hat{\tau}}_{.} = \frac{\sum_{j=1}^{n_r} \hat{\tau}_j}{n_r}$	$\frac{1}{n_r - 1} \sum_{j=1}^{n_r} \left(\hat{\tau}_j - \bar{\hat{\tau}}_{.}\right)^2$	$\frac{1}{n_r} \sum_{j=1}^{n_r} \hat{V}\left(\hat{\tau}_j \mid \beta_j\right)$
B1: Sample variance of reader-specific coefficient estimates **B3:** Average of variance estimators for reader-specific coefficient estimates Models A, B and D assume independence working correlation structure while model C assumes compound symmetry working correlation structure				

The second step involves fitting logistic models using GEE mechanism with readers considered as fixed effect and necessary interaction terms. Working correlation structure is set to compound symmetry. These models are used to obtain the variance estimate of the average coefficient estimate conditioning on readers ($\boldsymbol{B2}$) which can be easily obtained using appropriate ESTIMATE/LSMEANS/LSMESTIMATE statements in PROC GLIMMIX/GENMOD procedures in SAS. The robust standard error estimates will be consistent due to the virtue of using GEE.

Table 30 provides a summary of the fitted GEE models for each modeling scenario:

Table 30: Models fitted using GEE mechanism

Modeling Scenario	Fitted Logistic Model to overall data using GEE	Component B2
Model A	$\boldsymbol{logit(p_j) = \mu + \ \omega * Reader_j}$ $Corr(Y_{ij}, Y_{ij'}) = \rho$	$\hat{V}(\bar{\hat{\mu}}_{.}\lvert\boldsymbol{\beta})$
Model B	$\boldsymbol{logit(p_{jD}) = \mu + \eta * D + \ \omega * Reader_j + \kappa * (D * Reader_j)}$ $Corr(Y_{ij}, Y_{ij'}) = \rho$ remains similar for diseased and non-diseased subjects	$\hat{V}(\bar{\hat{\eta}}_{.}\lvert\boldsymbol{\beta})$
Model C	$\boldsymbol{logit(p_{jM}) = \mu + \delta * M + \ \omega * Reader_j + \upsilon * (M * Reader_j)}$ $\boldsymbol{Corr(Y_{ijM}, Y_{ij'M'}) = Corr(Y_{ijM}, Y_{ijM'}) = Corr(Y_{ijM}, Y_{ij'M}) = \rho}$ for a single subject (correlations can be allowed to differ across both modalities by using "group=modality" statement in PROC GLMMIX, SAS)	$\hat{V}\left(\bar{\hat{\delta}}_{.}\lvert\boldsymbol{\beta}\right)$
Model D	$\boldsymbol{logit(p_{jX}) = \mu + \tau * X + \ \omega * Reader_j + \zeta * (X * Reader_j)}$ $\boldsymbol{Corr(Y_{ij}, Y_{ij'}) = \rho}$	$\hat{V}(\bar{\hat{\tau}}_{.}\lvert\boldsymbol{\beta})$

1. **B2**: Variance estimator of average regression coefficient
2. Working correlations structure is compound symmetry
3. ω is the fixed reader effect; η is the fixed effect for true disease status; κ is the fixed reader*truth interaction effect; δ is the fixed modality effect; υ is the fixed reader*modality interaction effect; τ is the fixed-slope; ζ is the fixed reader*X interaction effect

4.3 SIMULATION STUDY

4.3.1 Simulation Study Details

In the simulation study we considered the same data as was used for evaluating built in GG and RG models (Chapters 2 and 3). The true values for models' parameters were the same as for investigating the built-in RG model (as described in 3.5.2). Here we focus only medium-sized variance components for simplicity and easy comparison with the PL-RG approach.

For evaluating performance of the proposed approach we used the same summary indices as for evaluations in previous chapters. For construction of confidence interval we considered standard normal and t-distribution with containment degrees of freedom. T-distribution is a default for built-in RG approach and can be useful for the procedure based on the bias-reduced variance estimator.

The following notations are used for the tables presented in the "Simulation Study Results" section:

1. GG="Subject-Specific" parameter; RG="Half Marginal" parameter
2. Original=Original SE Estimator; Bias-Corrected=Bias-Corrected SE Estimator
3. Coverage (%)= Coverage of the 95% CI
4. Est=Average of fixed effect parameter estimate
5. MC SD=Monte Carlo Standard Deviation
6. SB=Standardized bias of fixed effect parameter estimate (using MC Standard Deviation)
7. SE=Average Estimated Standard Error for fixed effect parameter estimate
8. RBS (%)=Relative bias of estimated standard error for fixed-effect parameter estimate (using MC Standard Deviation)

9. Sim Used=Used simulations: (1) Simulations resulting in non-zero standard error estimates of fixed-effect (2) PROC LOGISITIC/GLIMMIX/GENMOD convergence criteria satisfied
10. If one or more readers satisfied conditions (1) and/or (2) entire simulation was discarded

4.3.2 Simulation Study Results

For model A with covariate-free setting [Table 31], we observed the following:

i. Standardized bias of the fixed effect estimate was well within 1 Monte Carlo SD. The fixed-effect estimates displayed asymptotically unbiased property, mostly with increasing number of subjects.

ii. In all scenarios, the relative bias of the SE based on the original SE estimator was positive which did not decrease with increasing readers and was unlike what was expected. The bias-corrected SE estimator was substantially less biased and decreased with increasing number of readers.

iii. Z-approximation based on original SE estimator led to coverage which was close to nominal. But, when used along with the bias corrected SE, led to slightly elevated type I error rate.

iv. T-based reference distribution with containment degrees of freedom along with original SE estimate led to over conservative coverage while the bias-corrected estimate led to some improvement.

v. We also noted the decreased number of simulations that converged especially when true GG probabilities were extreme (e.g., 0.1) especially with large number of readers. This seem to have resulted from data separation issues for specific readers which led to discarding the entire simulation that the reader/s were part of. It seems tempting to use information from the remaining readers but the proposed approach requires fitting an additional overall GEE model which from our experience does not converge in such cases. Hence, it is not a practical option. The reason for this is the need to combine the information from both models and use simulations only where we have information from both the models.

When comparing the results of this approach vs the PL-RG approach, we observed that we obtained appropriate coverage rates using both the approaches. For this simple model, the standardized bias of the fixed effect estimator was more or less similar. At the same time, the relative bias of bias-corrected SE estimator from the proposed approach was similar in magnitude and direction as compared to the PL-RG SE estimator. There were slightly more convergence issues with the proposed approach for smaller probability values whereas almost all simulations converged for the PL-RG approach.

Table 31 Simulation Results for Model A fitted used Proposed Method

		Proposed Approach					
		p^{GG}=0.1 μ^{GG}=-2.20 μ^{RG}=-1.66		p^{GG}=0.5 μ^{GG}=0 μ^{RG}=0		p^{GG}=0.7 μ^{GG}=0.85 μ^{RG}=0.63	
		n_1		n_1		n_1	
	n_r	55	100	55	100	55	100
t-Coverage containment df (%) (Original/BC)	**5**	99/98	98/97	98/97	98/97	98/97	98/97
	10	98/97	98/97	98/97	97/97	98/97	98/97
Z-Coverage (%) (Original/BC)	**5**	94/90	94/91	93/91	93/92	93/91	92/91
	10	96/94	96/94	95/95	95/94	95/94	95/94
Est (MC SD)	**5**	-1.69 (0.382)	-1.69 (0.348)	0.01 (0.326)	0.01 (0.303)	0.65 (0.341)	0.64 (0.305)
	10	-1.7 (0.296)	-1.68 (0.256)	0 (0.258)	0 (0.223)	0.65 (0.264)	0.64 (0.228)
SB	**5**	-0.08	-0.1	0.02	0.03	0.06	0.04
	10	-0.12	-0.09	0	0.01	0.09	0.07
Original SE (RBS (%)) Bias-Corrected SE (RBS (%))	**5**	0.405 (6) 0.366 (-4)	0.36 (4) 0.336 (-3)	0.338 (4) 0.318 (-2)	0.306 (1) 0.294 (-3)	0.354 (4) 0.332 (-3)	0.315 (3) 0.302 (-1)
	10	0.321 (8) 0.296 (0)	0.275 (7) 0.26 (2)	0.268 (4) 0.255 (-1)	0.235 (5) 0.228 (2)	0.274 (4) 0.26 (-1)	0.242 (6) 0.233 (2)
Sim Used	**5**	966	999	1000	1000	998	1000
	10	922	990	1000	1000	998	999

Simulation parameters: Subject Variance=2, Reader Variance=0.7

We observed the following when applying the proposed approach to model B [Table 32]:

i. Similar to model A, we saw small standardized bias of the fixed effect.

ii. The estimates using original SE estimator where heavily upward-biased which got worse when the true GG probabilities were close to 0 or 1. The bias-corrected SE estimator was substantially less biased.

iii. Using the z-reference distribution with the original SE estimator (although quite-upward biased) led to almost optimal coverage of the 95% CI. However, using it in conjunction with the bias corrected SE estimator led to coverage on the lower end but yet above 90%.

iv. The t-based approximation led to extremely over conservative coverage with the original SE estimator and still conservative coverage with it's biased corrected version.

v. Once again since we have a binary covariate, data from several readers suffered from data-separation issues which led to quit a bit of loss in the number of converged simulations used for computation.

On comparison with PL-RG model results, we observed good coverage rates using both techniques. The standardized bias had no clear trend and was similar for both approaches in some scenarios (TPF=0.6, FPF=0.6) while completely different in magnitude and direction for others (e.g., TPF=0.3, FPF=0.9). Finally, the relative bias of the bias-corrected SE estimator was more closer to the PL-RG SE estimator as compared to the original SE estimator.

Table 32 Simulation Results for Model B fitted used Proposed Method

			TPF^{GG}=0.5, FPF^{GG}=0.03 **η^{GG}=3.48** **η^{RG}=2.75**		**TPF^{GG}=0.3, FPF^{GG}=0.9** **η^{GG}=-3.04** **η^{RG}=-2.31**		**TPF^{GG}=0.6, FPF^{GG}=0.6** **η^{GG}=0** **η^{RG}=0**	
			n_0		**n_0**		**n_0**	
	n_r	**n_1**	**100**	**175**	**100**	**175**	**100**	**175**
t-Coverage containment df (%): (Original/BC)	**5**	**55**	100/98	100/98	100/99	100/99	100/99	100/98
	5	**100**	99/97	99/98	100/98	99/98	99/98	99/98
	10	**55**	99/97	99/97	99/97	99/97	98/97	98/97
	10	**100**	99/97	99/97	98/97	98/97	99/98	99/98
Z-Coverage (%):(Original/BC)	**5**	**55**	97/93	97/92	97/94	96/93	96/93	96/92
	5	**100**	97/92	96/92	96/92	95/92	96/93	94/91
	10	**55**	98/94	97/94	96/95	97/95	96/94	96/94
	10	**100**	97/94	97/94	96/94	96/95	97/95	96/94
Est (MC SD)	**5**	**55**	2.8 (0.351)	2.79 (0.333)	-2.35 (0.317)	-2.34 (0.298)	0 (0.288)	0.01 (0.274)
	5	**100**	2.8 (0.333)	2.79 (0.314)	-2.35 (0.294)	-2.34 (0.271)	0 (0.259)	0 (0.246)
	10	**55**	2.79 (0.284)	2.8 (0.265)	-2.35 (0.263)	-2.35 (0.241)	0 (0.24)	0.01 (0.228)
	10	**100**	2.79 (0.261)	2.8 (0.243)	-2.35 (0.235)	-2.34 (0.213)	0 (0.206)	0 (0.193)
SB	**5**	**55**	0.15	0.13	-0.13	-0.1	0	0.03
	5	**100**	0.17	0.14	-0.13	-0.1	0	0.01
	10	**55**	0.15	0.21	-0.17	-0.16	0	0.03
	10	**100**	0.17	0.21	-0.15	-0.15	0	-0.02
Original SE (RBS (%))	**5**	**55**	0.426 (21)	0.376 (13)	0.361 (14)	0.329 (11)	0.316 (10)	0.298 (9)
Bias-Corrected SE (RBS (%))			0.349 (-1)	0.32 (-4)	0.314 (-1)	0.291 (-2)	0.282 (-2)	0.267 (-3)
	5	**100**	0.403 (21)	0.346 (10)	0.327 (11)	0.29 (7)	0.28 (8)	0.26 (6)
			0.33 (-1)	0.294 (-6)	0.287 (-3)	0.258 (-5)	0.253 (-2)	0.237 (-4)
	10	**55**	0.338 (19)	0.302 (14)	0.29 (10)	0.266 (10)	0.258 (7)	0.243 (7)
			0.292 (3)	0.267 (1)	0.262 (0)	0.243 (1)	0.238 (-1)	0.225 (-1)
	10	**100**	0.315 (21)	0.277 (14)	0.262 (11)	0.233 (9)	0.226 (9)	0.209 (9)
			0.271 (4)	0.244 (1)	0.238 (1)	0.215 (1)	0.21 (2)	0.196 (1)
Sim Used	**5**	**55**	887	966	992	998	1000	1000
	5	**100**	887	966	994	998	1000	1000
	10	**55**	784	951	992	998	999	990
	10	**100**	784	951	993	995	1000	994

Simulation parameters: Subject Variance=1.9, Reader Variance=0.5, Reader*Truth Variance=0.14

In case of model C which uses an assessment-level covariate [Table 33], the following was observed:

i. Small standardized bias (<<50%) indicating that this bias has a minimal affect on the coverage.

ii. The original and the bias-corrected version of the SE estimator showed similar behavior as in models A and B. The original SE estimator was highly upward biased and this bias did not always decrease with increasing number of readers. On the other hand, the bias corrected one was less biased, mostly under biased and this bias usually decreased when number of readers went up from 5 to 10.

iii. The z-based CI using the original SE estimator led to more optimal coverage while the bias-corrected SE estimator led to poor coverage (e.g. <90%) when number of readers was small.

iv. The t-based CI with the original SE estimator led to over conservative coverage, while it's bias corrected counterpart led to coverage closer to nominal.

v. The simulation models also suffered from data separation issues resulting in loss of simulations. The loss was however less significant as compared to model B.

When comparing the results from this approach vs. the PL-RG approach, the coverage rates when using based on the proposed approach when optimal were very similar to the PL-RG approach. The standardized bias of the fixed-effect estimator was more or less close using both approaches except that it was slightly higher in magnitude (but under-biased) for the proposed approach when the GG true positive fraction rates was close to 0 or 1. When comparing the relative bias of the SE estimator, the PL-RG approach led to under-biased and bigger SE estimates in magnitude. There were slightly more convergence issues with the proposed method due to data separation issues and the usage of two different models simultaneously.

Table 33 Simulation Results for Model C fitted used Proposed Method

		Proposed Method					
		$p_1^{GG}=0.5$, $p_2^{GG}=0.6$ **$\delta^{GG}=-0.41$** **$\delta^{RG}=-0.26$**		**$p_1^{GG}=0.1$, $p_2^{GG}=0.7$** **$\delta^{GG}=-3.04$** **$\delta^{RG}=-1.96$**		**$p_1^{GG}=0.8$, $p_2^{GG}=0.8$** **$\delta^{GG}=0$** **$\delta^{RG}=0$**	
		n_1		**n_1**		**n_1**	
	n_r	**55**	**100**	**55**	**100**	**55**	**100**
t-Coverage containment df (%) (Original/BC)	5	99/96	98/96	99/97	98/96	99/96	98/96
	10	99/96	98/96	98/96	98/97	98/96	98/96
Z-Coverage (%) (Original/BC)	5	95/90	93/89	95/92	94/89	94/88	93/88
	10	96/93	96/94	96/93	95/93	96/92	96/94
Est (MC SD)	5	-0.27(0.291)	-0.27(0.277)	-2.02(0.351)	-2(0.323)	0(0.322)	-0.01(0.297)
	10	-0.26(0.208)	-0.26(0.193)	-2.01(0.266)	-1.99(0.24)	0(0.229)	0(0.205)
SB	5	-0.06	-0.04	-0.17	-0.13	-0.01	-0.02
	10	-0.04	-0.03	-0.19	-0.16	-0.01	0
Original SE (RBS (%))	5	0.328 (13)	0.29 (4)	0.395 (13)	0.332 (3)	0.349 (8)	0.308 (4)
Bias-Corrected SE (RBS (%))		0.292 (0)	0.266 (-4)	0.344 (-2)	0.299 (-8)	0.303 (-6)	0.279 (-6)
	10	0.233 (12)	0.21 (9)	0.294 (11)	0.253 (6)	0.254 (11)	0.222 (8)
		0.208 (0)	0.195 (1)	0.261 (-2)	0.232 (-3)	0.224 (-2)	0.204 (-1)
Sim Used	5	999	1000	968	993	986	997
	10	1000	1000	939	987	979	996

Simulation parameters: Subject Variance=3.71, Reader Variance=0.8672, Reader*Modality Variance=0.4

Finally, when dealing with a continuous covariate, as in model D [Table 34], we observed the following:

i. The standardized bias of the fixed-effect was within 1 MC SD but was a little higher >50% for bigger GG slope values e.g., 0.1 which did not affect the coverage much except that the z-based coverage using bias-corrected variance estimator dropped below 90%.

ii. The relative bias of the standard error estimates using original SE estimator was sometimes slightly upward biased and sometimes slightly downward biased. This bias did not always decrease with increasing number of readers. In contrast, the bias resulting from using the bias corrected alterative always led to substantial decrease in bias and always decreased as number of readers increased.

iii. The z-based coverage based on the original variance estimator led to more or less close to nominal coverage. However, with the bias corrected estimator we saw a further decrease in the coverage which was unacceptable especially as the slope got bigger.

iv. The t-based coverage using original variance estimator led to extremely over-conservative coverage. However, when this reference distribution was used along with the bias-corrected variance estimator, there was a slight improvement in the coverage which was still conservative.

v. Very minimal convergence issues were noted for this model.

Comparing this approach with the PL-RG approach, both approaches gave acceptable coverage rates. However, the standardized bias of the fixed-effect estimator was greater for the proposed approach especially for large slope values whereas the relative bias of the PL-RG SE estimator was mostly somewhere in between the original and biased corrected version of the SE estimator for the proposed approach. Overall there were far less convergence issues for this model using the new approach vs. the PL-RG estimation approach.

To assess the potential advantages for improving the fixed effect estimate for the proposed approach we tried to use the estimate from the fixed-reader GEE model while using the proposed variance estimator. Simulation results [Table 34, Table 35] comparing "Estimator 1" (based on using simple average of the reader-specific estimates which also happens to be similar to the estimate obtained from a GEE model with fixed-reader effect and independence correlation structure) and "Estimator 2" (using a GEE model with fixed-reader effect with compound symmetry correlation structure) showed that results based on using "Estimator 2" were less biased and slightly improved the CI coverage.

Table 34 Simulation Results for Model D fitted used Proposed Method (Estimator 1)

		Proposed Method							
		$\mu^{GG}=0$ $\tau^{GG}=0$ $\tau^{RG}=0$		**$\mu^{GG}=0$ $\tau^{GG}=.03$ $\tau^{RG}=0.022$**		**$\mu^{GG}=0$ $\tau^{GG}=.06$ $\tau^{RG}=0.044$**		**$\mu^{GG}=0$ $\tau^{GG}=0.1$ $\tau^{RG}=0.073$**	
		n_1		**n_1**		**n_1**		**n_1**	
	n_r	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**
t-Coverage containment df (%) (Original/BC)	**5**	99/98	98/97	99/98	98/97	99/98	99/97	100/97	99/97
	10	98/97	98/97	98/97	97/96	98/97	97/96	97/93	97/94
Z-Coverage (%) (Original/BC)	**5**	95/92	92/90	96/93	92/91	95/92	93/89	96/88	94/89
	10	95/94	95/94	95/94	94/93	96/94	95/94	94/88	94/89
Est (MC SD)	**5**	0(0.013)	0(0.012)	0.024(0.014)	0.023(0.013)	0.049(0.016)	0.047(0.014)	0.087(0.023)	0.082(0.016)
	10	0(0.01)	0(0.009)	0.024(0.011)	0.023(0.009)	0.05(0.012)	0.047(0.01)	0.087(0.018)	0.082(0.012)
SB	**5**	0	-0.01	0.17	0.11	0.36	0.27	0.64	0.53
	10	0.02	0.01	0.24	0.16	0.48	0.37	0.79	0.71
Original SE (RBS (%)) Bias-Corrected SE (RBS (%))	**5**	0.013 (3) 0.013 (-3)	0.012 (-2) 0.011 (-6)	0.014 (5) 0.013 (-3)	0.012 (-2) 0.012 (-7)	0.016 (4) 0.015 (-6)	0.013 (-1) 0.013 (-8)	0.022 (-3) 0.019 (-17)	0.017 (2) 0.015 (-9)
	10	0.01 (4) 0.01 (-1)	0.009 (2) 0.009 (-1)	0.011 (3) 0.01 (-3)	0.009 (1) 0.009 (-2)	0.013 (6) 0.012 (-3)	0.01 (4) 0.01 (-2)	0.017 (-4) 0.015 (-16)	0.013 (4) 0.012 (-6)
Sim Used	**5**	1000	1000	1000	1000	1000	1000	996	1000
	10	1000	1000	1000	1000	1000	1000	991	1000

1. Simulation parameters: Subject Variance=2, Reader Variance=0.2 Reader*LesionSize Variance=0.001
2. Continuous variable (LesionSize)~Unif(1,100) and has been centered at 50.5 mm during simulation and fitting process

Table 35 Simulation Results for Model D fitted used Proposed Method (Estimator 2)

		Proposed Method							
		μ^{GG}=0 **τ^{GG}=0** **τ^{RG}=0**		**μ^{GG}=0** **τ^{GG}=.03** **τ^{RG}=0.022**		**μ^{GG}=0** **τ^{GG}=.06** **τ^{RG}=0.044**		**μ^{GG}=0** **τ^{GG}=0.1** **τ^{RG}=0.073**	
		n_1		**n_1**		**n_1**		**n_1**	
	n_r	**55**	**100**	**55**	**100**	**55**	**100**	**55**	**100**
t-Coverage containment df (%) (Original/BC)	**5**	99/98	98/97	100/99	98/97	99/98	98/97	100/97	99/97
	10	98/97	98/97	98/97	97/96	98/97	97/96	98/95	98/95
Z-Coverage (%) (Original/BC)	**5**	95/92	92/90	96/93	92/91	95/92	93/89	96/88	94/89
	10	96/94	95/94	96/95	94/93	96/94	96/94	95/90	94/91
Est (MC SD)	**5**	0(0.013)	0(0.012)	0.024(0.013)	0.023(0.013)	0.049(0.015)	0.047(0.013)	0.087(0.022)	0.081(0.016)
	10	0(0.01)	0(0.009)	0.024(0.01)	0.023(0.009)	0.049(0.012)	0.047(0.01)	0.086(0.017)	0.081(0.012)
SB	**5**	0	-0.01	0.16	0.1	0.34	0.25	0.63	0.51
	10	0.02	0.01	0.22	0.14	0.44	0.32	0.76	0.66
Original SE (RBS (%)) Bias-Corrected SE (RBS (%))	**5**	0.013 (4) 0.013 (-2)	0.012 (-2) 0.011 (-5)	0.014 (6) 0.013 (-2)	0.012 (-2) 0.012 (-6)	0.016 (5) 0.015 (-5)	0.013 (-1) 0.013 (-7)	0.022 (1) 0.019 (-14)	0.017 (3) 0.015 (-8)
	10	0.01 (5) 0.01 (0)	0.009 (3) 0.009 (-1)	0.011 (5) 0.01 (-1)	0.009 (2) 0.009 (-2)	0.013 (9) 0.012 (0)	0.01 (5) 0.01 (-1)	0.017 (0) 0.015 (-13)	0.013 (6) 0.012 (-4)
Sim Used	**5**	1000	1000	1000	1000	1000	1000	996	1000
	10	1000	1000	1000	1000	1000	1000	991	1000

1. Simulation parameters: Subject Variance=2, Reader Variance=0.2 Reader*LesionSize Variance=0.001
2. Continuous variable (LesionSize)~Unif(1,100) and has been centered at 50.5 mm during simulation and fitting process

Table 36 Comparison of PL-RG Model estimates with Proposed Method estimates based on real dataset

Model	Sample Size	Parameter	PL-RG model on logit scale (Estimate ± SE)	Proposed Method on logit scale* (Estimate ± SE)
A (modality=1)	$n_r = 7$ $n_1 = 55$	$logit(TPF) = \tilde{\mu}$	0.0705 ± 0.3044	0.0763 ± 0.3191^{Orig} 0.0763 ± 0.3072^{BC}
A (modality=2)	$n_r = 7$ $n_1 = 55$	$logit(TPF) = \tilde{\mu}$	0.3187 ± 0.2826	0.3452 ± 0.2969^{Orig} 0.3452 ± 0.2861^{BC}
B (modality=1)	$n_r = 7$ $n_0 = 175$ $n_1 = 55$	$\ln(OR) = \tilde{\eta}$	2.8448 ± 0.3086	2.9759 ± 0.3707^{Orig} 2.9759 ± 0.3369^{BC}
B (modality=2)	$n_r = 7$ $n_0 = 175$ $n_1 = 55$	$\ln(OR) = \tilde{\eta}$	2.8170 ± 0.2595	2.9760 ± 0.3388^{Orig} 2.9760 ± 0.3072^{BC}
C	$n_r = 7$ $n_1 = 55$	$\ln(OR) = \tilde{\delta}$	-0.2687 ± 0.1159	-0.2689 ± 0.1523^{Orig} -0.2689 ± 0.1083^{BC}

Orig: Original Standard Error Estimate

BC: Bias corrected Standard Error Estimate

*The fixed effect estimates for models A-C are exactly the same as empirical estimates

We compared the PL-RG approach with the proposed approach based on the real-life data [43] as illustrated in Table 36. We observed that the fixed-effect estimates from the proposed approach closely agreed with those from the PL-RG approach, although a bit larger. For models A, B and C, they were exactly similar to the empirical RG estimates. The original standard errors were slightly greater than the ones from PL-RG approach. Overall, the two approaches led to much similar results.

4.4 SUMMARY AND DISCUSSION

The proposed approach offers an acceptable alternative to the PL-RG approach for estimating half-marginal models in multi-reader studies. The primary advantage of this approach is the probabilistic nature and explicit estimation based on non-iterative combination of results from robust GEE approaches. The fixed-effect estimator is also consistent and produces the exact same estimate as the empirical estimates when using simple models with binary covariates or no covariates. This is very much unlike the PL-RG approach which is often criticized for it's non-probabilistic nature and possible consistency problem of its estimates. On the other hand, both the proposed approach and built in PL-RG technique allow us to make statistical inferences with confidence, and the fixed-effect estimator from the PL-RG approach often leads to more precise estimates in the considered scenarios.

The simple and well-grounded formulation of proposed estimators of fixed effects and their variance creates a solid foundation for developing improved approach for estimating useful half-marginal models in the future. However, the currently proposed approach has limitations. First of all, it requires fitting and implementing two separate types of models unlike fitting a single model for the PL-RG approach which is more convenient. In many cases that leads to lower implementation than for the built-in PL-RG approach (PROC GLIMMIX, SAS). Furthermore, there is a greater chance of data separation issues (more so due to the presence of categorical covariates) resulting from fitting reader-specific models and GEE models with readers as fixed-effects creating convergence issues not typically seen with PL-RG models.

In the future developments, I plan to address these issues by using more efficient and robust techniques for estimating fixed effects. It would also be of interest to extend the proposed technique for estimating variance components, which are important for understanding the sources of variability in the considered multi-reader studies. These components are generally useful in designing future studies and also may be useful in improved estimation methods. For the biased-reduced variance estimator, it might be worthwhile to investigate the distribution of the corresponding test-statistic and possibly derive the Satterthwaite-like degrees of freedom. This could possibly lead to further improvement in coverage especially when fitting more complex models.

APPENDIX A: IMPLEMENTATION OF THE SUBJECT-SPECIFIC MODELS IN SAS

Model A

```
proc glimmix data=nodule;
       class subject reader;
       model y(event="1")=/dist=binary;
       random reader subject;
run;
```

Model B

```
proc glimmix data=nodule;
        class truth subject reader;
        model y(event="1")=truth/dist=binary;
        random subject reader reader*truth;
run;
```

Different variability for “diseased” and “non-diseased” subjects can be modeled using the option “/group=truth” in the “random” statement.

Model C

```
proc glimmix data=nodule
        class modality subject reader;
        model y(event="1")=modality/dist=binary;
        random subject reader reader*modality;
run;
```

Model D

```
proc glimmix data=nodule method=RSPL ;
        class subject reader ;
        model y(event="1")= X/dist=binary;
        random subject reader reader*X;
run;
```

Estimation of these models using Laplace approximation is achieved by using option METHOD=LAPLACE (METHOD=RSPL is the default option engaging the PL estimation).

APPENDIX B: IMPLEMENTATION OF THE HALF MARGINAL MODELS IN SAS

Model A

```
proc glimmix data=nodule method=RSPL;
        class subject reader;
        model y(event="1")=/dist=binary;
        random reader;
        random intercept/subject=subject type=cs residual;
run;
```

Model B

```
proc glimmix data=nodule order=data method=RSPL;
        class truth subject reader;
        model y(event="1")=truth/dist=binary;
        random reader reader*truth;
        random intercept/subject=subject type=cs residual;
run;
```

Model C

```
proc glimmix data=nodule method=RSPL;
        class modality subject reader;
        model y(event="1")=modality/dist=binary;
        random reader reader*modality;
        random intercept/subject=subject type=cs residual;
run;
```

Model D

```
proc glimmix data=nodule method=RSPL ;
        class subject reader ;
        model y(event="1")= X/dist=binary;
        random reader reader*X;
        random intercept/subject=subject type=cs residual;
run;
```

APPENDIX C: ACCURACY OF ESTIMATED VARIANCE COMPONENTS FOR MODEL A

We tested the accuracy of each variance component (reader and subject variability) of model A for different variance component configuration (small, medium and large in magnitude). As observed in tables below, PL estimated technique resulted in substantial bias in each variance component as compared to the LA estimated technique. To compute the relative bias, we used the true variance parameters values as set in simulations.

Table 37 Model A: Variance Estimation (Small variance components)

		PL Estimation*						LAPLACE Estimation*					
		p=0.1		p=0.5		p=0.7		p=0.1		p=0.5		p=0.7	
		μ=-2.20		μ=0		μ=0.85		μ=-2.20		μ=0		μ=0.85	
		n_1		n_1		n_1		n_1		n_1		n_1	
	n_r	55	100	55	100	55	100	55	100	55	100	55	100
RB of Subject Variance (%)	5	-35	-36	-31	-32	-33	-33	5	-1	-11	-12	-12	-12
	10	-28	-28	-19	-20	-20	-21	-6	-5	-5	-5	-4	-5
Sim Used (Subject Var)	5	955	991	1000	1000	998	1000	952	990	1000	1000	997	1000
	10	996	1000	1000	1000	1000	1000	996	1000	1000	1000	999	1000
RB of Reader Variance (%)	5	81	23	7	-10	18	-3	49	-4	5	-10	12	-8
	10	25	6	-9	-12	-4	-11	21	0	-2	-6	1	-7
Sim Used (Reader Var)	5	638	714	702	822	708	808	573	621	678	796	670	780
	10	743	827	878	949	841	931	740	810	876	947	842	930

1. Simulation parameters (variance components in logit scale): Subject Variance=1, Reader Variance=0.1
2. p=TPF/Sensitivity, n_r=Number of readers, n_1=Number of diseased subjects
3. Statistics are computed on logit scale
4. RB: Relative Bias of Variance Components
5. Sim Used: Number of simulatons used for calculations out of 1000 simulations
6.* : Sim Used=Used simulations: (1) Nonzero positive standard errors in convergent simulations
(2) PROC GLIMMIX convergence criteria satisfied

Table 38: Model A: Variance Estimation (Medium variance components)

		PL Estimation*						LAPLACE Estimation*					
		p=0.1		p=0.5		p=0.7		p=0.1		p=0.5		p=0.7	
		μ=-2.20		μ=0		μ=0.85		μ=-2.20		μ=0		μ=0.85	
		n_1		n_1		n_1		n_1		n_1		n_1	
	n_r	55	100	55	100	55	100	55	100	55	100	55	100
RB of Subject Variance (%)	5	-45	-46	-39	-39	-40	-41	19	-7	-7	-9	-6	-10
	10	-34	-35	-24	-25	-26	-27	-4	-7	-1	-3	-2	-4
Sim Used (Subject Var)	5	998	999	1000	1000	1000	1000	997	999	1000	1000	1000	1000
	10	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
RB of Reader Variance (%)	5	-18	-26	-28	-29	-29	-28	-11	-28	-18	-21	-20	-22
	10	-23	-22	-21	-19	-22	-20	-12	-15	-9	-9	-10	-10
Sim Used (Reader Var)	5	915	963	964	986	956	981	905	948	961	986	954	979
	10	989	1000	997	1000	998	999	991	1000	998	1000	998	999

1. Simulation parameters (variance components in logit scale): Subject Variance=2, Reader Variance=0.7
2. p=TPF/Sensitivity, n_r=Number of readers, n_1=Number of diseased subjects
3. Statistics are computed on logit scale
4. RB: Relative bias of variance component
5. Sim Used: Number of simulatons used for calculations out of 1000 simulations
6.* : Sim Used=Used simulations: (1) Nonzero positive standard errors in convergent simulations
(2) PROC GLIMMIX convergence criteria satisfied

Table 39: Model A: Variance Estimation (Large variance components)

		PL Estimation*						LAPLACE Estimation*					
		p=0.1		p=0.5		p=0.7		p=0.1		p=0.5		p=0.7	
		μ=-2.20		μ=0		μ=0.85		μ=-2.20		μ=0		μ=0.85	
		n_1		n_1		n_1		n_1		n_1		n_1	
	n_r	55	100	55	100	55	100	55	100	55	100	55	100
RB of Subject Variance (%)	5	-51	-52	-44	-44	-46	-45	20	-11	-6	-8	-7	-9
	10	-37	-37	-29	-28	-31	-30	-3	-5	-3	-2	-3	-3
Sim Used (Subject Var)	5	1000	1000	1000	1000	1000	1000	999	999	1000	1000	1000	1000
	10	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
RB of Reader Variance (%)	5	-32	-34	-34	-35	-33	-35	-18	-28	-18	-22	-18	-23
	10	-27	-24	-23	-22	-23	-22	-11	-13	-7	-9	-7	-9
Sim Used (Reader Var)	5	974	988	994	997	989	993	970	984	994	997	988	995
	10	1000	1000	1000	1000	999	1000	1000	1000	1000	1000	999	1000

1. Simulation parameters (variance components in logit scale): Subject Variance=3, Reader Variance=1.5
2. p=TPF/Sensitivity, n_r=Number of readers, n_1=Number of diseased subjects
3. Statistics are computed on logit scale
4. RB: Relative Bias of Variance Components
5. Sim Used: Number of simulatons used for calculations out of 1000 simulations
6.* : Sim Used=Used simulations: (1) Nonzero positive standard errors in convergent simulations
(2) PROC GLIMMIX convergence criteria satisfied

APPENDIX D: MAGNITUDE OF VARIABILITY COMPONENTS

In order to gauge the magnitude of variance components on logit scale that can ultimately be useful for deriving parameters for reasonable modeling settings, we use the simplest model setting (model A) to perform our illustration. Suppose the between-reader variance is 0.7 and the between-subject variance is 2.0 on the logit scale. To better understand the magnitude of between-reader variability, we can compute the probability/sensitivity corresponding to a low-performing reader (e.g. at 5th percentile of the distribution of the reader random factor) and the sensitivity corresponding to a high-performing reader (e.g. at the 95th percentile). This can be computed exploiting the fact that random effects are normally distributed i.e. $\alpha_i \sim N(0,2)$ for subjects and $\beta_j \sim N(0,0.7)$ for readers and $\mu = logit^{-1}(p) = 0$. In particular, the reader-specific probabilities can now be computed by evaluating the following general integral:

$$p_k = \int logit^{-1}(\mu + \alpha_i + x_k) f(\alpha_i) d\alpha_i,$$

where $x_k = k^{th}$ percentile of the reader random effect distribution. This leads to the following quantities:

Probability of an average subject being classified as positive by a (i) low performing reader (5th percentile) = 0.296 (ii) average performing reader (50th percentile) = 0.499 and (iii) high performing reader (95th percentile) = 0.771

Similarly, in order to assess the between-subject variability, we can estimate the smallest and the largest proportion of readers who will label a specific subject “positive” (or equivalently the probability of a subject with a certain severity of the condition being classified as “positive” by an average reader). Namely, by evaluating the following integral:

$$p_k = \int logit^{-1}\left(\mu + x_k + \beta_j\right) f\left(\beta_j\right) d\beta_j,$$

where $x_k = k^{th}$ percentile of the subject random effect distribution, we obtain the following estimates:

- Probability of a subject with subtle condition (5th percentile) being classified as positive by an average reader = 0.051
- Probability of a subject with average condition (50th percentile) being classified as positive by an average reader = 0.498
- Probability of a subject with obvious condition (95th percentile) being classified as positive by an average reader = 0.952

An alternative way to gauge between-subject variability is to compute the correlation coefficient for the response of two readers (e.g. assuming the readers are very close to the average i.e. have reader-effects as "0").

Since the chosen variance components for this illustration came from fitting subject-specific models to the real observer study [43], it is easy to verify the integral-based computed probabilities by means of computing the empirical estimates of these probabilities using the same dataset. Note that the marginal sensitivity is 0.516 based on the subset of this data (modality=1, $n_1 = 55$, $n_r = 7$). Typically when the sensitivity is 0.5, the subject-specific and marginal sensitivities coincide.

Now, we compute the empirical probability estimates:

The percentiles of the distribution of empirical proportions for each reader ($n_r = 7$) are as follows:

- 5th percentile/minimum observation (low-performing reader) = 0.2909
- 50th percentile (average-performing reader) = 0.4545
- 95th percentile/maximum observation (high-performing reader) = 0.7454

The percentiles of the distribution of empirical proportions for each subject $(n_1 = 55)$ are as follows:

- 5th percentile/3rd observation (subject with subtle condition) = 0
- 50th percentile (subject with average condition) = 0.428
- 95th percentile/53rd observation (subject with obvious condition) = 1

We observed that the empirical estimates are very close to the ones computed numerically. Thus, we illustrated a simple tool to understand the magnitude of the chosen variance components in a cross-correlated study.

We also used similar approach to assess reasonable variance parameter configurations for model D i.e. $\left(\sigma_\alpha^2, \sigma_\beta^2, \sigma_X^2\right)$ given the fixed effect $\mu = 0$ and $\tau = 0, 0.03, 0.06, 0.1$. This simulation model is defined as follows:

$$logit\left(p_{ij}\right) = \mu + \tau * X_i + \alpha_i + \beta_j + \gamma_j * X_i,$$

where $X_i^* \sim$ Unif(1,100) is the lesion size, X_i is centered at 50.5mm i.e. $X_i = X_i^* - 50.5$, $\alpha_i \sim N(0, \sigma_\alpha^2)$ is the subject random effect, $\beta_j \sim N\left(0, \sigma_\beta^2\right)$ is the reader random effect and $\gamma_j \sim N(0, \sigma_X^2)$ is the reader*covariate interaction random effect. Next, we performed the following steps:

Table 40: Obtain value for variance components which lead to reasonable reader-specific probabilities

Model & Distribution of Linear Predictor		$\boldsymbol{\tau}$	$\boldsymbol{X = -49.5}$ $\boldsymbol{X^* = 1}$	$\boldsymbol{X = 0}$ $\boldsymbol{X^* = 50.5}$	$\boldsymbol{X = 49.5}$ $\boldsymbol{X^* = 100}$
			Reader-specific probability range (5th – 95th percentile of distribution of linear predictor)		
$\boldsymbol{logit\left(p_{ijX}\right) = \mu + \tau * X_i + \beta_j = \psi_{ij}}$ $\psi_{ij} \sim N(\mu + \tau * X_i, \sigma_\beta^2)$	$\boldsymbol{\sigma_\beta^2 = 0.2}$	0	0.32 – 0.67	0.32 – 0.67	0.32 – 0.67
		0.03	0.09 - 0.32	0.32 - 0.67	0.67 – 0.90
		0.06	0.02 – 0.09	0.32 – 0.67	0.90 - 0.97
		0.1	0.003 – 0.01	0.32 – 0.67	0.98 – 0.99
$\boldsymbol{logit\left(p_{ijX}\right) = \mu + \tau * X_i + \beta_j + \gamma_j * X_i = \psi_{ij}}$ $\psi_{ij} \sim N(\mu + \tau * X_i, \sigma_\beta^2 + \sigma_X^2 * {X_i}^2)$	$\boldsymbol{\sigma_\beta^2 = 0.2}$ $\boldsymbol{\sigma_X^2 = 0.001}$	0	0.06 – 0.93	0.06 – 0.93	0.06 – 0.93
		0.03	0.01 – 0.76	0.32 – 0.67	0.23 – 0.98
		0.06	0.003 – 0.43	0.32 – 0.67	0.57 – 0.99
		0.1	0 – 0.09	0.32 – 0.67	0.90 – 0.99

Step 1: Chose $\sigma_\beta^2 = 2$ that lead to realistic probabilities for a low and high performing reader for varying values of lesion size. Additionally, we chose $\sigma_X^2 = 0.001$ which together with $\sigma_\beta^2 = 2$ lead to acceptable range of probabilities.

Step 2: Since subjects are expected to vary a lot naturally, it makes sense to choose a slightly bigger value for σ_α^2 as compared to reader variance. Values were chosen similar to those in models A, B and C.

APPENDIX E: IMPLEMENTATION OF THE PROPOSED APPROACH IN SAS

The proposed approach can be implemented in the following seven steps (A through G):

A. Fit a GEE model for each reader separately and record the reader-specific coefficient, e.g., $\widehat{\theta}_j^{\,s}$ where $s = 1, \ldots ,1000$ is the index to denote simulation

Model A SAS code:

```
proc logistic data=data;
by sim reader;
model rating_binary(event="1")=;
estimate "Int" int 1/cl ilink;
ods output  Estimates=est_log;
run;
```

Model B SAS code:

```
proc logistic data=data;
by sim reader;
class truth(ref='0') / param = glm;
model rating_binary(event="1")=truth;
lsmeans truth/ilink cl diff;
ods output  Diffs=diff_log;
run;
```

Model C SAS code:

```
proc glimmix data=data order=data empirical;
by sim reader;
class modality;
model rating_binary(event="1")=modality/dist=binary;
random _residual_/subject=subject type=cs;
lsmeans modality/ilink cl diff;
ods output diffs=diff;
run;
```

Model D SAS code:

```
proc logistic data=data;
by sim reader;
model rating_binary(event="1")=X;
ods output ParameterEstimates=parm;
run;
```

B. Compute average of the estimated coefficient (logit scale) across the readers which gives us the model-averaged estimate of the desired parameter for model A, B, C and D (based on “Estimator 1”):

$$\bar{\hat{\theta}}_{\cdot}^{s} = \frac{1}{n_r}\sum_{j=1}^{n_r}\hat{\theta}_{j}^{s}$$

For model D, we can also use estimate based on “Estimator 2” i.e. average slope coefficient using the following SAS statements:

```
 proc genmod data=data;
 by sim;
 class  subject reader;
 model rating_binary(event="1")=X reader reader*X/dist=bin ;
 repeated subject=subject/ type=cs;
 estimate "avg slope" X 1/e;
 ods output Estimates=est_all_cs;
run;
```

C. Compute sample variance of the estimated coefficient (logit scale) across the readers:

$$\frac{1}{n_r - 1}\sum_{j=1}^{n_r}\left(\hat{\theta}_{j}^{s} - \bar{\hat{\theta}}_{\cdot}^{s}\right)^2$$

D. Compute the average of the estimated variance of the estimated coefficient (logit scale) across the readers:

$$\sum_{j=1}^{n_r}\frac{\hat{V}^{s}\left(\hat{\theta}_{j}^{s}|\beta_j\right)}{n_r}$$

E. Record the variance of the average coefficient for fixed readers $\hat{V}^{s}\left(\bar{\bar{\theta}}_{\cdot}^{s}|\boldsymbol{\beta}\right)$. For this, we fit a marginal model using GEE technique. This average coefficient and its corresponding standard error can be computed by formulating the linear combination that we are interested in and then writing appropriate “estimate” or “lsmestimate” statements within the SAS GLIMMIX/GENMOD procedure. In particular, we use the following SAS code for each type of modeling scenario:

Model A SAS code:

```
proc glimmix data=data empirical;
by sim;
class subject reader;
model rating_binary(event="1")=reader /dist=binary;
random _residual_/subject=subject type=cs;
estimate "Int" int 1/cl ilink;
ods output  Estimates=est_all;
run;
```

Model B SAS code:

```
proc glimmix data=data empirical;
by sim;
class truth subject reader;
model rating_binary(event="1")=truth reader truth*reader/dist=binary;
random _residual_/subject=subject type=cs;
lsmestimate  truth*reader "avg diff" -1 -1 -1 -1 -1 -1 -1 1 1 1 1 1 1 1/divisor=7 e;
ods output LSMEstimates=lsm_all;
run;
```

Model C SAS code:

```
proc glimmix data=data empirical;
by sim;
class modality subject reader;
model rating_binary(event="1")=modality reader modality*reader/dist=binary;
random _residual_/subject=subject type=cs;
lsmestimate  modality*reader "avg diff" 1 1 1 1 1 1 1 -1 -1 -1 -1 -1 -1 -1/divisor=7 e;
ods output LSMEstimates=lsm_all ;
run;
```

Model D SAS code:

```
proc genmod data=data;
by sim;
class subject reader;
model rating_binary(event="1")=X reader reader*X/dist=bin ;
repeated subject=subject/ type=cs;
estimate "avg slope" X 1/e;
ods output Estimates=est_all_cs;
run;
```

F. Combine the various components of the original and bias-corrected variance estimator using the following formulas:

$$\hat{V}^s\left(\bar{\hat{\theta}}_.\right)^{original} = \frac{1}{n_r} * \frac{1}{n_r - 1}\sum_{j=1}^{n_r}\left(\hat{\theta}_J^{\,s} - \bar{\hat{\theta}}_.^{\,s}\right)^2 + \hat{V}^s\left(\bar{\hat{\theta}}_.|\boldsymbol{\beta}\right)$$

$$\hat{V}^s\left(\bar{\hat{\theta}}_.\right)^{bc} = \frac{1}{n_r} * \frac{1}{n_r - 1}\sum_{j=1}^{n_r}\left(\hat{\theta}_J^{\,s} - \bar{\hat{\theta}}_.^{\,s}\right)^2 + \frac{n_r}{n_r - 1} * \hat{V}^s\left(\bar{\theta}_.|\boldsymbol{\beta}\right) - \frac{1}{n_r - 1}\sum_{j=1}^{n_r}\frac{\hat{V}^s\left(\hat{\theta}_J^{\,s}|\beta_j\right)}{n_r}$$

$\hat{V}^s\left(\bar{\hat{\theta}}_.\right)^{bc} = \frac{1}{n_r} * \frac{1}{n_r-1}\sum_{j=1}^{n_r}\left(\hat{\theta}_J^{\,s} - \bar{\hat{\theta}}_.^{\,s}\right)$ in case of using the constraint on variance estimator

G. Create z-based and t-based (containment df = $n_r - 1$) 95% confidence intervals:

$\bar{\hat{\theta}}_.^{\,s} \pm z_{0.975} * \sqrt{\hat{V}^s\left(\bar{\hat{\theta}}_.\right)^{original}}$ and $\bar{\hat{\theta}}_.^{\,s} \pm z_{0.975} * \sqrt{\hat{V}^s\left(\bar{\hat{\theta}}_.\right)^{bc}}$

$\bar{\hat{\theta}}_.^{\,s} \pm t_{0.975,df} * \sqrt{\hat{V}^s\left(\bar{\hat{\theta}}_.\right)^{original}}$ and $\bar{\hat{\theta}}_.^{\,s} \pm t_{0.975,df} * \sqrt{\hat{V}^s\left(\bar{\hat{\theta}}_.\right)^{bc}}$

APPENDIX F: SALAMANDER DATA ILLUSTRATION

As mentioned earlier, the Salamander mating data [1] is a popular binary cross-correlated dataset which has been used by several authors [25, 28, 37, 40, 29] to illustrate and compare results from different estimation techniques. The experiment was conducted in the summer of 1986 and involved 40 animals, 20 rough butt (RB) and 20 whiteside (WS) salamanders, with equal number of males and females. The bernoulli response was the success or failure of a mating between two salamanders. There were four mating types possible – RB/RB, RB/WS, WS/RB and WS/WS (female/male). The following GLMM model was fitted:

$$logit\ Pr\left(\boldsymbol{Y}_{ij} = 1 | \boldsymbol{f}, \boldsymbol{m}\right) = \boldsymbol{X}_{ij}\beta + f_i + m_j,$$

where $\beta = \left(\beta_0, \beta_{WSf}, \beta_{WSm}, \beta_{WSf*WSm}\right)^T$are the fixed-effects determined by salamanders' populations and gender, $f_i \sim N\left(0, \sigma_f^2\right)$ are female random effects assumed independent of $m_j \sim N(0, \sigma_m^2)$ which are the male random effects, $i = 1, \dots, 20$ is female index and $j = 1, \dots, 20$ is male index. $WSf = 1$ if the observation is from a Whiteside female and zero otherwise.

In Table 41, we assembled the results from fitting this model using PL, LA (standard LA), Gibbs Sampling [40], Metropolis-Hastings algorithm using PROC MCMC in SAS [60], Modified LA [37] and Improved LA [28].

Table 41: Salamander data estimates from different GLMM estimation methods

Method	$\hat{\beta}_0$	$\hat{\beta}_{WSf}$	$\hat{\beta}_{WSm}$	$\hat{\beta}_{WSf*WSm}$	$\hat{\sigma}_f^2$	$\hat{\sigma}_m^2$
PL	1.16	-2.57	-0.37	2.80	1.41	0.09
LA	1.34	-2.94	-0.42	3.18	1.58	0.07
Modified LA [37]	1.37	-3.02	-0.44	3.27	1.72	0.19
Improved LA [28]	1.36	-2.99	-0.44	3.24	1.72	0.15
Gibbs Sampling [40]	1.48	-3.25	-0.50	3.62	2.35	0.14
Metropolis-Hastings Algorithm (PROC MCMC, SAS) [60]	1.96	-4.43	-0.76	4.76	5.71	2.41

Table 41 illustrates the amount of underestimation of the variance components under the PL, LA and other approaches for the GLMM estimation. The Modified LA, Improved LA, Gibbs Sampling and Metropolis-Hastings algorithm provided larger estimates than the standard LA and PL approaches. The fixed effect estimates from the PL approach were also attenuated comparing to other approaches.

BIBLIOGRAPHY

1. McCullagh P, Nelder JA. Generalized Linear Models. *London: Chapman and Hall*, 1989.

2. Zhou XH, Obuchowski NA, McClish, DK. Statistical Methods in Diagnostic Medicine. *John Wiley & Sons Inc.*: New York, 1982.

3. Toledano AY, Gastonis C. Ordinal Regression Methodology For ROC Curves Derived From Correlated Data. *Statistics in Medicine* 1996; **15**: 1807-1826.

4. Song HH. Analysis of correlated ROC areas in diagnostic testing. *Biometrics* 1997; **53**: 370-382.

5. Lee MLT, Rosner BA. The average area under correlated receiver operating characteristic curves: a nonparametric approach based on generalized two sample Wilcoxon statistics. *Applied Statistics* 2001; **50**: 337–344.

6. Bandos AI, Rockette HE, Gur D. A permutation test for comparing ROC curves in multireader studies a multi-reader ROC, permutation test. *Academic Radiology* 2006; **13**: 414-420.

7. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. *Investigative Radiology* 1992; **27**: 723-731.

8. Obuchowski NA, Rockette HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations. *Communications in Statistics - Simulation and Computation* 1995; **24**: 285-308.

9. Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis. *Academic Radiology* 2000; **7**: 341-349.

10. Hillis SL, Berbaum KS. Monte Carlo validation of the Dorfman-BerbaumMetz method using normalized pseudovalues and less data-based model simplification. *Academic Radiology* 2005; **12**: 1534-1541.

11. Gallas BD. One-Shot Estimate of MRMC Variance: AUC. *Academic Radiology* 2006; **13** (3): 353-362.

12. Bandos AI, Rockette HE, Gur D. Exact Bootstrap Variances of the Area Under ROC Curve. *Communications in Statistics -Theory and Methods* 2007; **36**: 2443-2461.

13. Hillis SL, Berbaum KS, Metz CE. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. *Academic Radiology* 2008; **15**: 647-661.

14. Ishwaran H, Gastonis CA. A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *The Canadian Journal of Statistics* 2000; **28** (4): 731-750.

15. Wang F, Gastonis CA. Hierarchical models for ROC curve summary measures: Design and analysis of multi-reader, multi-modality studies of medical tests. *Statistics in Medicine* 2008; **27**: 243-256.

16. Toledano AY. Three methods for analysing correlated ROC curves: a comparison in real data sets from multi-reader, multi-case studies with a factorial design. *Statistics in Medicine* 2003; **22**: 2919-2933.

17. Song X, Zhou XH. A marginal model approach for analysis of multi-reader multi-test receiver operating characteristic (ROC) data. *Biostatistics* 2005; **6**: 303-312.

18. Zeger SL, Liang KY, Albert PS. Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics* 1988; **44**(4): 1049-1060.

19. Hricak H, Constantine G, Coakley F *et al.*. Early Invasive Cervical Cancer: CT and MR imaging in preoperative evaluation – ACRIN/GOG comparative study of diagnostic performance and interobserver variability. *Radiology* 2007; **245**: 491-498.

20. Obuchowski NA, Beiden SV, Berbaum KS *et al*.. Multireader, Multicase Receiver Operating Characteristic Analysis: An Empirical Comparison of Five Methods. *Academic Radiology* 2004; **11**: 980-995.

21. Wolfinger R and O'Connell M. Generalized Linear Mixed Models: A Pseudo-Likelihood Approach. *Journal of Statistical Computation and Simulation* 1993; **48**: 233-243.

22. Hoffman EB, Sen PK, Weinberg CR. Within-Cluster Resampling. *Biometrika* 2001; **88**: 1121-1134.

23. Burnham KP and Anderson DR. Model Selection and multimodel inference: a practical information-theoretic approach. *Springer* 2nd Edition, 2002.

24. Rubin DB. Multiple Imputation for Nonresponse in Surveys. *John Wiley & Sons Inc.*: New York, 1987.

25. Capanu M, Gonen M, Begg CB. An assessment of estimation methods for generalized linear mixed models with binary outcomes. *Statistics in Medicine* 2013; **32** (26): 4550-4566.

26. Pinheiro JC, Bates DM. Approximation to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics* 1995; **4**(1): 12-35.

27. Joe H. Accuracy of Laplace approximation for discrete response mixed models. *Journal of Computational Statistics and Data Analysis* 2008; **52**: 5066-5074.

28. Ruli E, Sartori N, Ventura L. Improved Laplace approximation for marginal likelihoods. *Electronic Journal of Statistics* 2016; **10**: 3986-4009.

29. Breslow NE, Clayton DG. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* 1993; **88**: 9-25.

30. Lindstrom MJ, Bates DM. Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics* 1990; **46**: 673-687.

31. Breslow NE, Lin X. Bias Correction in Generalized Linear Mixed Models With a Single Component of Dispersion. *Biometrika* 1995; **82**: 81-91.

32. Pinheiro J, Chao E. Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized linear Models. *Journal of Computational and Graphical Statistics* 2006; **15**: 58-81.

33. Ng E, Carpenter JR, Goldstein H *et al.*. Estimation in generalized linear mixed models with binary outcomes by simulated maximum likelihood. *Statistical Modelling* 2006; **6**: 23-42.

34. Raudenbush S, Yang M, Yosef M. Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Approximation. *Journal of Computational and Graphical Statistics* 2000; **9**: 141-157.

35. Bellio R, Varin C. A pairwise likelihood approach to generalized linear models with crossed random effects. *Statistical Modelling* 2005; **5**: 217-227.

36. Lin X, Breslow NE. Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion. *Journal of the American Statistical Association* 1996; **91**: 1007-1016.

37. Shun Z. Another Look at the Salamander mating Data. A Modified Laplace Approximation Approach. *Journal of the American Statistical Association* 1997; **92**: 341-349.

38. Shun Z, McCullagh P. Laplace Approximation of High Dimensional Integrals. *Journal of the Royal Statistical Society* 1995; **57**: 749-760.

39. Withanage N, Leon A, Rudnisky C. Joint Estimation of Multiple Disease-Specific Sensitivities and Specificities via Crossed Random Effects Models for Correlated Reader-Based Diagnostic Data: Application of Data Cloning. *Statistics in Medicine* 2014; **34**: 3916-3928.

40. Karim RM, Zeger SL. Generalized Linear Models with Random Effects; Salamander Mating Revisited. *Biometrics* 1992; **48**: 631-644.

41. Yoonsang K, Choi YK, Emery S. Logistic Regression with multiple Random effects: A simulation study of estimation methods and statistical packages. *The American Statistician* 2013; **67**(3): 171-182.

42. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. *Oxford University Press*: New York, 2003.

43. Slasky BS, Gur D, Good WF *et al.*. Receiver Operating Characteristic Analysis of Chest Image Interpretation with Conventional, Laser-printed, and High-Resolution Workstation Images. *Radiology* 1990; **174**:775-80.

44. Stroup WW. Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. *CRC Press*.

45. Neuhaus JM, Kalbfleisch JD, Hauck WM. A Comparison of Cluster-Specific and Population-Averaged Approaches for Analyzing Correlated Binary Data. *International Statistical Review* 1991; **59:** 25-35.

46. Hoeffding W. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* 1948; **19**: 293-325.

47. Arvesen JN. Jackknifing U-Statistics. *Annals of Mathematical Statistics* 1969; **40**: 2076-2100.

48. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Area under Two or More Correlated Receiver Operating Characteristics Curves: A Nonparametric Approach. *Biometrics* 1988; **44**: 837-845.

49. Sen PK. On some convergence properties of U-statistics. *Calcutta Statistical Association Bulletin* 1960; **10**:1-18.

50. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* 1975; **12**: 387-415.

51. Wieand S, Gail MH, James BR *et al.*. A family of non-parametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 1988; **76**: 585–592.

52. Gallas BD, Pennello GA, Myers KJ. Multireader multicase variance analysis of binary data. *Journal of Optical Society of America* 2007; **24**(12): B70-80.

53. Bandos AI, Rockette HE, Gur D. Exact bootstrap variances of the area under ROC curve. *Communications in Statistics-Theory and Methods* 2007; **36**: 2443-2461.

54. Dodd LE, Pepe MS. Semiparametric Regression for the Area under the Receiver Operating Characteristic Curve. *Journal of the American Statistical Association* 2003; **98**: 409-417.

55. Dodd LE, Pepe MS. Partial AUC Estimation and Regression. *Biometrics* 2003; **59**: 614-623.

56. Gallas BD, Bandos AI, Samuelson FW *et al.*. A framework for random-effects ROC analysis: Biases with the bootstrap and other variance estimators. *Communications in Statistics - Theory and Methods* 2009; **38**(15): 2586-2603.

57. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC studies. *Statistics in Medicine* 2007; **26**(3): 596-619.

58. Chen W, Wunderlich A, Petrick N and Gallas BD. Multireader multicase reader studies with binary agreement data: simulation, analysis, validation, and sizing. *Journal of Medical Imaging* 2014; **1**(3): 031011.

59. Brown H, Prescott R. *Applied Mixed Models in* Medicine. John Wiley & Sons, Ltd, 2006.

60. Chen F, Brown G, Stokes M. Fitting Your Favorite Mixed Models with PROC MCMC. In *Proceedings of the SAS Global Forum 2016 Conference*. Cary, NC: SAS Institute Inc. https://support.sas.com/resources/papers/proceedings16/SAS5601-2016.pdf.