DIGITALCOMMONS
@WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 16 | Issue 2 Article 15

December 2017

Modeling Agreement between Binary Classifications of Multiple Raters in R and SAS

Aya A. Mitani
Boston University, amitani@bu.edu

Kerrie P. Nelson
Boston University, kerrie@bu.edu

Recommended Citation

Mitani, A. A., & Nelson, K. P. (2017). Modeling Agreement between Binary Classifications of Multiple Raters in R and SAS. Journal of Modern Applied Statistical Methods, 16(2), 277-309. doi: 10.22237/jmasm/1509495300

Modeling Agreement between Binary Classifications of Multiple Raters in R and SAS

Cover Page Footnote

The authors are grateful for the support provided by grant R01-CA-17246301 from the United States National Institutes of Health. The AIM study was supported by the American Cancer Society, made possible by a generous donation from the Longaberger Company's Horizon of Hope®Campaign (SIRSG-07-271, SIRSG-07- 272, SIRSG-07-273, SIRSG-07-274, SIRSG-07-275, SIRGS-06-281, SIRSG-09- 270, SIRSG-09-271), the Breast Cancer Stamp Fund, and the National Cancer Institute Breast Cancer Surveillance Consortium (HHSN261201100031C). The cancer data used in the AIM study was supported in part by state public health departments and cancer registries in the U.S., see http://www.bcsc- research.org/work/acknowledgement.html. We also thank participating women, mammography facilities, radiologists, and BCSC investigators for their data. A list of the BCSC investigators is provided at: http://www.bcsc-research.org/.

Journal of Modern Applied Statistical Methods
November 2017, Vol. 16, No. 2, 277-309.
doi: 10.22237/jmasm/1509495300

ISSN 1538 − 9472

Modeling Agreement between Binary Classifications of Multiple Raters in R and SAS

Aya A. Mitani
Boston University
Boston, MA

Kerrie P. Nelson
Boston University
Boston, MA

Cancer screening and diagnostic tests often are classified using a binary outcome such as diseased or not diseased. Recently large-scale studies have been conducted to assess agreement between many raters. Measures of agreement using the class of generalized linear mixed models were implemented efficiently in four recently introduced R and SAS packages in large-scale agreement studies incorporating binary classifications. Simulation studies were conducted to compare the performance across the packages and apply the agreement methods to two cancer studies.

Keywords: Agreement, binary classifications, Cohen's kappa, Fleiss' kappa, generalized linear mixed model, multiple raters

Introduction

Assessing the strength of agreement between physicians' ratings of screening test results is of primary interest because an effective diagnostic procedure is dependent upon high levels of consistency between raters. However, in practice, substantial discrepancies are often observed between physicians' ratings and is considered a major issue in many common screening tests including mammography and diagnosis of invasive bladder cancer (Beam, Conant, & Sickles, 2002; Compérat et al., 2013; Elmore, Wells, Lee, Howard, & Feinstein, 1994; Onega et al., 2012). This has motivated large-scale studies to examine accuracy and agreement between physicians' ratings and to investigate factors that may play an influential role on the consistency of ratings, precipitating a pressing need for statistical methods of agreement that can flexibly accommodate classifications of a large number of raters.

Aya Mitani is a graduate student in the Department of Biostatistics. Email her at: amitani@bu.edu. Kerrie Nelson is a Research Associate Professor of Biostatistics. Email her at: kerrie@bu.edu.

The outcome of a patient's screening test may be classified using a binary categorical scale (for example, diseased or not diseased) based upon the physician's (subjective) interpretation of the screening test result. For example, mammographic results are often categorized as requiring recall or no recall of a patient for further testing and bladder cancer images may be classified as indicating invasive or non-invasive cancer (Compérat et al., 2013). In this paper we focus on large-scale agreement studies where more than two raters' classifications are made using a binary categorical scale.

When multiple raters participate in a large-scale agreement study, only a limited number of methods are available to assess agreement between their binary ratings in a unified and comprehensive approach. Summary measures include Fleiss' measure of agreement and Shrout and Fleiss' intraclass correlation coefficient (ICC) (Fleiss & Cuzick, 1979; Fleiss, 1971; Shrout & Fleiss, 1979). Modeling approaches include a Bayesian generalized linear mixed model (GLMM) with nested random effects and an approach based upon GLMMs with crossed random effects (Hsiao, Chen, & Kao, 2011; Nelson & Edwards, 2008, 2010). Log linear models, another modeling approach, are best-suited for modeling agreement between two or three raters (Agresti, 1989; Tanner & Young, 1985).

Due to a lack of statistical methods that can easily be implemented in practice for studies with multiple raters, clinical research papers tend to instead focus on comparing agreement using pairwise approaches (i.e. comparing between each pair of raters at a time) which can be inefficient, lending itself to several summary measures and often complex or disjointed interpretation of agreement (Ciatto et al., 2005; Compérat et al., 2013; Epstein, Allsbrook, Amin, Egevad, & ISUP Grading Committee, 2005; Ooms et al., 2007).

Until recently, various modeling approaches such as Nelson and Edwards' (2008) GLMM-based method have been challenging to implement due to a lack of availability in standard statistical software packages for modeling GLMMs and a necessity for sophisticated programming skills. However, recent advances in statistical software packages including R (R Core Team, 2014) and SAS (Cary, NC: SAS Institute) have led to much improved and efficient procedures for fitting complex models including GLMMs with crossed random effects. In this paper we demonstrate how Fleiss' kappa for multiple raters and Nelson and Edwards' GLMM modeling approach can easily be implemented in four R packages and in SAS software to assess agreement in large-scale studies with binary classifications.

The aim of this study is to compare the performance of the different software packages using extensive simulation studies to assess the impact of normally and non-normally distributed (symmetric and skewed) random effects and sample size on parameter estimation and the calculation of the agreement statistics. It is motivated by two large-scale agreement studies. The first is a study of 119 community radiologists assessing 109 mammograms as recall or no recall conducted by the Breast Cancer Surveillance Consortium (BCSC) (Onega et al., 2012). The second study conducted by Compérat et al. (2013) involved 8 pathologists reviewing 25 bladder cancer specimens for the presence or absence of invasive cancer. For each of these two studies we implement the different agreement approaches described above in each of the four statistical software packages and assess levels of agreement between the multiple raters. We also demonstrate how the classifications of individual raters can be assessed from their random effect terms.

Models and Measures of Agreement for Multiple Raters

GLMM Approach An approach based upon GLMMs with a crossed random effects structure can be implemented to assess levels of agreement between multiple raters' binary classifications (Nelson & Edwards, 2008, 2010). This approach, unlike many others, is intended to accommodate the ratings of multiple raters, does not grow increasingly complex as the number of raters increases, and can accommodate missing data where some raters do not classify every test result (Ibrahim & Molenberghs, 2009). Derived from this model is a chance-corrected measure of agreement which incorporates data from the entire sample of subjects. Its value, unlike Cohen's kappa statistics, is robust to the underlying prevalence of the disease. A brief description of the method is following; full details can be found in Nelson and Edwards (2008, 2010). Our setup assumes a sample of J raters ($j = 1,\dots, J$) each independently classifying a sample of I subjects ($i = 1,\dots, I$) generating the set of binary outcomes Y_{ij}, each taking the value 0 or 1.

The binary GLMM with a probit link function and crossed random effects models the probability that a subject's test result is classified as a success, $\Pr(Y_{ij} = 1)$ as follows:

$$\Phi^{-1}\left(\Pr\left(Y_{ij}=1 \mid u_i, v_j\right)\right)=\eta+u_i+v_j \tag{1}$$

where η is the intercept and u_i and v_j are the random effects for the i^{th} subject and the j^{th} rater, respectively. The subject random effects u_i ($i = 1,\dots, I$) and the rater random effects v_j ($j = 1,\dots, J$) are assumed normally distributed with mean 0 and variances σ_u^2 and σ_v^2, respectively. A positive random effect value for u_i indicates a test result that is more likely than other test results to be classified as a success over many raters. A positive value for v_j suggests a rater who is liberal in classifying a subject as a success over their classification of many such test results. The chance-corrected model-based kappa has been derived previously and takes the form

$$\kappa_m = 1 - 4\int_{-\infty}^{+\infty} \Phi\left(\frac{z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\left\{1-\Phi\left(\frac{z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\right\}\phi(z)\,dz \quad 0 \le \kappa_m \le 1 \tag{2}$$

with its variance derived using the multivariate delta method as

$$\operatorname{var}(\hat{\kappa}_m) = 16\Bigg\{\left[\int_{-\infty}^{+\infty}\left(\frac{1}{2\hat{\rho}(1-\hat{\rho})}\right)\left(\frac{z\sqrt{\hat{\rho}}}{\sqrt{1-\hat{\rho}}}\right)\phi\left(\frac{z\sqrt{\hat{\rho}}}{\sqrt{1-\hat{\rho}}}\right)\left[1-2\Phi\left(\frac{z\sqrt{\hat{\rho}}}{\sqrt{1-\hat{\rho}}}\right)\right]\phi(z)\,dz\right]^2$$
$$\times\left[\left(\frac{\hat{\sigma}_v^2+1}{\left(\hat{\sigma}_T^2\right)^2}\right)^2\left(\frac{2\hat{\sigma}_u^4}{I}\right)+\left(\frac{\hat{\sigma}_u^2}{\left(\hat{\sigma}_T^2\right)^2}\right)^2\left(\frac{2\hat{\sigma}_v^4}{I}\right)\right]\Bigg]\Bigg\}$$

where $\sigma_T^2 = \sigma_u^2 + \sigma_v^2 + 1$ and $\rho = \sigma_u^2 / \sigma_T^2$. Full details on the derivation of κ_m and its variance can be found in Nelson and Edwards (2008, 2010). The summary measure of agreement κ_m takes values between 0 and 1 and is interpreted in a similar manner to Cohen's original kappa where a value close to 0 indicates little or no chance-corrected agreement and values closer to 1 reflect strong chance-corrected agreement between raters (Cohen, 1968; Landis & Koch, 1977).

The marginal likelihood function for the GLMM model takes the form:

$$\mathrm{L}(\boldsymbol{\theta};\mathbf{Y}) =$$
$$\int_{\mathbf{u}}\int_{\mathbf{v}}\left\{\prod_{i=1}^{I}\prod_{j=1}^{J}\Phi\left(\eta+u_i+v_j\right)\left[1-\Phi\left(\eta+u_i+v_j\right)\right]\right\}\times\left(\prod_{i=1}^{I}\frac{e^{-\frac{u_i^2}{2\sigma_u^2}}}{\sqrt{2\pi\sigma_u^2}}\right)\left(\prod_{j=1}^{J}\frac{e^{-\frac{v_j^2}{2\sigma_v^2}}}{\sqrt{2\pi\sigma_v^2}}\right)d\mathbf{u}d\mathbf{v}$$

where **Y** is the vector of all the binary classifications of all raters.

The inclusion of the crossed random effects leads to a high-dimensional likelihood function, thus no closed form solution for maximizing the marginal likelihood function is available. Hence, approximate maximum likelihood methods are explored for estimating the parameters. Adaptive Gaussian quadrature is not a viable technique for obtaining approximate maximum likelihood estimates due to the large number of random effects. Instead, estimates of the parameters $\boldsymbol{\theta} = \left(\eta, \sigma_u^2, \sigma_v^2\right)$ can be obtained by fitting the GLMM using an approximate maximum likelihood approach such as the Monte-Carlo expectation-maximum (MCEM) algorithm provided in McCulloch (1997) and Kuk and Cheng (1997). These methods based on Monte-Carlo Markov-Chain (MCMC) (Karim & Zeger, 1992; Kuk & Cheng, 1997; McCulloch, 1997) are feasible in obtaining approximate maximum likelihood estimates for these GLMM models, however they often take a large amount of computational programming and running time and are sometimes unstable, not reaching convergence. Recently a multivariate Laplacian approximation technique, which is computationally very efficient and stable, has been implemented in R and SAS for fitting GLMMs with crossed random effects. In the multivariate Laplacian approximation method, large-sample approximate standard errors are estimated by taking the square-roots of the diagonals of matrix **H** at convergence, i.e.

$$\mathrm{se}\left(\hat{\boldsymbol{\theta}}\right) = \sqrt{\mathrm{diag}\left\{-\left[\mathbf{H}\left(\hat{\boldsymbol{\theta}}\right)\right]^{-1}\right\}}$$

where

$$\mathbf{H} = \frac{\partial^2 \, \mathrm{l}\left(\boldsymbol{\theta}; \mathbf{u}, \mathbf{v}, \mathbf{y}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\boldsymbol{t}}}$$

is the second-order derivative of the log-likelihood function l(**θ**; **u**, **v**, **y**) evaluated at the approximate maximum likelihood estimates of **θ** and is generated during the model-fitting process.

Fleiss Kappa for Multiple Raters Fleiss (1971) described a generalized Kappa statistic which extends Scott's pi (Scott, 1955) in order to accommodate multiple raters and multiple categories. Later, Fleiss and Cuzick (1979) introduced a version of their kappa statistic for binary classifications with unequal number of ratings per test result. Briefly, it is structured as follows: For I subjects ($i = 1,\ldots, I$) under study, let n_i denote the number of raters rating the ith subject and let x_i denote the number of positive ratings on the ith subject. Defining $p_i = x_i / n_i$ as the proportion of positive ratings for each subject,

$$\bar{n} = \frac{\sum_i n_i}{I}$$

as the mean number of raters for each subject, and

$$\bar{p} = \frac{\sum_i x_i}{(I\bar{n})}$$

as the overall proportion of positive ratings, the Fleiss' kappa for agreement takes the form

$$\hat{\kappa}_{\mathrm{F}} = 1 - \frac{\sum_i n_i p_i (1 - p_i)}{I(\bar{n} - 1)\bar{p}(1 - \bar{p})} \tag{3}$$

with variance

$$\mathrm{Var}(\hat{\kappa}_{\mathrm{F}}) = \frac{2(\bar{n}_{\mathbf{H}} - 1)}{I\bar{n}_{\mathbf{H}}(\bar{n} - 1)^2} + \frac{(\bar{n} - \bar{n}_{\mathbf{H}})[1 - 4\bar{p}(1 - \bar{p})]}{I\bar{n}\bar{n}_{\mathbf{H}}(\bar{n} - 1)^2 \bar{p}(1 - \bar{p})} \tag{4}$$

where $\bar{n}_{\mathrm{H}}$ is defined as the harmonic mean number of raters for each subject,

$$\bar{n}_{\mathrm{H}} = \frac{I}{\sum_i \frac{1}{n_i}}$$

When the number of raters per subject is constant, $\hat{\kappa}_F$ is equivalent to the Fleiss kappa statistic introduced by Fleiss in 1971 (1971; Fleiss, Nee, & Landis, 1979; Fleiss & Cuzick, 1979). Fleiss' kappa take values between 0 and 1 and are interpreted in a similar manner to Cohen's original kappa (Cohen, 1968), where 0 indicates no chance-corrected agreement and values closer to 1 suggest strong chance-corrected agreement between the raters. For further details on this summary agreement measures, see Fleiss (1971) and Fleiss and Cuzick (1979). A potential drawback of Fleiss' kappa includes vulnerability to marginal prevalence issues in a similar manner to Cohen's kappa.

Statistical Software Packages in SAS and R

Until recently GLMMs with crossed random effects have been challenging to implement in standard software packages, instead requiring sophisticated programming skills and often computationally intensive algorithms (Kuk & Cheng, 1997; McCulloch, 1997). However, recent advances in SAS and R allow for these models to be fit efficiently by using packages or procedures that do not require programming skills. Four of the available procedures that are capable of fitting GLMMs with crossed random effects allowing for a probit link function in R and SAS are (we will briefly discuss each in turn):

a) R – clmm function in ORDINAL package
b) R – glmer function in LME4 package
c) R – MCMCglmm package
d) SAS – GLIMMIX procedure

ORDINAL Package in R The ORDINAL package (Christensen, 2013) was recently added to R and is primarily intended for fitting cumulative mixed models such as ordered regression models, proportional odds and proportional hazards models for grouped survival times, and ordered logit/probit models. The *clmm* function in the ORDINAL package allows GLMMs with crossed random effects to be fitted with a probit link function. Estimation procedures include the Laplace approximation and Gaussian quadrature but we are restricted to the Laplace method to fit our model of interest with crossed random effects. While this package is primarily intended to fit ordinal models, it also provides an efficient approach for estimating parameters in a binary GLMM. For fitting our GLMM of interest, the probit link function and the random effects structure can be specified in the model formula. Solutions to the random effects for subjects and raters are computed based on the conditional modes, the points at which the conditional density of the estimated random effects are maximized. We are not aware of any studies comparing the performance of the ORDINAL package to that of other packages such as LME4.

LME4 Package in R The *glmer* function in LME4 package is perhaps the most widely-used function to fit GLMMs in R. Its default approximation method is the Laplace approximation and the function accommodates crossed random effects. To fit the model of interest, family = binomial(link = “probit”) and the random effects structure are specified in the model formula. Similarly to the ORDINAL package, the solution to the random effects are computed based on the conditional modes.

MCMCglmm Package in R The above packages use a frequentist approach to fit GLMMs. The MCMCglmm package uses a Bayesian approach and can fit GLMMs with crossed random effects and a probit link function (Hadfield, 2010). Priors for the fixed effects and variance structures for the random effects and residuals need to be specified. In MCMCglmm, the prior distribution for the fixed effects are assumed multivariate normal with the user specifying the parameters, and the prior distribution for both the R-structure for the error distribution and the G-structure for the random effects variance covariance matrices are assumed inverse-Wishart, again with the user specifying the parameters (Hadfield, 2015). The function *posterior.mode* or *posterior.mean* is used to obtain solutions to the random effects for each subject and rater.

GLIMMIX Procedure in SAS In a similar manner to the ORDINAL and LME4 packages in R, the GLIMMIX procedure in SAS relies on the Laplace approximation for estimation of GLMMs with crossed random effects. The solution to the random effects are again computed based on the conditional modes.

Another procedure in SAS that fits GLMMs is the NLMIXED procedure. The NLMIXED procedure estimates the parameters by integral approximation methods through adaptive Gaussian quadrature. However, at present, the procedure cannot accommodate a crossed random effects structure so it will not be examined here.

Methodology

Although the LME4, MCMCglmm, and PROC GLIMMIX packages were described for estimation in various binary GLMM models (Kim, Choi, & Emery, 2013; Li, Lingsma, Steyerberg, & Lesaffre, 2011; Zhang et al., 2011), the performance of the ORDINAL package has not yet been reported for binary outcomes nor for the calculation of agreement measures. Our focus in this paper is to explore the use of these four aforementioned packages in R and SAS to calculate the measures of agreement for multiple raters classifying test results using a binary scale. To achieve this, we conducted extensive simulation studies to compare the performance of the four packages with regards to estimation of GLMM model parameters and the summary agreement measures. One important motivation for conducting these simulation studies is to ensure that reasonably unbiased estimates of the model-based measure of agreement κ_m are obtained from the existing packages.

Simulation studies were conducted under scenarios that varied in sample size (number of subjects and raters), random effects components, distributions of the random effects, and the choice of priors for the MCMCglmm Bayesian method. The various simulation scenarios we explored are displayed in Table 1. Part I of the simulations had normally-distributed random effects, while parts II and III had non-normally-distributed random effects. In part II, the random effects were symmetric (mixture of two normal distributions and uniform distribution) and, in part III, at least one of the random effects were skewed (exponential, Gamma or chi-squared distribution). For each part of the simulations, we evaluated four scenarios. The first scenario (Scenario 1) resembled the BCSC breast cancer data set to verify that our methods perform well in this setting and others (η = −0.1, $\sigma_u^2 = 1.5$, $\sigma_v^2 = 0.2$). In Scenario 2, the variance of the rater random effects was set to be larger than the variance of the subject random effects (η = 1, $\sigma_u^2 = 1$, $\sigma_v^2 = 5$). In Scenario 3, the variance of the subject random effects was set to be larger than the variance of the rater random effects (η = 1, $\sigma_u^2 = 5$, $\sigma_v^2 = 1$). In Scenario 4, the variances of both random effects were set as large $\eta = 1$, $\sigma_u^2 = 10$, $\sigma_v^2 = 10$). Regardless of the random effects distribution, the variances of the subject and rater random effects were kept constant for each scenario (i.e. for Scenario 1, the variance of the subject random effects was set as 1.5 for normal, non-normal symmetric, and skewed distributed random effects). Within each scenario, one was larger in sample size with 150 subjects and 100 raters (Scenario #a) while the other was smaller with 100 subjects and 50 raters (Scenario #b). [Table 1]

For each simulation scenario in part I (normally distributed random effects), one thousand datasets were generated using R in the following manner: First, *I* subject random effects and *J* rater random effects were randomly generated from $\mathrm{N}\left(0,\sigma_u^2\right)$ and $\mathrm{N}\left(0,\sigma_v^2\right)$ distributions, respectively. For each (*ij*)th observation, the probability of the *j*th rater correctly classifying the *i*th subject was generated according to the ordinal probit GLMM

$$p_{ij} = \Pr\left(Y_{ij} = 1 \mid u_i, v_j\right) = \Phi\left(\eta + u_i + v_j\right), \quad i = 1,\ldots,I;\ j = 1,\ldots,J$$

using the *qnorm* function in R.

Table 1. List of parameters used to generate simulated data sets for each scenario (Number of simulations per scenario = 1,000)

Scenario	*I*	*J*	True η	Distribution of u_i*	Distribution of v_j*
					I. Normally distributed random effects
1a	150	100	-0.1	N(0, 1.5)	N(0, 0.2)
1b	100	50			
2a	150	100	1	N(0, 1)	N(0, 5)
2b	100	50			
3a	150	100	1	N(0, 5)	N(0, 1)
3b	100	50			
4a	150	100	1	N(0, 10)	N(0, 10)
4b	100	50			
					II. Non-normally distributed random effects (Symmetric)
1a	150	100	-0.1	0.5N(-1, 0.5) + 0.5N(1, 0.5)	Unif(-0.775, 0.775)
1b	100	50			
2a	150	100	1	0.5N(-0.8, 0.36) + 0.5N(0.8, 0.36)	Unif(-3.87, 3.87)
2b	100	50			
3a	150	100	1	0.5N(-2, 1) + 0.5N(2, 1)	Unif(-1.73, 1.73)
3b	100	50			
4a	150	100	1	0.5N(-3, 1) + 0.5N(3, 1)	Unif(-5.48, 5.48)
4b	100	50			
					III. Non-normally distributed random effects (Skewed)
1a	150	100	-0.1	$\mathrm{Exp}\left(\sqrt{1/1.5}\right)$	$\mathrm{Gamma}\left(4, \sqrt{20}\right)$
1b	100	50			
2a	150	100	1	N(0, 1)	Gamma(5, 1)
2b	100	50			
3a	150	100	1	Gamma(5, 1)	Unif(-1.73, 1.73)
3b	100	50			
4a	150	100	1	$\chi^2_{\mathrm{df}=5}$	N(0, 10)
4b	100	50			

Note: * Mean and variance shown for normal distributions, N(μ, σ^2)

A binary classification Y_{ij} was then randomly generated for each observation from the corresponding Bernoulli distribution with probability p_{ij}. To assess the impact of a misspecified random effects distribution in GLMM, we also generated data with non-normally-distributed subject and rater random effects (Litière, Alonso, & Molenberghs, 2008). In the symmetric non-normal random effects scenarios, the random effects of the subjects were randomly sampled from a symmetric mixture of two normal distributions with mean of 0 and the same variance as the corresponding normal distribution. Each i^{th} subject was assigned a number generated from uniform distribution with (0, 1) support. If the assigned number was less than 0.5, the random effect of the subjects was sampled from the first of the two normal distributions. Otherwise, the random effect of the subjects was sampled from the second of the two normal distributions. The rater random effects were randomly sampled from a uniform distribution with mean of 0 and the same variance as the corresponding normally distributed random effects. In the skewed random effects scenarios, the random effects of the subjects and raters were randomly sampled from a combination of various skewed distributions (exponential, Gamma, and chi-squared) and normal and uniform distributions. For the true random effects distribution to have mean 0, an assumption of GLMM, each of the skewed random effects distributions was centered by subtracting its true mean value. See Table 1 for the parameters and distributions of random effects used in each set of scenarios.

The binary GLMM in equation (1) was then fitted to each of the one thousand simulated datasets using each of the four statistical packages (PROC GLIMMIX, LME4, ORDINAL, and MCMCglmm). With the MCMCglmm package, two different sets of priors were used for each scenario. We specified the variances of the subject and rater random effect terms to follow an inverse-Wishart (IW) distribution, which is comprised of two parameters: the scale parameter V, and the degree of freedom parameter ν, also referred to as the degree of belief parameter. For the first set of priors, denoted by “MCMCglmm1”, we let the variance of the random effects follow an IW distribution with $V = 1$ and $\nu = 1$, and for the second set of priors, denoted by “MCMCglmm10”, we let the variance of the random effects follow an IW distribution with $V = 10$ and $\nu = 1$. Under Scenario 1a, we also used the uninformative prior specification with $V = 1$ and $\nu = 0.002$ which is used frequently for variance structures (Hadfield, 2015).

The GLMM parameters of interest estimated for each dataset were η, σ_u^2, and σ_v^2. These parameter estimates were then used to compute the model-based measure of agreement, $\hat{\kappa}_m$, and its variance, $\operatorname{Var}\left(\hat{\kappa}_m\right)$. Fleiss’ agreement measure $\hat{\kappa}_{\mathrm{F}}$ was also calculated for each dataset.

Results

Simulation results from normally-distributed and symmetric non-normally-distributed random effects datasets are presented in Tables 2 and 3, respectively, for large sample size ($I = 150$, $J = 100$). Results from skewed random effects are presented in Supplementary Table 3. For each simulation scenario, the mean of the 1,000 estimates (Mean Estimate) and the mean of the 1,000 model-based standard errors (Mean SE) estimated for each of the model parameters, η, σ_u^2, and σ_v^2 from each of the four software packages are reported. The mean of the 1,000 estimates and standard errors for measure of agreement κ_m are also reported for each set of simulations. The coverage probability (the percent of times the 95% confidence interval for $\hat{\kappa}_m$ included the true κ_m value) of κ_m over the 1,000 simulated datasets is also reported for each of the four statistical packages, as well as the convergence rate of the GLMM based on the number of times the model was able to produce the standard errors for σ_u^2, and σ_v^2 estimates. Also, the mean estimated Fleiss' kappa ($\hat{\kappa}_{\mathrm{F}}$) and the mean standard error for each simulation scenario are reported. The focus is on results from scenarios with large sample size ($I = 150$, $J = 100$). Simulation results from scenarios with small sample size ($I = 100$, $J = 50$) followed a similar pattern to those from scenarios with large sample size. Full details of the simulation results of small sample size can be viewed in Supplementary Tables 1, 2, and 4.

GLMM Parameter Estimates

Minimal biases were observed in the estimation of η across the four packages when the random effects were normally distributed. Slightly larger biases were observed under the scenarios with non-normal random effects and when one of the variance components, σ_u^2 or σ_v^2, was 5 and the other was 1 (Scenarios 2 and 3). These biases tended to be larger under the MCMCglmm package for both sets of priors. Biases in the estimation of η were largest under the scenarios with skewed random effects but varied little among the different packages. [Supplementary Table 3] Due to the model format used in its package, the η estimates produced from the ORDINAL package have an opposite sign from those produced from other packages. To make the comparison between packages easier, we present η estimates with consistent signs in the tables.

Observe more variability in biases of the random effects variance component estimates between the different packages. Generally, with normally-distributed random effects, ORDINAL, LME4, and PROC GLIMMIX tended to slightly underestimate σ_u^2 and σ_v^2 while MCMCglmm1 and MCMCglmm10 tended to overestimate them. For example, under Scenario 1a, σ_u^2 were 1.492, 1.500, and 1.493 for ORDINAL, LME4, and PROC GLIMMIX, respectively, while they were 1.530 and 1.613 for MCMCglmm1 and MCMCglmm10, respectively. [Table 2] For the symmetric non-normal random effects, most packages overestimated σ_u^2, and σ_v^2 under Scenarios 1 and 2. Under Scenario 3 ($\sigma_u^2 = 5$ and $\sigma_v^2 = 1$), ORDINAL, LME4, and PROC GLIMMIX estimated σ_v^2 with minimal bias (0.993, 0.999, and 0.993, respectively) but overestimated σ_u^2 (5.816, 5.758, and 5.816, respectively). MCMCglmm1 and MCMCglmm10 also overestimated σ_u^2 (6.263 and 6.386, respectively). [Table 3] For the skewed random effects, all packages tended to overestimate the larger of the two variances under Scenarios 1, 2, and 3. Under Scenario 4, all packages underestimated σ_u^2 while σ_v^2 was estimated with smaller biases. [Supplementary Table 3]

The ORDINAL package and the GLIMMIX procedure produced identical GLMM parameter estimates to the third decimal place confirming that these two packages employ virtually identical multivariate Lapacian procedures. With the exception of LME4, the other three packages exhibited very stable estimation procedures with usually a 100% convergence success rate over each set of 1,000 simulated data sets, for both normally- and non-normally-distributed random effects. The LME4 package proved to be consistently less stable compared to all the other packages, with convergence rates ranging from 79.8% to 99.9%. In particular, convergence rate for LME4 tended to be worse for simulation scenarios with large random effects variances and for non-normally-distributed random effects distribution (symmetric and skewed). The average time to fit one GLMM for the larger data set was 9, 8, 109, 104, and 27 seconds for ORDINAL, LME4, MCMCglmm1, MCMCglmm10, and PROC GLIMMX, respectively, indicating that all four packages were able to fit these models in a computationally efficient manner.

Agreement Measures

The parameter κ_m was estimated with minimal bias in all simulation scenarios and across all four packages and various values of σ_u^2 and σ_v^2 when the random effects were normally distributed. In general, observe slightly larger bias under simulations with non-normally-distributed random effects compared to those with normally-distributed random effects (symmetric and skewed).

Shown in Figure 1 is the relationship between mean absolute bias and coverage probability based on the 1,000 κ_m estimates for each statistical package and for each simulation scenario. The results from PROC GLIMMIX are omitted because they were identical to those from ORDINAL. The dotted line across the horizontal axis represent bias at 0 and the dotted line across the vertical axis represent coverage probability at 95%. An ideal situation is when the estimate falls on the intersection between the two dotted lines. In general, the mean absolute bias was lowest under scenarios with normal random effects, slightly larger under scenarios with symmetric non-normal random effects, and largest under scenarios with skewed random effects. For scenarios with normally-distributed random effects, the coverage probabilities were consistently close to the anticipated 95% (90-95% for all packages). [Table 2] For scenarios with symmetric non-normal random effects, coverage probabilities were slightly higher than the anticipated 95% under Scenario 1 (97.1-98.1%) and Scenario 4 (98.7%-98.7%), while they were slightly lower than anticipated under Scenario 2 (82.6%-93.0%) and Scenario 3 (64.8%-87.5%). [Table 3] For scenarios with skewed random effects, the coverage probabilities were lower, especially under the extreme case scenarios, Scenario 1 and Scenario 3, where both the random effects distributions were highly skewed. More specifically, under Scenario 1 where the subject and rater random effects followed an exponential distribution and a Gamma distribution respectively, coverage probability ranged from 37.5% to 40.9% amongst all packages. Under Scenario 3 where the subject and rater random effects followed a Gamma distribution and a uniform distribution respectively, coverage probability ranged from 52.9% to 65.2% amongst all packages. [Figure 1; Supplementary Table 3]

Note the largest differences in mean absolute bias and coverage probability between the four packages under Scenario 3, when $\sigma_u^2 = 5$ and $\sigma_v^2 = 1$. For symmetric non-normal random effects, ORDINAL (same as PROC GLIMMIX) and LME4 yielded lower mean absolute biases (0.007 and 0.010, respectively) and higher coverage probabilities (84.1% and 87.5%, respectively) compared with MCMCglmm1 (mean absolute bias = 0.019, coverage probability = 64.8%) and MCMCglmm10 (mean absolute bias = 0.121, coverage probability = 74.0%). However, for skewed random effects, MCMCglmm1 and MCMCglmm10 yielded lower mean absolute biases (0.041 and 0.045, respectively) and higher coverage probabilities (65.2% and 59.9%, respectively) compared to ORDINAL/PROC GLIMMIX (mean absolute bias = 0.048, coverage probability = 54.5%) and LME4 (mean absolute bias = 0.049, coverage probability = 52.9%).

Table 2. Mean estimates and mean standard errors (SEs) from 1,000 simulations for the probit GLMM and agreement statistics computed from each statistical package with normally distributed random effects, I = 150 and J = 100

			Statistical Package				
			ORDINAL	**LME4**	**MCMCglmm1**	**MCMCglmm10**	**PROC GLIMMIX**
Scenario	**Parameter**	**Truth**	**Mean Estimate (Mean SE)**	**Mean Estimate (Mean SE)**	**Mean Estimate (Mean SE)**	**Mean Estimate (Mean SE)**	**Mean Estimate (Mean SE)**
1a	GLMM parameters:						
	η	-0.1	-0.103 (0.110)	-0.103 (0.110)	-0.103 (0.112)	-0.104 (0.119)	-0.103 (0.110)
	σ_u^2	1.5	1.492 (0.186)	1.500 (0.184)	1.530 (0.193)	1.613 (0.203)	1.493 (0.186)
	σ_v^2	0.2	0.198 (0.031)	0.199 (0.031)	0.213 (0.033)	0.316 (0.047)	0.198 (0.031)
	Agreement measures:						
	Model-based Kappa, κ_m	0.375	0.373 (0.022)	0.374 (0.033)	0.376 (0.022)	0.370 (0.023)	0.373 (0.022)
	Fleiss Kappa, κ_F				0.373 (0.001)		
	Coverage probability of κ_m (%)		93.2	93.3	92.4	94.2	93.2
	GLMM convergence rate (%)		99.9	100.0	100.0	100.0	100
2a	GLMM parameters:						
	η	1	1.012 (0.238)	1.010 (0.236)	1.015 (0.243)	1.020 (0.247)	1.012 (0.238)
	σ_u^2	1	0.999 (0.125)	1.006 (0.127)	1.014 (0.128)	1.085 (0.135)	0.999 (0.125)
	σ_v^2	5	4.791 (0.774)	4.771 (0.698)	5.061 (0.827)	5.214 (0.852)	4.791 (0.774)
	Agreement measures:						
	Model-based Kappa, κ_m	0.091	0.095 (0.013)	0.096 (0.013)	0.093 (0.013)	0.096 (0.014)	0.095 (0.013)
	Fleiss Kappa, κ_F				0.083 (0.001)		
	Coverage probability of κ_m (%)		94.1	94.6	93.1	94.7	94.1
	GLMM convergence rate (%)		100.0	99.0	100.0	1000.0	100

Table 2 (continued)

			Statistical Package				
			ORDINAL	LME4	MCMCglmm1	MCMCglmm10	PROC GLIMMIX
Scenario	Parameter	Truth	Mean Estimate (Mean SE)	Mean Estimate (Mean SE)	Mean Estimate (Mean SE)	Mean Estimate (Mean SE)	Mean Estimate (Mean SE)
3a	GLMM parameters:						
	η	1	0.995 (0.211)	0.992 (0.208)	0.999 (0.215)	1.002 (0.219)	0.995 (0.211)
	σ_u^2	5	4.849 (0.657)	4.815 (0.637)	5.122 (0.703)	5.230 (0.718)	4.849 (0.657)
	σ_v^2	1	0.998 (0.149)	1.005 (0.151)	1.023 (0.155)	1.124 (0.169)	0.998 (0.149)
	Agreement measures:						
	Model-based Kappa, κ_m	0.506	0.500 (0.025)	0.498 (0.025)	0.507 (0.025)	0.502 (0.026)	0.500 (0.025)
	Fleiss Kappa, κ_F				0.497 (0.001)		
	Coverage probability of κ_m (%)		91.9	92.1	90	90.8	91.9
	GLMM convergence rate (%)		100	96.3	100	100	100
4a	GLMM parameters:						
	η	1	0.999 (0.409)	1.003 (0.409)	0.999 (0.412)	0.999 (0.415)	0.999 (0.409)
	σ_u^2	10	10.013 (1.273)	10.101 (1.361)	10.191 (1.302)	10.275 (1.305)	10.013 (1.273)
	σ_v^2	10	9.913 (1.501)	10.009 (1.563)	10.151 (1.558)	10.258 (1.566)	9.912 (1.501)
	Agreement measures:						
	Model-based Kappa, κ_m	0.316	0.319 (0.031)	0.319 (0.031)	0.318 (0.031)	0.318 (0.031)	0.319 (0.031)
	Fleiss Kappa, κ_F				0.312 (0.001)		
	Coverage probability of κ_m (%)		94.6	94.7	93.6	93.9	94.6
	GLMM convergence rate (%)		100	93.5	100	100	100

Table 3. Mean estimates and mean standard errors from 1,000 simulations for the probit GLMM and agreement statistics computed from each statistical package with symmetric non-normally distributed random effects, *I* = 150 and *J* = 100

			Statistical Package				
			ORDINAL	**LME4**	**MCMCglmm1**	**MCMCglmm10**	**PROC GLIMMIX**
Scenario	**Parameter**	**Truth**	**Mean Estimate (Mean SE)**	**Mean Estimate (Mean SE)**	**Mean Estimate (Mean SE)**	**Mean Estimate (Mean SE)**	**Mean Estimate (Mean SE)**
1a	GLMM parameters:						
	η	-0.1	-0.103 (0.112)	-0.104 (0.112)	-1.104 (0.114)	-0.104 (0.120)	-0.103 (0.112)
	σ_u^2	1.5	1.554 (0.189)	1.564 (0.188)	1.588 (0.196)	1.669 (0.205)	1.554 (0.189)
	σ_v^2	0.2	0.200 (0.031)	0.201 (0.031)	0.214 (0.033)	0.317 (0.048)	0.200 (0.031)
	Agreement measures:						
	Model-based Kappa, κ_m	0.375	0.381 (0.022)	0.382 (0.022)	0.383 (0.022)	0.377 (0.023)	0.381 (0.022)
	Fleiss Kappa, κ_F				0.421 (0.001)		
	Coverage probability of κ_m (%)		97.5	97.3	97.1	98.1	97.5
	GLMM convergence rate (%)		100.0	100.0	100.0	100.0	100.0
2a	GLMM parameters:						
	η	1	1.104 (0.260)	1.095 (0.254)	1.112 (0.265)	1.119 (0.269)	1.104 (0.260)
	σ_u^2	1	0.999 (0.125)	1.006 (0.132)	1.015 (0.128)	1.086 (0.136)	0.999 (0.125)
	σ_v^2	5	5.655 (0.921)	5.609 (0.804)	6.073 (0.995)	6.222 (1.018)	5.655 (0.921)
	Agreement measures:						
	Model-based Kappa, κ_m	0.091	0.084 (0.012)	0.085 (0.012)	0.081 (0.012)	0.084 (0.012)	0.084 (0.012)
	Fleiss Kappa, κ_F				0.063 (0.001)		
	Coverage probability of κ_m (%)		91.3	93	82.6	91.1	91.3
	GLMM convergence rate (%)		100.0	92.3	100.0	100.0	100.0

Table 3 (continued)

			Statistical Package				
			ORDINAL	**LME4**	**MCMCglmm1**	**MCMCglmm10**	**PROC GLIMMIX**
Scenario	**Parameter**	**Truth**	**Mean Estimate (Mean SE)**	**Mean Estimate (Mean SE)**	**Mean Estimate (Mean SE)**	**Mean Estimate (Mean SE)**	**Mean Estimate (Mean SE)**
3a	GLMM parameters:						
	η	1	0.995 (0.211)	0.992 (0.208)	0.999 (0.215)	1.002 (0.219)	0.995 (0.211)
	σ_u^2	5	4.849 (0.657)	4.815 (0.637)	5.122 (0.703)	5.230 (0.718)	4.849 (0.657)
	σ_v^2	1	0.998 (0.149)	1.005 (0.151)	1.023 (0.155)	1.124 (0.169)	0.998 (0.149)
	Agreement measures:						
	Model-based Kappa, κ_m	0.506	0.500 (0.025)	0.498 (0.025)	0.507 (0.025)	0.502 (0.026)	0.500 (0.025)
	Fleiss Kappa, κ_F				0.497 (0.001)		
	Coverage probability of κ_m (%)		91.9	92.1	90	90.8	91.9
	GLMM convergence rate (%)		100.0	96.3	100.0	100.0	100.0
4a	GLMM parameters:						
	η	1	1.025 (0.394)	1.030 (0.395)	1.028 (0.397)	1.027 (0.400)	1.025 (0.394)
	σ_u^2	10	8.970 (1.141)	9.058 (1.242)	9.091 (1.158)	9.173 (1.164)	8.970 (1.141)
	σ_v^2	10	9.413 (1.434)	9.493 (1.488)	9.666 (1.487)	9.789 (1.505)	9.413 (1.434)
	Agreement measures:						
	Model-based Kappa, κ_m	0.316	0.307 (0.031)	0.307 (0.031)	0.305 (0.031)	0.305 (0.031)	0.307 (0.031)
	Fleiss Kappa, κ_F				0.324 (0.001)		
	Coverage probability of κ_m (%)		98.7	98.7	97.8	98.1	98.7
	GLMM convergence rate (%)		100.0	92.1	100.0	100.0	100.0

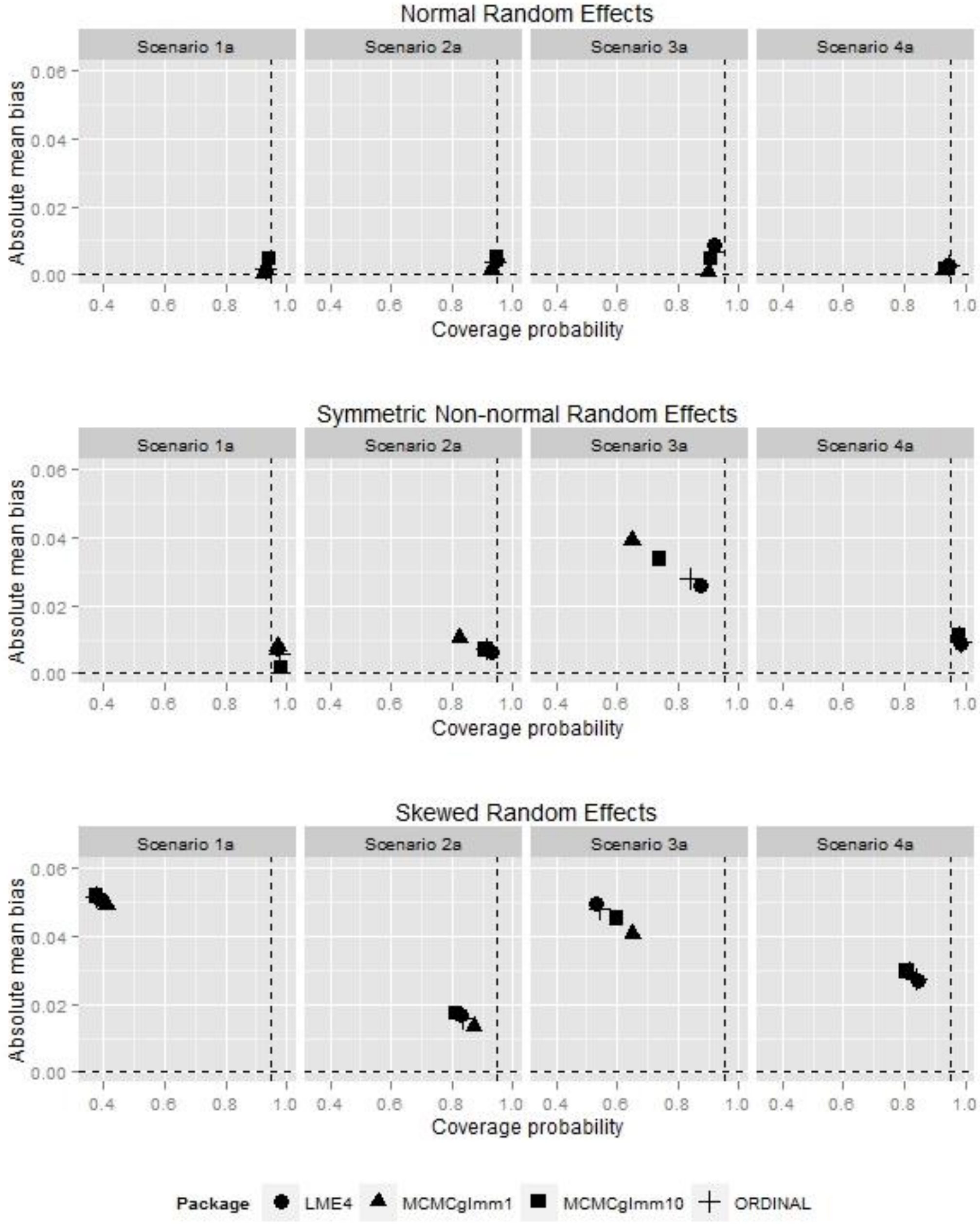

Figure 1. Absolute mean bias and coverage probability of estimated model-based kappa for each statistical package by scenario

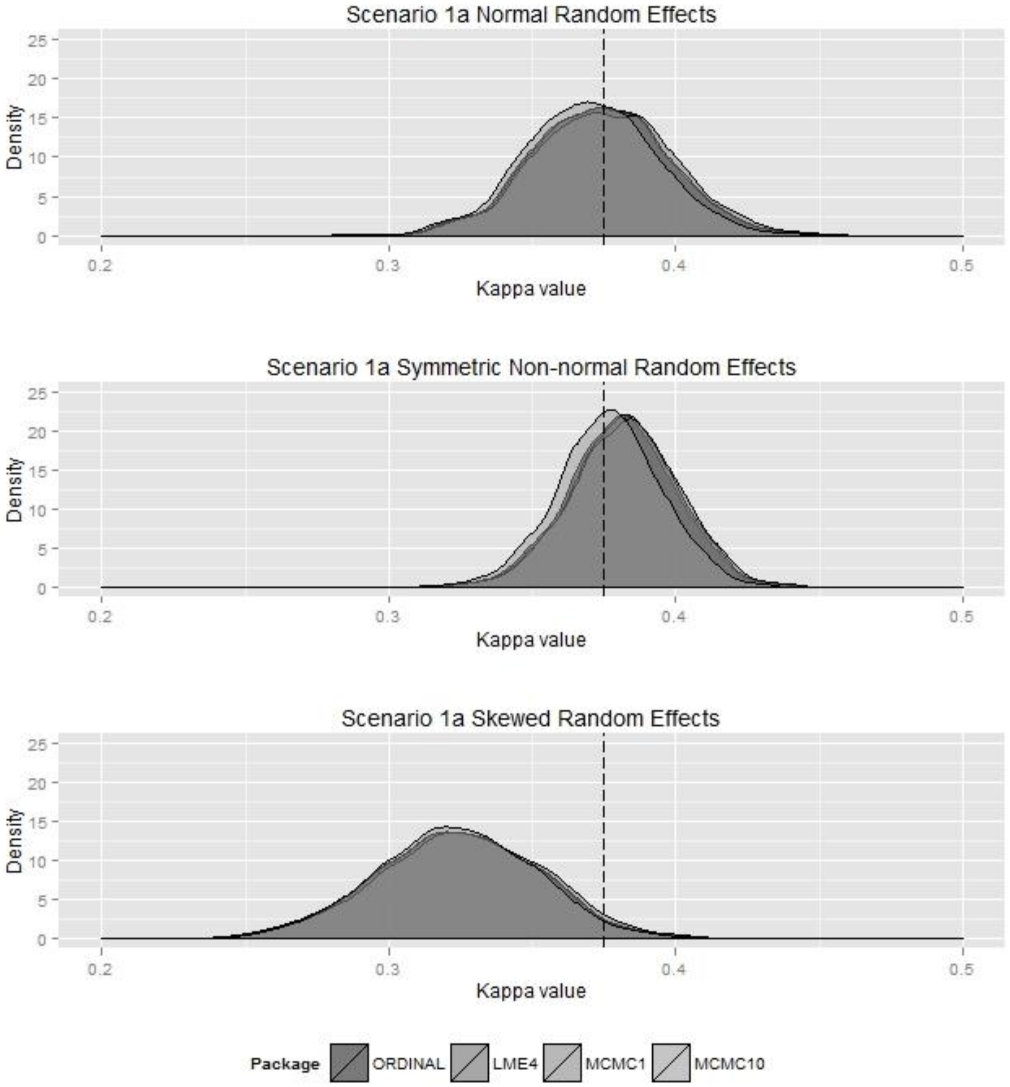

Figure 2. Density of model-based kappa measure of agreement estimates from each statistical package by varying sample size and random effects distribution for scenario 1

Interestingly, small to moderate biases in the GLMM parameter estimates had little noticeable impact on the estimates of the agreement measure κ_m. For example, under one of the scenarios with normally-distributed random effects (Scenario 1a), the estimates for σ_u^2 and σ_v^2 under the ORDINAL package were 1.492 and 0.198, respectively, while under MCMCglmm10, they were 1.613 and 0.316. Even with such seemingly different estimates, both packages produced similar κ_m estimates (0.373 under ORDINAL and 0.370 under MCMCglmm10).

Shown in Figure 2 are the density of κ_m estimates from the simulation scenario with $\sigma_u^2 = 1.5$ and $\sigma_v^2 = 0.2$ (Scenario 1; normal, symmetric non-normal, and skewed random effects distributions). Again, the results from PROC GLIMMIX are omitted because they were identical to those from ORDINAL. The densities of κ_m estimates obtained from all set of simulations were examined using plots, and found to be symmetric and reasonably bell-shaped, centered around the true value of κ_m for normal and symmetric non-normal random effects distributions. For skewed random effects distribution, the density of κ_m estimates appeared to be symmetric and bell-shaped but off-centered with a wider spread. Within each type of random effects distributions, the densities of κ_m estimates were extremely similar across the four packages. Similar densities of κ_m estimates were obtained from other simulation scenarios.

The empirical standard errors, computed as the standard deviation of the 1000 estimated κ_m, were comparable to the means of the model-based standard errors (Mean SE) presented in Tables 2 and 3. In general, when the random effects distribution was normal or skewed, the empirical standard errors were equal to or slightly larger than the model-based standard errors. On the other hand, when the random effects distribution was symmetric non-normal, the empirical standard errors were equal to or smaller than the model-based standard errors.

Fleiss' kappa estimates ($\hat{\kappa}_{\mathrm{F}}$) were comparable to model-based kappa estimates ($\hat{\kappa}_m$) in the majority of scenarios under normally distributed random effects. When the random effects distribution was symmetric non-normal, we observed slightly larger differences between $\hat{\kappa}_{\mathrm{F}}$ and $\hat{\kappa}_m$. For example, under symmetric non-normal Scenario 1a ($\sigma_u^2 = 1.5$ and $\sigma_v^2 = 0.2$), the mean of $\hat{\kappa}_{\mathrm{F}}$ was 0.421, while the means of $\hat{\kappa}_m$ ranged from 0.377 to 0.383 depending on the package. [Table 3] Under the scenarios with skewed random effects, the mean $\hat{\kappa}_{\mathrm{F}}$ and $\hat{\kappa}_m$ were also comparable except under Scenario 3 ($\sigma_u^2 = 5$ and $\sigma_v^2 = 1$) where the mean of $\hat{\kappa}_{\mathrm{F}}$ was 0.438 while the means of $\hat{\kappa}_m$ ranged from 0.459 to 0.466 depending on the package. The mean standard errors of $\hat{\kappa}_{\mathrm{F}}$ computed using equation (4) were extremely small, ranging from 0.001 to 0.003 depending on the sample size. However, the empirical standard errors for Fleiss' kappa ranged from

0.026 to 0.055, suggesting that the theoretical standard error potentially underestimates the variability of Fleiss' kappa statistic. This is a topic that needs to be further examined.

Applications to Large-Scale Cancer Studies

Mammogram Screening Study One of the two data sets used for illustration is from a previously-published study conducted by the BCSC, the Assessing and Improving Mammography (AIM) study, where radiologists evaluated whether a subject should be recalled or not based upon their screening mammogram results (Onega et al., 2012). In brief, the AIM study recruited 119 radiologists and obtained a set of 130 mammograms from 6 breast screening registries. The investigators developed 4 mammogram test sets, each containing 109 mammograms sampled from a set of 130 mammograms. Each test set varied by cancer prevalence and case difficulty, and included more cancer cases than a standard screening set; thus recall rates cannot be compared to a standard screening study. Participating radiologists were randomly assigned to one of the test sets and classified the mammograms in their test set. The primary outcome measured on each patient was a binary measure of whether the patient should be recalled for further testing versus no recall. See Onega et al. for further details on the AIM study design.

The aims are to assess the levels of agreement between the study radiologists using the two measures of agreement and to compare these results between the four available statistical packages. The data set was fit in all four packages.

Table 4. Estimates and standard errors for the probit GLMM and agreement statistics computed from each statistical package on the AIM data set

	Statistical Package				
	ORDINAL	**LME4**	**MCMCglmm1**	**MCMCglmm10**	**PROC GLIMMIX**
Parameters	**Estimate (SE)**	**Estimate (SE)**	**Estimate (SE)**	**Estimate (SE)**	**Estimate (SE)**
η	-0.124 (0.114)	-0.125 (0.114)	-0.121 (0.113)	-0.116 (0.125)	-0.124 (0.114)
σ_u^2	1.431 (0.192)	1.444 (0.189)	1.494 (0.205)	1.559 (0.218)	1.431 (0.192)
σ_v^2	0.195 (0.029)	0.195 (0.029)	0.207 (0.033)	0.295 (0.040)	0.195 (0.029)
κ_m (95% CI)	0.367	0.368	0.373	0.368	0.367
	(0.321-0.413)	(0.322-0.414)	(0.326-0.420)	(0.321-0.415)	(0.321-0.413)
κ_F (95% CI)			0.358		
			(0.356-0.361)		

Table 5. Estimates and standard errors for the probit GLMM and agreement statistics computed from each statistical package on bladder cancer data set

	Statistical Package				
	ORDINAL	**LME4**	**MCMCglmm1**	**MCMCglmm10**	**PROC GLIMMIX**
Parameters	**Estimate (SE)**	**Estimate (SE)**	**Estimate (SE)**	**Estimate (SE)**	**Estimate (SE)**
η	0.490 (0.460)	0.499 (0.461)	0.622 (0.502)	0.613 (0.763)	0.490 (0.460)
σ_u^2	3.137 (1.492)	3.156 (1.452)	6.114 (1.898)	5.853 (3.345)	3.137 (1.492)
σ_v^2	0.369 (0.274)	0.366 (0.275)	0.723 (0.575)	2.508 (1.587)	0.369 (0.274)
κ_m (95% CI)	0.490 (0.375-0.605)	0.492 (0.377-0.607)	0.570 (0.449-0.691)	0.430 (0.259-0.601)	0.490 (0.375-0.605)
κ_F (95% CI)			0.465 (0.391-0.539)		

Table 4 presents the estimated parameters with the standard errors from the GLMM model, the model-based kappa values with 95% CI, and the Fleiss kappa value ($\hat{\kappa}_F$) with 95% CI for this study. The version of Fleiss' kappa for unequal number of raters per subject was used because subjects' mammograms were classified by different number of raters. The model-based kappa $\hat{\kappa}_m$ produced slightly higher estimates compared to Fleiss' kappa in all four packages. For the model-based approaches, ORDINAL, LME4, MCMCglmm10, and PROC GLIMMIX produced extremely comparable results ($\hat{\kappa}_m$ = 0.367, 0.368, 0.368, and 0.367, respectively) indicating fair agreement between the radiologists. The kappa value obtained from MCMCglmm1 was slightly higher ($\hat{\kappa}_m$ = 0.373), but not enough to alter the inference and conclusion of the agreement. Fleiss' kappa ($\hat{\kappa}_F = 0.358$) was estimated slightly lower than the model-based kappa estimates $\hat{\kappa}_m$.

One of the simulation scenarios (Scenario 1) was designed to resemble the BCSC breast cancer data set. Under normally distributed random effects, the biases and coverage probabilities of $\hat{\kappa}_m$ were comparable between the packages. [Figure 1] Slightly more variability in bias was observed under non-normally-distributed random effects. Bias of $\hat{\kappa}_m$ obtained from MCMCglmm1 was the highest (0.008) while the bias obtained from MCMCglmm10 was the lowest (0.002). [Figure 1]

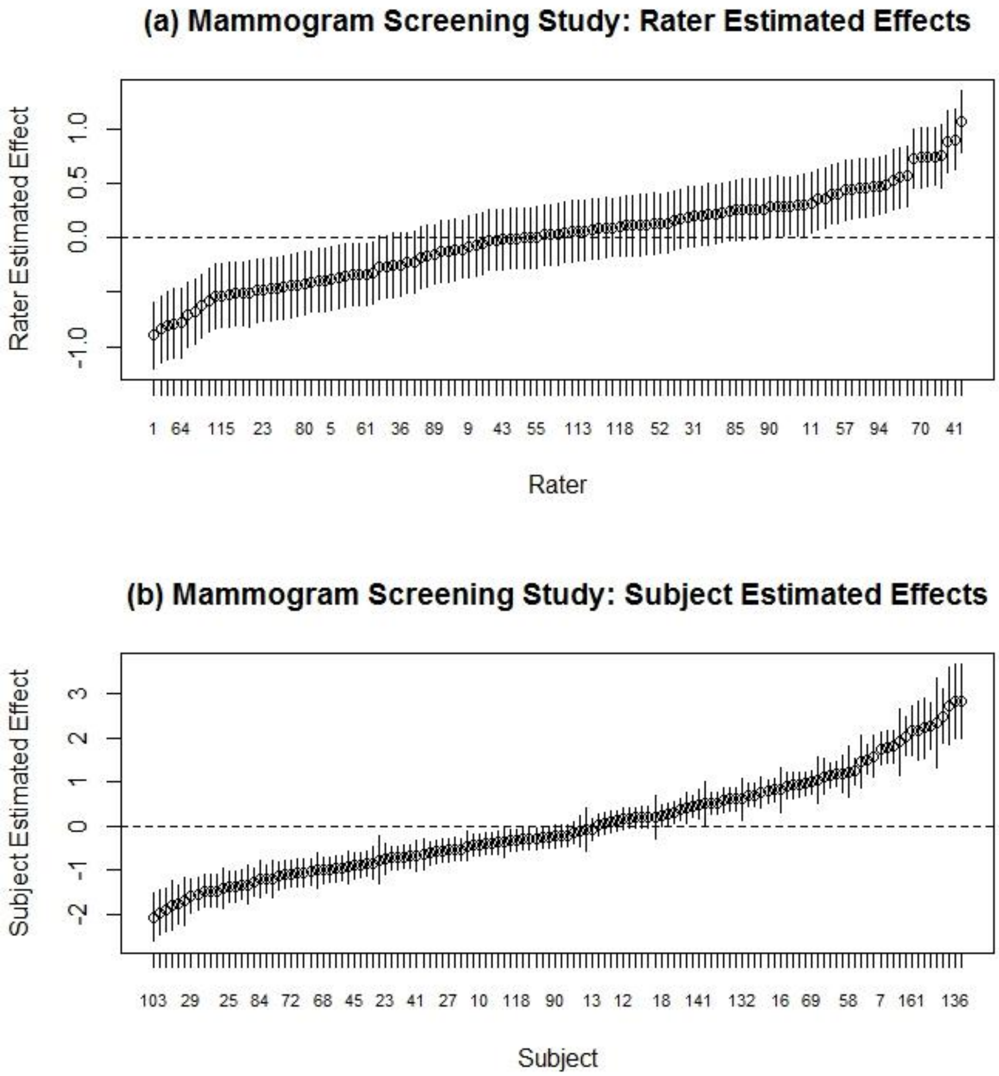

Figure 3. Rater- and subject-specific random effects from breast cancer data set

Bladder Cancer Study The second data set used for illustration is a study carried out by Compérat et al. (2013) which assessed agreement among eight genitourinary pathologists reviewing twenty-five bladder cancer specimens. Each pathologist provided a binary classification for each specimen according to whether or not they considered the sample to be non-invasive or invasive bladder cancer.

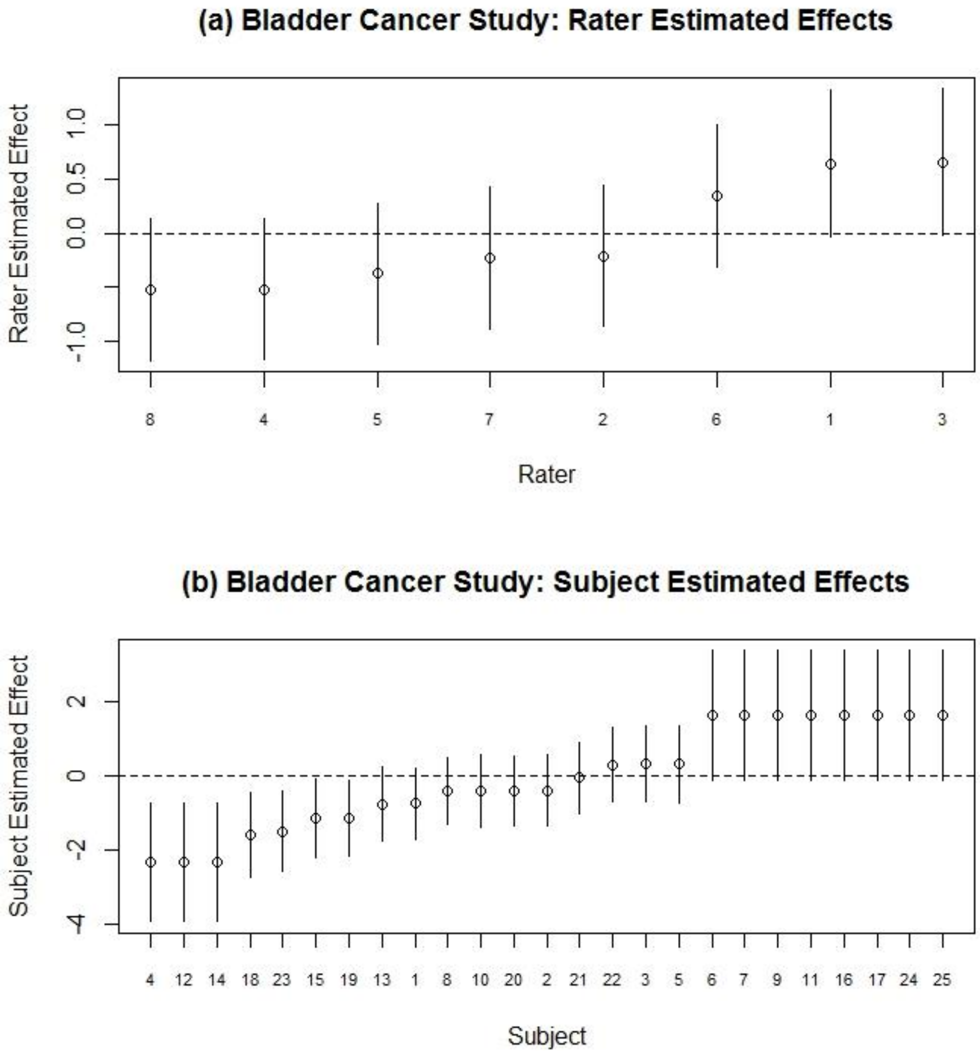

Figure 4. Rater- and subject-specific random effects from bladder cancer data set

This data set was fit using the four packages and calculated the two agreement measures ($\hat{\kappa}_m$, $\hat{\kappa}_\text{F}$). Model-based kappa estimates $\hat{\kappa}_m$ obtained from ORDINAL, LME4, and MCMCglmm packages with the smaller prior were higher compared to the Fleiss' kappa estimate ($\hat{\kappa}_\text{F} = 0.465$), which corroborate the original study value of moderate agreement between study pathologists. [Table 5] Results from the MCMCglmm package yielded an especially higher kappa estimate with the smaller prior ($\hat{\kappa}_m = 0.570$) and a lower kappa estimate with the larger prior ($\hat{\kappa}_m = 0.430$) relative to the estimates from the other packages. Compared to the previous AIM data set example, this data set provided a wider range of $\hat{\kappa}_m$ computed by the different packages, with the lowest and highest kappa values as 0.430 (MCMCglmm1) and 0.570 (MCMCglmm10), respectively. In a similar manner to our simulations, ORDINAL, LME4, and PROC GLIMMIX provided equivalent kappa estimates. However, all packages indicated that the pathologists had moderate agreement.

Unique Characteristics of Raters and Test Results Each statistical package can generate subject- and rater-specific random effects based on the conditional modes of the conditional distributions for the random effects. These solutions to the random effects are useful in understanding the behavior of individual raters if, for example, a rater is liberal or conservative in their classification of the test results. We present the solutions to the random effects from the ORDINAL package, and similar solutions were obtained from PROC GLIMMIX.

Presented in Figure 3 are the rater-specific random effects with 95% CI and the subject-specific random effects with 95% CI for the AIM study. Radiologists with large positive random effects values tended to recall mammograms more aggressively compared to other raters. However, radiologists with large negative random effects values were less likely to recall mammograms relative to other raters. For example, the radiologist with ID 22 who had the largest rater random effect ($\hat{v}_{22} = 1.07$) recalled 71% of the mammograms that he/she classified while the average recall rate among all radiologists was 43%. [Figure 3a] The subject-specific random effects ranged from -2.08 to 2.82. Large positive random effects values indicate mammograms with a high probability of recall while large negative values indicate mammograms with low probability of recall. Values that are close to 0 indicate mammograms with ambiguous results and suggest that the disease status on these mammograms was less well-defined than others. For example, subjects with IDs 136 and 147 had the largest random effects ($\hat{u}_{136} = \hat{u}_{147} = 2.82$) and they both had a recall rate of 100% while subject with ID 103 with the smallest random effect ($\hat{u}_{103} = -2.08$) had a recall rate of 2%. [Figure 3b]

Displayed in Figure 4 are the random effects conditional modes for the bladder cancer study. The rater-specific random effects were all moderate in value, ranging from -0.527 to 0.657. Relative to other pathologists, pathologists 1 and 3 were more likely to categorize the specimens as invasive (more liberal) while pathologists 8 and 4 were less likely to do so (more conservative). [Figure 4a] The subject-specific random effects ranged from -2.345 to 1.614. Subjects with large positive values of random effects (IDs 6-25) suggest having a more clear indication of invasive cancer compared to other subjects. On the other hand, subjects with large negative values of random effects (IDs 4-14) suggest that their samples indicate a non-invasive cancer. [Figure 4b] Note that many rater- and subject-specific random effects are equal to others due to the small number of raters and test results in this study.

Conclusion

The performance of four different packages in R and SAS was compared in the estimation of parameters for the binary GLMM and for two available measures of agreement between multiple raters. The GLMM parameter estimates were similar between the four packages when the random effects were normally distributed, especially between the packages that use a frequentist approach (ORDINAL, LME4, and PROC GLIMMIX). For one of the scenarios (Scenario 1a), the Bayesian package (MCMCglmm) was explored further by altering the belief parameter (v) to 0.002 which is used regularly in the prior specification of the random effects variance structure (Hadfield, 2015). Changing the specification of the priors had a minimal impact on the estimation of the random effects parameters and on the agreement statistic in the Bayesian package (MCMCglmm). When the random effects were non-normally distributed (both symmetric and skewed), we observed more variability in the GLMM parameter estimates between the four packages. However, we observed considerably smaller variability in the model-based agreement estimates even when the difference in the GLMM parameter estimates between the packages were relatively large.

It was shown in many studies misspecification of the random effects distributions do not seriously affect the estimation of the fixed effects. In computing the model-based kappa statistic from GLMM, however, the interest is in estimating the variances of the subject and rater random effects. Fewer studies have evaluated the impact of model misspecification on the random effects estimates and variance components. Through simulation, Agresti, Caffo, and Ohman-Strickland (2004) showed that extreme departure from Gaussian of the random effects may lead to loss of efficiency in the estimated variance of the random effects when fitting binary GLMM. If the true variance of the random effects is small, however, the problem of misspecification is negligible even if the true distribution is not Gaussian. In their simulation study, Litiere et al. (2008) assessed the impact of misspecified random effects distribution under binary GLMM on the maximum likelihood estimate of the random effects variance component. They observed that substantial bias can occur under misspecification even if the true variance of the random effects is small. On the other hand, McCulloch and Neuhaus (2011) showed that the estimation of random effects variance components is robust to misspecification of the random effects distribution. In our simulation study, we did observe slightly higher bias in the estimated variance of the random effects when the true random effects distribution were skewed compared to when the true random effects distribution was normal. This was more pronounced under the extreme scenarios where both the subject and rater random effects were non-normally distributed. Litiere et al. (2008) also noted that a more serious bias can be observed with more than one random effects in the model. However, the absolute bias in the model-based kappa estimates, which takes values between 0 and 1, was generally low (0.06 or less) even for these extreme scenarios across the four packages.

Typically used as an approach to measure reliability among multiple judges, the intra-class correlation coefficient (ICC) is another popular summary statistic for assessing agreement. Fleiss and Cuzick (1979) show that if the sample size is moderately large, ICC is "virtually identical" to kappa." (p. 539) Indeed, in our simulation study, we observed that Fleiss' kappa and ICC were identical to the second decimal place and hence only report the Fleiss' kappa as a comparison measure to the model-based agreement statistic.

In general, under normally distributed random effects, Fleiss' kappa estimates were smaller compared to the model-based kappa estimates, except in one scenario where Fleiss' kappa estimate was considerably larger than the model-based kappa estimates. Fleiss' kappa has several restrictions: First, it requires a constant number of ratings per subject. If the number of ratings per subject differs, then an alternate form of Fleiss' kappa is required to compute agreement. Second, Fleiss' kappa is prone to prevalence of success. If the success rate is low, Fleiss' kappa will underestimate the agreement between raters (Nelson & Edwards, 2008). Furthermore, although not discussed here, Fleiss' kappa cannot be extended to incorporate information about rater characteristics that may impact agreement. Lastly, in the simulation study, the standard errors of estimated Fleiss' kappa statistics computed using equation (4) were much smaller compared to the empirical standard errors. However, this issue needs to be further examined.

This study has some limitations. The assessment was restricted to four packages in R and SAS because of their popularity and accessibility. Other packages available in estimating GLMM with a crossed random effects structure such as MLwiN, WinBUGS, and Stata were not included.

This study has several strengths. First, the data generated for these simulation studies included realistic scenarios including the implementation of non-normally distributed random effects. In fact, the data set generated for one of the simulation scenarios was based on a real-life data set from the AIM study. Second, to our knowledge, this is the first study where the relatively new ORDINAL package was compared with existing packages on the performance of fitting GLMM with a crossed random effects structure for binary responses. The ORDINAL package is extremely stable, unlike the LME4 package, computationally efficient, and its parameter estimates were identical to those of PROC GLIMMIX in SAS. Lastly, the straightforward and reliable implementation of model-based measure of agreement ($\hat{\kappa}_m$) using existing packages was demonstrated. Model-based measure of agreement is robust to missing and unbalanced data, where not every subject's test result is rated by each rater.

Among frequentist R users, the ORDINAL package is recommended over the LME4 package for its stability and computational efficiency regardless of sample size and distribution of random effects. The GLIMMIX procedure in SAS produced nearly identical results to the ORDINAL package. For those who prefer Bayesian analysis, the MCMCglmm package performs well in fitting binary GLMM with a crossed random effects structure and for computing model-based agreement statistics. Although there was very little variability in the model-based agreement measures using different sets of priors, performing sensitivity analyses is recommended by altering the prior specification of the random effects distribution. A useful advantage of the Bayesian package implemented here (MCMCglmm) is its flexibility in incorporating a known characteristic of the data set to the model through the use of priors and its robustness to model misspecification when random effects distribution is skewed. Programs for fitting the binary GLMM with a crossed random effects structure for each of the four packages and an example data set are provided in supplementary materials. Full code for computing $\hat{\kappa}_m$ and its variance from GLMM parameter estimates for each package described in this paper is also included in the programs.

Overall, existing statistical software offer satisfactory packages or procedures for fitting binary GLMMs with a crossed random effects structure, and for estimation of agreement measures in large-scale agreement studies based upon multiple raters' binary classifications.

Acknowledgements

The authors are grateful for the support provided by grant R01-CA-17246301 from the United States National Institutes of Health. The AIM study was supported by the American Cancer Society, made possible by a generous donation from the Longaberger Company's Horizon of Hope®Campaign (SIRSG-07-271, SIRSG-07-272, SIRSG-07-273, SIRSG-07-274, SIRSG-07-275, SIRGS-06-281, SIRSG-09-270, SIRSG-09-271), the Breast Cancer Stamp Fund, and the National Cancer Institute Breast Cancer Surveillance Consortium (HHSN261201100031C). The cancer data used in the AIM study was supported in part by state public health departments and cancer registries in the U.S., see http://www.bcsc-research.org/work/acknowledgement.html. We also thank participating women, mammography facilities, radiologists, and BCSC investigators for their data. A list of the BCSC investigators is provided at: http://www.bcsc-research.org/.

Funding

This study was funded by the United States National Institutes of Health (grant number 1R01CA17246301-A1).

References

Agresti, A. (1989). A model for agreement between ratings on an ordinal scale. *Mathematical and Computer Modelling, 12*(9), 1188. doi: 10.1016/0895-7177(89)90272-0

Agresti, A., Caffo, B., & Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis, 47*(3), 639-653. doi: 10.1016/j.csda.2003.12.009

Beam, C. A., Conant, E. F., & Sickles, E. A. (2002). Factors affecting radiologist inconsistency in screening mammography. *Academic Radiology, 9*(5), 531-540. doi: 10.1016/S1076-6332(03)80330-6

Christensen, R. H. B. (2013). ordinal: Regression models for ordinal data (Version 2013.9-30) [R software package]. Retrieved from http://www.cran.r-project.org/package=ordinal

Ciatto, S., Houssami, N., Apruzzese, A., Bassetti, E., Brancato, B., Carozzi,… Scorsolini, A. (2005). Categorizing breast mammographic density:

Intra- and interobserver reproducibility of BI-RADS density categories. *The Breast*, *14*(4), 269-275. doi: 10.1016/j.breast.2004.12.004

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213-220. doi: 10.1037/h0026256

Compérat, E., Egevad, L., Lopez-Beltran, A., Camparo, P., Algaba, F., Amin, M.,… Van der Kwast, T. H. (2013). An interobserver reproducibility study on invasiveness of bladder cancer using virtual microscopy and heatmaps. *Histopathology*, *63*(6), 756-766. doi: 10.1111/his.12214

Elmore, J. G., Wells, C. K., Lee, C. H., Howard, D. H., & Feinstein, A. R. (1994). Variability in radiologists' interpretations of mammograms. *The New England Journal of Medicine*, *331*(22), 1493-1499. doi: 10.1056/NEJM199412013312206

Epstein, J. I., Allsbrook, W. C. J., Amin, M. B., Egevad, L. L., & ISUP Grading Committee. (2005). The 2005 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *The American Journal of Surgical Pathology*, *29*(9), 1228-1242. doi: 10.1097/01.pas.0000173646.99337.b1

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382. doi: 10.1037/h0031619

Fleiss, J. L., & Cuzick, J. (1979). The reliability of dichotomous judgments: Unequal numbers of judges per subject. *Applied Psychological Measurement*, *3*(4), 537-542. doi: 10.1177/014662167900300410

Fleiss, J. L., Nee, J. C. M., & Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, *86*(5), 974-977. doi 10.1037/0033-2909.86.5.974

Hadfield, J. D. (2010). MCMC methods for multi-response generalised linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, *33*(2), 1-22. doi: 10.18637/jss.v033.i02

Hadfield, J. D. (2015). *MCMCglmm course notes*. Retrieved from: https://cran.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf

Hsiao, C. K., Chen, P.-C., & Kao, W.-H. (2011). Bayesian random effects for interrater and test-retest reliability with nested clinical observations. *Journal of Clinical Epidemiology*, *64*(7), 808-814. doi: 10.1016/j.jclinepi.2010.10.015

Ibrahim, J., & Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *TEST, 18*(1), 1-43. doi: 10.1007/s11749-009-0138-x

Karim, M. R., & Zeger, S. L. (1992). Generalized linear models with random effects; Salamander mating revisited. *Biometrics, 48*(2), 631-644. doi: 10.2307/2532317

Kim, Y., Choi, Y.-K., & Emery, S. (2013). Logistic regression with multiple random effects: A simulation study of estimation methods and statistical packages. *The American Statistician, 67*(3), 37-41. doi 10.1080/00031305.2013.817357

Kuk, A. Y. C., & Cheng, Y. W. (1997). The Monte Carlo Newton-Raphson algorithm. *Journal of Statistical Computation and Simulation, 59*(3), 233-250. doi: 10.1080/00949657708811858

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. doi: 10.2307/2529310

Li, B., Lingsma, H. F., Steyerberg, E. W., & Lesaffre, E. (2011). Logistic random effects regression models: A comparison of statistical packages for binary and ordinal outcomes. *BMC Medical Research Methodology, 11*(77). doi: 10.1186/1471-2288-11-77

Litière, S., Alonso, A., & Molenberghs, G. (2008). The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Statistics in Medicine, 27*(16), 3125-3144. doi 10.1002/sim.3157

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statististical Association, 92*(437), 162-170. doi: 10.2307/2291460

McCulloch, C. E., & Neuhaus, J. M. (2011). Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics, 67*(1), 270-279. doi: 10.1111/j.1541-0420.2010.01435.x

Nelson, K. P., & Edwards, D. (2008). On population-based measures of agreement for binary classifications. *Canadian Journal of Statistics, 36*(3), 411-426. doi: 10.1002/cjs.5550360306

Nelson, K. P., & Edwards, D. (2010). Improving the reliability of diagnostic tests in population-based agreement studies. *Statistics in Medicine, 29*(6), 617-626. doi: 10.1002/sim.3819

Onega, T., Smith, M., Miglioretti, D. L., Carney, P. A., Geller, B. A., Kerlikowske, K.,… Yankaskas, B. (2012). Radiologist agreement for mammographic recall by case difficulty and finding type. *Journal of the American College of Radiology, 9*(11), 788-794. doi: 10.1016/j.jacr.2012.05.020

Ooms, E. A., Zonderland, H. M., Eijkemans, M. J. C., Kriege, M., Mahdavian Delavary, B., Burger, C. W., & Ansink, A. C. (2007). Mammography: Interobserver variability in breast density assessment. *The Breast*, *16*(6), 568-576. doi: 10.1016/j.breast.2007.04.007

R Core Team (2014). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical. Retrieved from: http://www.R-project.org/

Scott, W. A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly, 19*(3), 321-325. doi: 10.1086/266577

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428. doi: 10.1037/0033-2909.86.2.420

Tanner, M. A., & Young, M. A. (1985). Modeling agreement among raters. *Journal of the American Statistical Association, 80*(389), 175-180. doi: 10.2307/2288068

Zhang, H., Lu, N., Feng, C., Thurston, S. W., Xia, Y., Zhu, L., & Tu, X. M. (2011). On fitting generalized linear mixed-effects models for binary responses using different statistical packages. *Statistics in Medicine, 30*(20), 2562-2572. doi: 10.1002/sim.4265