Published in final edited form as:
*Stat Med*. 2010 March 15; 29(6): 617–626. doi:10.1002/sim.3819.

# Improving the reliability of diagnostic tests in population-based agreement studies

**Kerrie P. Nelson**[a,*,†] and **Don Edwards**[b]

[a]Massachusetts General Hospital and Harvard Medical School, Biostatistics Center, 50 Staniford Street, Suite 560, Boston, MA 02114, U.S.A

[b]Department of Statistics, University of South Carolina, SC 29208, U.S.A

## Abstract

Many large-scale studies have recently been carried out to assess the reliability of diagnostic procedures, such as mammography for the detection of breast cancer. The large numbers of raters and subjects involved raise new challenges in how to measure agreement in these types of studies. An important motivator of these studies is the identification of factors that contribute to the often wide discrepancies observed between raters' classifications, such as a rater's experience, in order to improve the reliability of the diagnostic process of interest. Incorporating covariate information into the agreement model is a key component in addressing these questions. Few agreement models are currently available that jointly model larger numbers of raters and subjects and incorporate covariate information. In this paper, we extend a recently developed population-based model and measure of agreement for binary ratings to incorporate covariate information using the class of generalized linear mixed models with a probit link function. Important information on factors related to the subjects and raters can be included as fixed and/or random effects in the model. We demonstrate how agreement can be assessed between subgroups of the raters and/or subjects, for example, comparing agreement between experienced and less experienced raters. Simulation studies are carried out to test the performance of the proposed models and measures of agreement. Application to a large-scale breast cancer study is presented.

### Keywords

agreement; model-based kappa; Cohen's kappa; intraclass kappa; generalized linear mixed model; crossed random effects

## 1. Introduction

The reliability of common medical procedures for the diagnosis of cancer and other diseases has become an important issue over the last few decades [1–4]. Subjective classifications of tests including mammograms, x-rays and biopsies are routinely carried out by physicians and other biomedical professionals, yet wide discrepancies have been observed between

[*]Correspondence to: Kerrie P. Nelson, Massachusetts General Hospital and Harvard Medical School, Biostatistics Center, 50 Staniford Street, Suite 560, Boston, MA 02114, U.S.A.
[†]kpnelson@partners.org

experts [1, 4–6]. Many studies have recently been carried out to measure the levels of agreement between experts in such settings, and to attempt to identify factors that may contribute to the observed discrepancies. Identification of influential factors provides valuable insight into how the reliability of diagnostic procedures might be improved [7–9]. These studies include, among others, a nationwide study carried out by Beam *et al.* [9] involving the classification of 148 mammograms by over 100 experts. Their study concluded that individual radiologists' current reading volume was not statistically associated with the accuracy in reading screening mammograms, but several other factors were associated. Miglioretti *et al.* [7] examined the variability between raters in their classification of over 35 000 mammograms. They found considerable variation in the interpretative performance of diagnostic mammography that was not explained by the characteristics of the patients whose mammograms were interpreted. Thirty-eight prostate cancer biopsies were classified by 41 pathologists in a Gleason grading study [10], where barely moderate agreement between raters was found. The assessment of agreement in these studies is challenging due to the large numbers of experts and subjects under study. Further, the modeling of additional information that may lead to the identification of influential factors raises a further challenge.

Currently available methods for modeling agreement and reliability among raters can be broadly categorized as summary measures and model-based approaches. Summary statistics include the very popular Cohen's kappa [3] and its extensions, which are often applied in biomedical studies to describe the overall level of chance-corrected agreement between two or more raters. The inclusion of covariate information when using Cohen's kappa has also been described [11–13]. However, it is well acknowledged that Cohen's kappa has a number of weaknesses in its usage [14, 15] and can lead to an inaccurate assessment of agreement, sometimes severely so. Cohen's kappa is applicable when several subjects are classified by two raters. However, when multiple raters are classifying several subjects, the multirater intraclass kappa described in Kraemer *et al.* [16] is the appropriate summary measure of reliability to use. In general, summary statistics provide a useful overall measure of agreement, but sustain a large loss of information due to outputting a single value about the agreement within a process.

Model-based approaches provide a more complete and broader framework for assessing factors influencing agreement, and thus perhaps improving reliability. These include log-linear models [17–21], latent class and trait models [20–24], and logistic regression models [25, 26]. Several of these approaches include the raters as fixed effects, and provide inference to the specific raters and items under study, not extending to agreement involving entire populations of such raters. Such methods work well when a small number of raters is involved, but become increasingly complex, frequently involving a large number of parameters when the classifications of more than two or three raters are included. Log-linear models and latent models can incorporate covariate information; however, few methods currently available are able to both incorporate covariate information and yield inference regarding the underlying diagnostic procedure.

For common medical procedures, it is desirable to make inference regarding the levels of agreement in the underlying populations of raters and subjects who are typically involved in the diagnostic procedure of interest, and to examine the influence of factors on the agreement process in such a setting. A number of approaches have been developed with this purpose in mind, including a population definition for the intraclass kappa [27], a flexible latent variable model proposed by Williamson and Manatunga [23] for ordinal classifications, and more recently, a generalized linear mixed model for modeling binary classifications [28], which is designed to examine agreement in a population-based setting. While Williamson and Manatunga's approach can potentially include any number of raters, a fixed effect term is included for each individual rater in the model and is thus best-suited to a small to a moderate number of raters.

Clinical factors are often associated with the prevalence of a disease and may influence the reliability of diagnostic procedures. In the breast cancer setting, a mammogram presenting an advanced stage breast cancer is easier to distinguish than many less advanced forms of breast cancer, leading to a better agreement between the raters. Prior knowledge regarding a subject's clinical history could also influence a rater's perception of the mammogram and consequently their classification. For example, in the classification of mammograms, the age of the woman is an important factor as the prevalence of breast cancer increases with a woman's age [29]. Other factors such as a rater's level of experience and type of training could affect the levels of agreement present.

The earlier paper by Nelson and Edwards [28] proposed a simple population-based agreement model and summary statistic that focuses on inference regarding agreement in a diagnostic procedure for large numbers of raters and subjects (assumed randomly sampled from their populations), where each rater in the sample of raters is assumed to rate each subject independently of the other raters. This paper focused on defining the general concepts of measuring reliability, and also how the class of generalized linear mixed models could be used as a suitable framework in such a setting. Attention was focused on the more theoretical aspects underlying the simplest model proposed through the use of probit and logit link functions.

In this paper we extend the simple population-based agreement model and summary measure of agreement proposed by Nelson and Edwards [28] to the measurement of agreement between subgroups of raters and/or subjects, for example, how agreement between experienced raters compares with that of less experienced raters, by the incorporation of important covariate information into the model. Such models can lead to the identification of factors that influence agreement and the reliability of the diagnostic procedure, where consistency between raters is an important prerequisite for a reliable procedure. In agreement models, the inclusion of covariate information requires careful consideration due to its influence on both the prevalence of the disease under study and on the agreement process, and yields valuable information regarding factors that influence both how a rater classifies a subject and the agreement between the raters.

The remainder of the paper is as follows: Section 2 introduces the basic population-based agreement model described in Nelson and Edwards [28] and presents the extensions to the model. In Section 3, the associated population-based summary measure of agreement and an extended version are described. Simulation studies are carried out in Section 4 to examine the effects of including important factors into the proposed model and the agreement summary statistic described in Sections 2 and 3. Application to a breast cancer study involving the classifications of a large set of mammograms is presented in Section 5. In Section 6 concluding remarks and discussion are made.

## 2. Models and measures of agreement

A natural choice for modeling agreement data when the underlying populations of raters and subjects and the diagnostic procedure are of interest is the class of generalized linear mixed models with a crossed random effect structure. We restrict our attention here to classifications made on a binary scale, for example, $y_{ij} = 1$ if the $i$th subject is rated as positive (for example, diseased) by the $j$th rater, and $y_{ij} = 0$ otherwise. It is assumed that the raters and the subjects are randomly selected from their respective populations, and that each of the $J$ raters classify all of the subjects under study (independently from the other $J - 1$ raters). Although binary outcomes can be modeled using a generalized linear mixed model with either a probit or logit link function, among others, we will focus here on the probit link function for ease of mathematics. Nearly identical results are achieved for the logistic link function [28]. The basic form of the agreement model is

$$\Phi^{-1}(p_{ij})=\eta+u_i+v_j, \quad (1)$$

for subject $i = 1, \ldots, I$ and rater $j = 1, \ldots, J$. The quantity $p_{ij} = \text{pr}(y_{ij} = 1)$ is the probability of the $i$th item being classified as positive by the $j$th rater, and the constant η refers to the intercept term of the model and can also be regarded as a measure of the prevalence of positives in the data. When η is large, the overall frequency of positives in the data is high. The terms $u_i$ and $v_j$ represent random effects for the $i$th subject and $j$th rater respectively, with assumed independent Normal $(0, \sigma_u^2)$ and Normal $(0, \sigma_v^2)$ distributions. A subject with a positive (negative) random effect is more (less) likely than other such subjects to be classified as positive over many raters. A large value of $\sigma_u^2$ is indicative of subjects that are easy to distinguish from one another. A rater with a positive (negative) random effect is more liberal (cautious) in classifying a subject as positive over many subjects. A large value of $\sigma_v^2$ suggests more variability between the raters within the population in how they classify the subjects.

### 2.1. Inclusion of covariates

Many agreement models, including log-linear and logistic regression models, utilize an agreement index as the response, where 0 = two raters disagree, 1 = two raters agree, in their classification of a single item [17, 25, 26, 30]. In the generalized linear mixed model considered here, and in Williamson and Manatunga’s agreement model [23], the response variable $y_{ij}$, instead, is the binary classification made on the $i$th subject by the $j$th rater. Then when a covariate is included into the model, careful consideration has to be given as to whether it is likely to directly impact the prevalence so that it is included as a fixed effect, or more likely to influence the agreement between the raters and is then included as a random effect term. Covariates that influence both the prevalence and the agreement can be included as both fixed and random effects.

Earlier studies [5, 7, 9, 10] demonstrated that certain factors including a radiologist's average volume of mammogram interpretations, a woman's age and time since previous mammogram, a physician's length of practice and type of training may influence how a rater classifies an item, and may affect the agreement between the raters. For example, raters who carry out more classifications on a routine basis may develop more experience and accuracy in reading mammograms, leading to a higher rate of agreement between the experienced raters. Factors related to subject (for example, stage of cancer) would lead to using a test in some clinical populations and not in others. Factors related to raters would change the form of the test in the same population. The basic agreement model in equation (1) is extended to

$$\Phi^{-1}(p_{ij})=\eta+\boldsymbol{\beta}_{\mathbf{1}}^{'}\boldsymbol{x}_i+\boldsymbol{\beta}_{\mathbf{2}}^{'}\boldsymbol{x}_j+\boldsymbol{z}_1^{'}\boldsymbol{u}_i+\boldsymbol{z}_2^{'}\boldsymbol{v}_j, \quad i=1,\ldots,I, j=1,\ldots,J, \qquad (2)$$

where $\boldsymbol{x}_i$ ($r$ × 1) is the vector of factors associated with the $i$th subject, and $\boldsymbol{x}_j$ ($s$ × 1) the vector of fixed effects associated with the $j$th rater. The associated vectors of parameters are $\boldsymbol{\beta}_{\mathbf{1}}$ ($r$ × 1) and $\boldsymbol{\beta}_{\mathbf{2}}$ ($s$ × 1), respectively. The vectors $\boldsymbol{z}_1$ ($p$ × 1) and $\boldsymbol{z}_2$ ($q$ × 1) represent the design vectors of the random effects for the random effect vectors $\boldsymbol{u}_i$ ($p$ × 1) and $\boldsymbol{v}_j$ ($q$ × 1), respectively, where $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ contain the associated random effects for the subjects and the raters. The random effects are assumed to follow multivariate normal distributions

$$\boldsymbol{u}\sim\mathrm{MVN}(\mathbf{0},\boldsymbol{\Sigma}_{\boldsymbol{u}}) \quad \text{and} \quad \boldsymbol{v}\sim\mathrm{MVN}(\mathbf{0},\boldsymbol{\Sigma}_{\boldsymbol{v}}),$$

where the covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ are of dimensions $p$ × $p$ and $q$ × $q$, respectively. More complex random effect structures can be employed if required and if sufficient data are available. The random effect vectors $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ can be predicted to describe the unique effects of each rater and subject included in the study.

## 3. A population-based measure of agreement

Cohen's kappa is a very popular summary measure of agreement due to its appealing simplicity of both calculation and interpretation. In this section, we describe and extend a simple population-based summary measure of agreement introduced in Nelson and Edwards [28] to include covariate information associated with the raters and subjects to try and improve the reliability of a diagnostic test, and to determine their influence on the prevalence and agreement. It is demonstrated how the extended summary statistic can be used to compare the agreement between the subgroups of raters and/or subjects, for example, the amount of agreement between the experienced and the inexperienced raters. The summary statistic is comparable to Cohen's kappa in style and interpretation, avoiding many of the weaknesses observed in Cohen's kappa use, while allowing for population-based inference. An advantage of the model-based kappa statistic is that all available data is utilized to estimate the parameters, even when the agreement between the sub-groups are of interest. A general overall measure of agreement based upon all the raters and subjects included in a study can be obtained by fitting the simplest form of the generalized linear mixed model presented in equation (1).

The simplest form of the model-based kappa measure of agreement as introduced in Nelson and Edwards [28] is

$$\kappa_{\mathrm{m}}=1-4\int_{-\infty}^{\infty}\Phi\left(\frac{z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\left[1-\Phi\left(\frac{z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\right]\phi(z)\mathrm{d}z, \qquad (3)$$

where $\rho=\sigma_u^2/(\sigma_u^2+\sigma_v^2+1)$ and $0\leq\kappa_{\mathrm{m}}\leq1$; the maximum likelihood estimates of the variances $\sigma_u^2$ and $\sigma_v^2$ are obtained from the corresponding generalized linear mixed model in equation (1). We can interpret the numerical value of our summary statistic in a similar manner to Cohen’s kappa, where a value close to 1 is suggestive of strong agreement, such that as given in Landis and Koch [31]. The approximate asymptotic variance of $\hat{\kappa}_M$ is derived using the multivariate delta theorem, and is estimated as

$$\widehat{\mathrm{var}}(\hat{\kappa}_{\mathrm{m}})\approx16\left[\left\{\int_{-\infty}^{\infty}\left(\frac{1}{2\hat{\rho}(1-\hat{\rho})}\left(\frac{z\sqrt{\hat{\rho}}}{\sqrt{1-\hat{\rho}}}\right)\phi\left(\frac{z\sqrt{\hat{\rho}}}{\sqrt{1-\hat{\rho}}}\right)\left[1-2\Phi\left(\frac{z\sqrt{\hat{\rho}}}{\sqrt{1-\hat{\rho}}}\right)\right]\right)\phi(z)\mathrm{d}z\right\}^2\right.$$
$$\left.\times\left[\left(\frac{\hat{\sigma}_v^2-\hat{\sigma}_w^2}{(\hat{\sigma}_T^2)^2}\right)\left(\frac{2\hat{\sigma}_u^4}{I}\right)+\left(\frac{\hat{\sigma}_u^2}{(\hat{\sigma}_T^2)^2}\right)^2\left(\frac{2\hat{\sigma}_v^4}{J}\right)\right]\right].$$

### 3.1. Inclusion of covariates

In equation (4), we present a measure of agreement (extended from the original $\kappa_{\mathrm{m}}$ in equation (3)) between two raters $j$ and $j'$ each classifying the $i$th subject, which accounts for the fixed effects vector $\boldsymbol{x}_j$, where the two raters may have different values for at least one of the covariates of interest contained in $\boldsymbol{x}_j$. For example, $\kappa_{\mathrm{m}}$ in equation (4) provides a summary measure of agreement between two randomly chosen raters, who have different amounts of experience in reading mammograms.

$$\kappa_{\mathrm{m}}=1-4\int_{-\infty}^{\infty}\Phi\left(\frac{z\sqrt{\rho}+\frac{\beta_1'x_i+\beta_2'x_j}{\sqrt{\sigma_T^2}}}{(1-\rho)^{1/2}}\right)\left[1-\Phi\left(\frac{z\sqrt{\rho}+\frac{\beta_1'x_i+\beta_2'x_{j'}}{\sqrt{\sigma_{T'}^2}}}{(1-\rho')^{1/2}}\right)\right]\phi(z)\mathrm{d}z \qquad (4)$$

where

$$\rho=\sigma_1^2/(\sigma_1^2+\sigma_2^2+1)\ \text{ and }\ \rho'=\sigma_1^2/(\sigma_1^2+\sigma_{2'}^2+1). \qquad (5)$$

The terms $\sigma_1^2$ and $\sigma_2^2$ represent the variances of the sums of the random effect components $\boldsymbol{z}_1'\boldsymbol{u}_i$ and $\boldsymbol{z}_2'\boldsymbol{v}_j$ for the *i*th subject and *j*th rater, respectively, that is, the variability associated with the rater and the subject random effects such that $\sigma_1^2{=}\mathrm{var}(z_1'u_i){=}z_1'\Sigma_u z_1$ and $\sigma_2^2{=}\mathrm{var}(z_2'v_j){=}z_2'\Sigma_v z_2$. Similarly, $\sigma_{2'}^2{=}\mathrm{var}(z_2'v_j')$ for the *j*′th rater. The quantity $\sigma_T^2{=}\sigma_1^2{+}\sigma_2^2{+}1$ is a measure of the total variability present in the model ($T$ for total), given the covariate values of the *i*th subject and the *j*th rater. Similarly, $\sigma_T^2$ is the total variability in the model given the covariate values of the *i*th subject and the *j*′th rater. In equation (5) an extended version of ρ is presented. It is defined as a measure of subject distinguishability relative to the variability between two raters with the same covariate information, given the *i*th subject's covariate information. Similarly, the term ρ′ specifies the value of ρ for another rater *j*′ who may have a different set of covariate values from rater *j*, thus leading to a different value of $\sigma_2^2$, and consequently ρ. The formula for ρ′ is presented in equation (5).

With the inclusion of covariates, the exact form of the variance of $\kappa_\mathrm{m}$ is dependent upon the random effect design vectors $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ and the assumed correlation structures of the random effects. The model-based kappa statistic $\kappa_\mathrm{m}$ is a function of the variance components contained in $\boldsymbol{\Sigma_u}$ and $\boldsymbol{\Sigma_v}$, and the regression coefficient vectors $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$, where vector $\boldsymbol{\theta}$ contains all of the individual components. The asymptotic variance of $\kappa_\mathrm{m}$ can then be obtained using the multivariate delta theorem, and estimated as

$$\widehat{\mathrm{var}}(\hat{\kappa}_\mathrm{m}){=}16\mathrm{var}\left(\int_{-\infty}^{\infty}\Phi\left(\frac{z\sqrt{\hat{\rho}}{+}\frac{\hat{\beta}_1' x_i{+}\hat{\beta}_2 x_i}{\sqrt{\hat{\sigma}_T^2}}}{(1-\hat{\rho})^{1/2}}\right)\left[1-\Phi\left(\frac{z\sqrt{\hat{\rho}}{+}\frac{\hat{\beta}_1' x_i{+}\hat{\beta}_2 x_{j'}}{\sqrt{\hat{\sigma}_{T'}^2}}}{(1-\hat{\rho}')^{1/2}}\right)\right]\phi(z)\mathrm{d}z\right) \quad (6)$$

$$={\frac{16}{IJ}}(\boldsymbol{h}\boldsymbol{\Sigma}^{-1}\boldsymbol{h}') \quad \text{where } \boldsymbol{h}{=}\left(\frac{\delta\kappa_\mathrm{m}}{\delta\theta_1},\ldots,\frac{\delta\kappa_\mathrm{m}}{\delta\theta_l}\right) \quad (7)$$

is the vector of derivatives of $\kappa_\mathrm{m}$ with respect to the variance components contained in the $\boldsymbol{\Sigma_u}$ and $\boldsymbol{\Sigma_v}$ matrices and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, and $\boldsymbol{\Sigma}$ ($l$ × $l$) is the variance–covariance matrix of all the variance components and $\boldsymbol{\beta}$ contained in $\boldsymbol{\theta}$. An application of this equation is presented in the Appendix.

## 4. Simulation studies

Simulation studies were carried out to examine the effects of including factors of interest on the estimation in the models and measures of agreement described. The simulations were based upon three probit generalized linear mixed models of increasing complexity as follows:

$$\text{Model (a)}:\Phi^{-1}(p_{ij})=\eta+u_i+v_j,$$

$$\text{Model (b)}:\Phi^{-1}(p_{ij})=\eta+\beta x_j+u_i+v_j,$$

$$\text{Model (c)}:\Phi^{-1}(p_{ij})=\eta+\beta x_j+u_i+v_j+z_2 v_{1j},$$

where $i = 1, \ldots, I$, $j = 1, \ldots, J$ with $I$ and $J$ each set at 50. The term $\eta$ is the intercept and the response variable $y_{ij}$ represents the classification made by the $j$th rater on the $i$th subject, equaling 0 for a subject classified as not diseased and 1 otherwise. The random effects $u_i$ and $v_j$ in models (a) and (b) are assumed to be normally distributed as $\mathrm{N}(0, \sigma_u^2)$ and $\mathrm{N}(0, \sigma_v^2)$, respectively. The additional random effect $v_{1j}$ in model (c) is associated with a factor $z_2$ likely to influence the agreement between the raters, such as a rater's level of experience ($z_2$ = 1 for an experienced rater, and 0 otherwise). For simulation purposes, $z_{1j}$ is randomly generated from a Bin($n$ = 1, $p$ = 0.5) distribution, $j = 1, \ldots, J$, and the random effect term $v_{1j}$ is also assumed to be normally distributed with variance $\sigma_{v1}^2$ and correlated with the other rater random effect $v_j$. The covariance matrix for model (c) thus takes the form

$$\begin{pmatrix} u_i \\ v_j \\ v_{1j} \end{pmatrix} \sim \mathrm{MVN}\left(\mathbf{0}, \begin{bmatrix} \sigma_u^2 & 0 & 0 \\ 0 & \sigma_v^2 & \rho_v\sigma_v\sigma_{v_1} \\ 0 & \rho_v\sigma_v\sigma_{v_1} & \sigma_{v_1}^2 \end{bmatrix}\right). \quad (8)$$

In models (b) and (c), $x_j$ represents a fixed covariate value for the $j$th rater, (for example, the experience level of the $j$th rater (1 = high, 0 = low). In the simulations, $x_j$ is randomly generated from a Bin(1,0.5) distribution. Two different values for the intercept, $\eta$ = 1 and 3, were included in the simulations. The regression coefficient $\beta$ was set at 0.5. Different values of the variance components were also included to assess the effects of increasing variability: for all three models, $\sigma_u^2, \sigma_v^2$ and $\sigma_{v_1}^2$ were set at (1,1,1) and (5,5,5), and for model (c), the correlation $\rho_V$ between random effects $v_j$ and $v_{1j}$ was set at 0.25. Due to the computational intensity involved, the number of simulations was restricted to the fitting of 100 randomly generated data sets for each simulation scenario.

Each data set was fitted using a Monte-Carlo expectation–maximization algorithm (MCEM) developed by McCulloch [32] (see also Kuk and Cheng [33]) to obtain almost-exact maximum likelihood estimates of the parameters in the generalized linear mixed model. A description of this algorithm is provided in Nelson and Edwards [28].

Starting values of the parameters in the vector $\theta$ were set at $\eta$ = 0.5, $\beta$ = 0.05 and the random effect parameters, $\sigma_u^2, \sigma_v^2, \sigma_{v1}^2$ and $\rho_V$ were all set at 0.5 for simplicity. Other sets of starting values were also tested, and resulted in the same estimated parameters in each case. Convergence criteria within the algorithm was set at max($|\theta_k - \theta_{k-1}|$<0.0001, where $k$ = 1,

…, $K$ is the number of parameters to be estimated, and $\theta$ contains the parameters to be estimated.

The model-based kappa statistic was calculated for each individual data set, based upon the estimated values of the parameters from the corresponding generalized linear mixed model, and for model (a), a version of Cohen's kappa for multiple raters [31] was calculated.

Tables I and II display the mean estimates of the parameters and their associated standard errors from the simulation studies are described. We observe that the parameters $\eta$ and $\beta$ are estimated on average with little bias in each of the 12 models examined. Similarly, the almost-exact maximum likelihood mean estimates of the variance components $\sigma_u^2$ and $\sigma_v^2$ are nearly unbiased, although there is a slight increase in the level of bias observed for the variance components for model (c). The mean estimates of the correlation coefficient $\rho_V$ are underestimated for each simulation scenario for model (c), sometimes severely so. In the simplest model (a), the mean estimated Cohen's kappa is only slightly lower in value than the corresponding model-based kappa statistic when the prevalence term $\eta = 1$; however, the mean estimated Cohen's kappa is noticeably smaller than the mean estimated model-based kappa for $\eta = 3$ (Cohen's $\hat{\kappa} = 0.09$, $\hat{\kappa}_m = 0.21$) and the associated standard errors are large enough to suggest that there may be no significant difference between these two estimated measures of agreement.

## 5. Application to a breast cancer study

An agreement study was carried out by Beam *et al.* [9] where 148 randomly selected mammograms were classified by a large number of physicians randomly selected from a group of 294 physicians from the USA. The mammograms included both diseased and non-diseased cases. Data on a number of covariates, including the subject's age, the number of mammograms read in the previous year by each rater and the number of years of experience of the raters, were collected. The subjects' ages ranged from 40 to 85 years. The classifications were made using the BIRADS scale, and have been dichotomized here so that an outcome of 0 represents a mammogram classified as non-diseased, and 1 as diseased. Full details on the data collection can be found in Beam *et al.* [9]. Data on 104 randomly sampled physicians were analyzed using the models and measure of agreement described in Sections 2 and 3 and a summary of the pairwise agreement is presented in Table III. The age of each woman on whom the mammograms was taken was included as an indicator variable $x_i$, where $x_i = 1$ for a subject less or equal to 60 years of age. The level of experience of a physician was included as an indicator variable $z_2 = 1$ if a physician had 10 or more years experience of rating mammograms and 0 otherwise. These are two examples of models that can be fit to this data set; there are other models that could also be feasible.

Two generalized linear mixed models with a probit link function were fitted to this data set using McCulloch's MCEM algorithm; one model (i) to assess the overall agreement present between all the raters and subjects included in the study, and a second model (ii) to assess agreement after accounting for a subject's age and each rater's level of experience. These models are

$$\text{Model (i)}:\Phi^{-1}(p_{ij})=\eta+u_i+v_j,$$

$$\text{Model (ii)}:\Phi^{-1}(p_{ij})=\eta+\beta x_i+u_i+v_j+z_2 v_{1j}.$$

Details on the forms of $\rho_v$, var($\hat{\kappa}_m$) and the likelihood function $L(\boldsymbol{\theta}, \boldsymbol{y})$ for model (ii) are presented in the Appendix.

Table IV presents the parameter estimation and model-based and Cohen's kappa statistics for these two models. We observe a negative intercept for both models, reflecting the fact that over half of the mammograms were classified as not having cancer present. The negative regression coefficient for subject's age suggests that the odds in favor of a younger patient (less than 60 years) being classified as having a diseased mammogram is approximately 45 per cent of that of an older patient (over 60 years old). The variability observed between the subjects $(\sigma_u^2)$ is larger than the variability observed between the raters $(\sigma_v^2)$. Cohen's kappa for overall agreement between all raters was estimated at $\hat{\kappa} = 0.60$, whereas the model-based kappa $\hat{\kappa}_m = 0.53$ (se = 0.08). The agreement between highly experienced raters classifying mammograms of younger women ($z_2 = 1$ and $x_i = 1$) is $\hat{\kappa}_m = 0.53$ (se = 0.07), whereas chance-corrected agreement between less experienced raters is estimated as $\hat{\kappa}_m = 0.50$ (se = 0.08), suggesting little difference between these two estimated measures of agreement.

## 6. Discussion

In this paper, we have extended a model-based approach and measure of agreement, which are appropriate for use in large studies with the goal of improving diagnostic tests that wish to examine the reliability between many raters each classifying the same set of many subjects. This population-based approach, which is based upon the class of generalized linear mixed models allows for inference regarding the underlying diagnostic procedure, and for conclusions to be made regarding the populations of raters and subjects who are typically involved in the medical testing procedure of interest.

Extending the basic model and measure of agreement presented by Nelson and Edwards [28] to include covariate information that may affect both the prevalence of the disease and the agreement process yields valuable information about the roles of raters and subjects. For example, the significance of the level of experience of raters when classifying test items, such as mammograms, can be more closely examined. Other information can also be easily incorporated in this modeling approach, such as factors related to the subject's clinical history. The use of generalized linear mixed models allows for missing observations and unbalanced data, and easily incorporates large numbers of raters and subjects, unlike most other methods for assessing agreement. Specific details regarding the performance of individual raters included in the study can also be obtained easily by estimating the random effects.

Obtaining almost-exact maximum likelihood estimates of the parameters in the generalized linear mixed models requires the use of a computationally intensive algorithm, such as McCulloch's MCEM algorithm [32] or an equivalent [33]. At present, software packages do not have the capacity to obtain almost-exact maximum likelihood estimates for generalized linear mixed models with crossed random effects structures. The model-based kappa statistic can also be calculated using standard software such as R and SAS; a computer function written in the freely available software package R included in Table V.

As is the case for any class of models where the outcomes of interest are binary, larger data sets are required when more factors are to be included into the model, since more parameters are required to be estimated. This also ensures successful convergence of the MCEM algorithm in the parameter estimation process.

It is assumed in the current model setting that the random effects of the subjects and raters are normally distributed. Previous research [34] has shown that for a generalized linear mixed model with a Poisson link function, fixed effect parameters are estimated with little or no bias, whereas variance components are estimated with more bias and variability when non-normal random effects are modeled assuming normality.

Further extensions to the model include multiple independent classifications of each item by the individual raters. This can be useful when we wish to examine the variability of ratings made by any rater for a particular item, also known as 'intra-rater' variability. This can be flexibly included in the framework of the generalized linear mixed model and summary statistic proposed, and is a topic for future research. Other future directions for extending this class of models in the assessment of agreement include development of a model to allow for ordinal classification scales, and to investigate the inclusion of non-normal random effect distributions for the raters and items.

## Acknowledgments

The authors are grateful for the support provided by the following grant from the United States' Institutes of Health R03CA114783-01A1. We thank Dr Craig Beam for kindly providing us with the breast cancer data set. We also gratefully acknowledge helpful advice from Charles McCulloch, Patrick Graham, Robert Best and the comments from the editors and the referees.

Contract/grant sponsor: United States' Institutes of Health; contract/grant number: R03CA114783-01A1

## APPENDIX A

## DETAILS ON MODEL FITTING FOR APPLICATION

For model (i), the random effects $u_i$ and $v_j$, $i = 1, \dots, I$ and $j = 1, \dots, J$, are assumed to be normally distributed with zero means and variances $\sigma_u^2$ and $\sigma_v^2$, respectively. The random effects in model (*ii*) are assumed to be distributed as in Section 5. For model (ii), the term ρ takes the form

$$\rho = \frac{\sigma_1^2}{(\sigma_2^2 + \sigma_1^2 + 1)} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2 + z_2^2 \sigma_{v_1}^2 + 2 z_2 \rho_v \sigma_v \sigma_{v1} + 1}$$

and $\sigma_T^2=\sigma_u^2+\sigma_v^2+z_1^2\sigma_{v_{ij}}^2+2z_1\rho_v\sigma_v\sigma_{v1}+1$. The statistic $\hat{\kappa}_m$ is calculated using the estimated quantities, $\hat{\sigma}_u^2, \hat{\sigma}_v^2, \hat{\sigma}_{v1}^2$ and $\hat{\rho}$ from the fitted model. To calculate the variance for $\kappa_m$, let $\boldsymbol{\theta}=(\sigma_u^2, \sigma_v^2, \sigma_{v1}^2, \rho_v, \beta)$. Since $\kappa_m$ is a function of $\theta$, the multivariate delta theorem can be applied directly. The variance of $\hat{\kappa}_m$, var($\hat{\kappa}_m$) can be obtained using equation (7), where vector $\boldsymbol{h}$ takes the following form:

$$\boldsymbol{h}=\left(\frac{\delta\kappa_m}{\delta\sigma_u^2}, \frac{\delta\kappa_m}{\delta\sigma_v^2}, \frac{\delta\kappa_m}{\delta\sigma_{v1}^2}, \frac{\delta\kappa_m}{\delta\rho_v}, \frac{\delta\kappa_m}{\delta\beta}\right).$$

The matrix $\Sigma$ is of dimension 5 × 5 since the vector $\theta$ contains five elements, and takes the form

$$\boldsymbol{\Sigma}=-\left(E\left[\frac{\delta^2 logL}{\delta\boldsymbol{\theta}\delta\boldsymbol{\theta}}\right]\right)^{-1},$$

where the likelihood function of interest here is

$$L(\boldsymbol{\theta};\boldsymbol{y})=\prod_{i=1}^{I}\prod_{j=1}^{J}p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}}\frac{1}{2\pi\sigma_v\sigma_{v_1}\sqrt{1-\rho^2}}\exp\left\{\frac{-1}{2(1-\rho_v^2)}\left(\frac{v_j^2}{\sigma_v^2}+\frac{v_{1j}^2}{\sigma_{v1}^2}-\frac{2\rho_v v_j v_{1j}}{\sigma_v\sigma_{v1}}\right)\right\}\times\frac{1}{\sqrt{2\pi\sigma_u^2}}\exp\left\{\frac{-u_i^2}{2\sigma_u^2}\right\}$$

with $p_{ij} = \Phi(\eta+\beta x_i+u_i+v_j+z_2 v_{1j})$.

## References

1. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. The New England Journal of Medicine. 1994; 331(22):1493–1499. [PubMed: 7969300]
2. Yerushalmy J. The importance of observer error in the interpretation of photofluorograms and the value of multiple readings. International Tuberculosis Year Book. 1956; 24:110–124.
3. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960; 20:37–46.
4. Holmquist ND, McMahan CA, Williams OD. Variability in classification of carcinoma in situ of the uterine cervix. Archives of Pathology. 1967; 84:334–345. [PubMed: 6045443]
5. Sickles EA, Miglioretti DL, Ballard-Barbash R, Geller BM, Leung JWT, Rosenberg RD. Performance benchmarks for diagnostic mammography. Radiology. 2005; 235:775–790. [PubMed: 15914475]
6. Beam CA, Conant EF, Sickles EA. Factors affecting radiologist inconsistency in screening mammography. Academic Radiology. 2002; 9:531–540. [PubMed: 12458879]
7. Miglioretti DL, Smith-Bindman R, Abraham L, Brenner RJ, Carney PA, Bowles EJA, Bowles EJA, Buist DSM, Elmore JG. Radiologist characteristics associated with interpretive performance of diagnostic mammography. Journal of the National Cancer Institute. 2007; 99:1854–1863. [PubMed: 18073379]
8. Elmore JG, Carney PA. Does practice make perfect when interpreting mammography? Journal of the National Cancer Institute. 2002; 94:321–323. [PubMed: 11880465]
9. Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. Journal of the National Cancer Institute. 2007; 95:282–290.

10. Allsbrook WC, Mangold KA, Johnson MH, Lane RB, Lane CG, Epstein JI. Interobserver reproducibility of Gleason grading of prostatic carcinoma. Human Pathology. 2001; 32(1):81–88. [PubMed: 11172299]
11. Lipsitz SR, Laird NM, Brennan TA. Simple moment estimates of the $\kappa$-coefficient and its variance. Applied Statistics. 1994; 43:309–323.
12. Barlow W. Measurement of interrater agreement with adjustment for covariates. Biometrics. 1996; 52:695–702. [PubMed: 10766505]
13. Klar N, Lipsitz SR, Ibrahim JG. An estimating equations approach for modelling kappa. Biometrical Journal. 2000; 1:45–58.
14. Maclure M, Willett WC. Misinterpretation and misuse of the Kappa statistic. American Journal of Epidemiology. 1987; 126(2):161–169. [PubMed: 3300279]
15. Nelson JC, Pepe MS. Statistical description of interrater variability in ordinal ratings. Statistical Methods in Medical Research. 2000; 9:475–496. [PubMed: 11191261]
16. Kraemer HC, Periyakoil VS, Noda A. Kappa coefficients in medical research. Statistics in Medicine. 2002; 21:2109–2129. [PubMed: 12111890]
17. Tanner MA, Young MA. Modeling agreement among raters. Journal of the American Statistical Association. 1985; 80(389):175–180.
18. Agresti A. A model for agreement between ratings on an ordinal scale. Biometrics. 1988; 44:539–548.
19. Goodman LA. Simple models for the analysis of association in cross classifications having ordered categories. Journal of the American Statistical Association. 1979; 74:537–552.
20. Coull BA, Agresti A. Generalized log-linear models with random effects, with application to smoothing contingency tables. Statistical Modelling. 2003; 3:251–271.
21. Graham P. Modeling covariate effects in observer agreement studies: the case of nominal scale agreement. Statistics in Medicine. 1995; 14:299–310. [PubMed: 7724915]
22. Dawid AP, Skene AM. Maximum likelihood estimation of observer error-rates using the EM algorithm. Applied Statistics. 1979; 28:20–28.
23. Williamson JM, Manatunga AK. Assessing interrater agreement from dependent data. Biometrics. 1997; 54:707–714.
24. Uebersax JS, Grove WM. Latent class analysis of diagnostic agreement. Statistics in Medicine. 1990; 9:559–572. [PubMed: 2190288]
25. Coughlin SS, Pickle LW, Goodman MT, Wilkens LR. The logistic modeling of interobserver agreement. Journal of Clinical Epidemiology. 1992; 45(11):1237–1241. [PubMed: 1432004]
26. Lipsitz SR, Parzen M, Fitzmaurice GM, Klar N. A two-stage logistic regression model for analyzing inter-rater agreement. Psychometrika. 2003; 68(2):289–298.
27. Kraemer HC. Ramifications of a population model for $\kappa$ as a coefficient of reliability. Psychometrika. 1979; 44(4):461–472.
28. Nelson KP, Edwards D. On population-based measures of agreement for binary classifications. Canadian Journal of Statistics. 2008; 36(3):411–426.
29. Holford TR, Cronin KA, Mariotto AB, Feuer EJ. Changing patterns in breast cancer incidence trends. Journal of the National Cancer Institute Monographs. 2006; 36:19–25.
30. Agresti, A. An Introduction to Categorical Data Analysis. New York: Wiley; 1996.
31. Landis JR, Koch GG. Measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159–174. [PubMed: 843571]
32. McCulloch CE. Maximum likelihood algorithms for generalized linear mixed models. Journal of the American Statistical Association. 1997; 437:162–170.
33. Kuk AYC, Cheng YW. The Monte Carlo Newton–Raphson algorithm. Journal of Statistical Computing and Simulation. 1997; 59:233–250.
34. Nelson KP, Leroux BG. Properties and comparison of estimation methods in a log-linear generalized linear mixed model. Journal of Statistical Computation and Simulation. 2008; 78(3):367–384.

## Table I

Simulation results for the probit generalized linear mixed model and $\kappa_m$ based upon 100 data sets for $I = 50$ and $J = 50$ for two different sets of parameter values $\boldsymbol{\theta}=(\eta, \beta, \sigma_u^2, \sigma_v^2 \sigma_{v_1}^1, \rho_v)$.

| Parameter | True value | Model (a) | Model (b) | Model (c) |
|---|---|---|---|---|
| (*i*) $\eta = 1$, $\sigma_{u_0}^2 = \sigma_v^2 = \sigma_{v_1}^2 = 1$ | | | | |
| $\eta$ | 1 | 0.95 (0.17) | 0.99 (0.24) | 1.02 (0.36) |
| $\beta$ | 0.5 | — | 0.46 (0.25) | 0.48 (0.41) |
| $\sigma_u^2$ | 1 | 0.97 (0.20) | 0.94 (0.26) | 0.94 (0.21) |
| $\sigma_v^2$ | 1 | 1.02 (0.23) | 1.09 (0.31) | 0.97 (0.32) |
| $\sigma_{v_1}^2$ | 1 | — | — | 1.37 (0.46) |
| $\rho_{v01}$ | 0.25 | — | — | −0.0004 (0.001) |
| $\kappa$ | | 0.19 (0.04) | | |
| $\kappa_m$ | | 0.21 (0.03) | | |
| (*i*) $\eta = 1$, $\sigma_{u_0}^2 = \sigma_{v_0}^2 = \sigma_{v_1}^2 = 5$ | | | | |
| $\eta$ | 1 | 0.89 (0.29) | 0.71 (0.48) | 1.049 (0.50) |
| $\beta$ | 0.5 | — | 0.38 (0.67) | 0.43 (0.66) |
| $\sigma_u^2$ | 5 | 4.90 (0.84) | 4.65 (0.98) | 4.66 (0.92) |
| $\sigma_v^2$ | 5 | 5.26 (1.26) | 5.09 (1.36) | 4.28 (0.82) |
| $\sigma_{v_1}^2$ | 5 | — | — | 6.64 (2.24) |
| $\rho_v$ | 0.25 | — | — | 0.0004 (0.005) |
| $\kappa$ | | 0.28 (0.05) | | |
| $\kappa_m$ | | 0.29 (0.04) | | |

The three models fitted are: (a) $\Phi^{-1}(p_{ij}) = \eta+u_i+v_j$, (b) $\Phi^{-1}(p_{ij}) = \eta+\beta x_j+u_i+v_j$ and (c) $\Phi^{-1}(p_{ij}) = \eta+\beta x_j+u_i+v_j+z_2 v_{1j}$; $x_i \sim$ Bin(1,0.5), $z_2 \sim$ Bin(1,0.5). Cohen's kappa = $\kappa$ and model-based kappa = $\kappa_m$. Mean parameter estimates are presented with associated standard errors in parentheses.

## Table II

Simulation results for the probit generalized linear mixed model and model-based kappa statistic $\kappa_m$ based upon 100 data sets for $I$ = 50 and $J$ = 50 using two different sets of parameter values $\boldsymbol{\theta}=(\eta,\beta,\sigma_u^2,\sigma_v^2\sigma_{v_1}^1,\rho_v)$.

| Parameter | True value | Model (a) | Model (b) | Model (c) |
|---|---|---|---|---|
| (*ii*) η =3, $\sigma_{u0}^2=\sigma_{v0}^2=\sigma_{v_1}^2=1$ | | | | |
| η | 3 | 2.95 (0.25) | 3.05 (0.41) | 3.02 (0.41) |
| β | 0.5 | — | 0.51 (0.40) | 0.42 (0.08) |
| $\sigma_u^2$ | 1 | 1.00 (0.37) | 1.05 (0.30) | 0.76 (0.30) |
| $\sigma_v^2$ | 1 | 1.09 (0.38) | 1.14 (0.36) | 1.3 (0.52) |
| $\sigma_{v_1}^2$ | 1 | — | — | 1.25 (0.94) |
| $\rho_v$ | 0.25 | — | — | −0.004 (0.002) |
| κ | | 0.09 (0.04) | | |
| $\kappa_m$ | | 0.21 (0.05) | | |
| (*ii*) η = 3, $\sigma_{u0}^2=\sigma_{v0}^2=\sigma_{v_1}^2=5$ | | | | |
| η | 3 | 2.93 (0.33) | 2.87 (0.49) | 2.71 (0.29) |
| β | 0.5 | — | 0.45 (0.56) | 0.54 (0.31) |
| $\sigma_u^2$ | 5 | 5.00 (1.12) | 4.44 (1.04) | 3.99 (0.56) |
| $\sigma_v^2$ | 5 | 5.39 (1.51) | 4.34 (0.51) | 4.44 (1.59) |
| $\sigma_{v_1}^2$ | 5 | — | — | 5.21 (3.25) |
| $\rho_v$ | 0.25 | — | — | 0.003 (0.01) |
| κ | | 0.24 (0.06) | | |
| $\kappa_m$ | | 0.29 (0.05) | | |

The three models fitted are: (a) $\Phi^{-1}(p_{ij}) = \eta+u_i+v_j$, (b) $\Phi^{-1}(p_{ij}) = \eta+\beta x_j+u_i+v_j$ and (c) $\Phi^{-1}(p_{ij}) = \eta+\beta x_j+u_i+v_j+z_2v_{1j}$; $x_i$ ~ Bin(1,0.5), $z_2$ ~ Bin(1,0.5). Cohen's kappa = κ and model-based kappa = $\kappa_m$. Mean parameter estimates are presented with associated standard errors in parentheses.

## Table III

Summary of the pairwise agreement between 104 randomly selected physicians each independently classifying 148 slides for the presence ($y_{ij} = 1$) or absence ($y_{ij} = 0$) of breast cancer [9].

| | | Physician B | | |
|---|---|---|---|---|
| | **Category** | **Non-diseased** | **Diseased** | **Total** |
| Physician A | Non-diseased | 460 951 | 64 531 | 525 482 |
| | Diseased | 74 467 | 192 739 | 267 206 |
| | Total | 535 418 | 257 270 | 792 688 |

## Table IV

Results for the breast cancer data set.

| Parameter | Model (a) | Model (b) |
|---|---|---|
| $\eta$ | −0.83 (0.15) | −0.13 (0.02) |
| $\beta$ | — | −0.37 (0.03) |
| $\sigma_u^2$ | 3.54 (0.45) | 3.45 (0.40) |
| $\sigma_v^2$ | 0.25 (0.04) | 0.25 (0.03) |
| $\sigma_{v_1}^2$ | — | 0.24 (0.003) |
| $\rho_v$ | — | 0.001 (0.01) |
| Cohen’s kappa $\kappa$ | 0.60 | |
| Model-based kappa $\kappa_m$ | 0.53 (0.08) | |
| $\kappa_m(z_2 = 0, x_i = 1)$ | — | 0.50 (0.08) |
| $\kappa_m(z_2 = 1, x_i = 1)$ | — | 0.53 (0.07) |

The two models fitted are: (a) $\Phi^{-1}(p_{ij}) = \eta+u_i+v_j$ and (b) $\Phi^{-1}(p_{ij}) = \eta+\beta x_i+ u_i+v_j+z_2 v_{1j}$; $x_i$ is an indicator variable for $i$th subject’s age (1 for subjects less than or equal to 60 years of age, 0 for subjects greater than 60 years of age). The term $z_2 = 0$ for an inexperienced rater, and 1 for an experienced rater. Parameter estimates are presented with standard errors in parentheses.

## Table V

Function in *R* for calculating the two-rater model-based kappa statistic and its variance.

```
twokappafn = function(sigma2u, beta1)
{
integrand1 = function(z)
 {
  term1=pnorm(sqrt(sigma2u)*z − beta1)
  term2= pnorm(sqrt(sigma2u)*z + beta1)
fullintegrand= term1*term2*dnorm(z)
}
result1 = integrate (integrand1, lower=−100, upper=100)
integrand2=function(z)
{
  term3=(1-pnorm(sqrt(sigma2u)*z − beta1))
  term4=(1-pnorm(sqrt(sigma2u)*z + beta1))
fullintegrand2=term3*term4*dnorm(z)
}
result2= integrate(integrand2, lower=−100, upper=100)
res2=result2 $value
# two-rater model-based kappa
2*(res1 + res2)−1
}
# Calculation of variance: need to enter values for varbeta1, varsigmasqu,
sigmasqu and beta1.
integrand1a = function(z)
{term=(0.5*(1/sqrt(sigmasqu))*dnorm(sqrt(sigmasqu)*z +
beta1/2)*pnorm(sqrt(sigmasqu)*z − beta1/2)
+ 0.5*(1/sqrt (sigmasqu))*pnorm(sqrt(sigmasqu)*z +
beta1/2)*dnorm(sqrt(sigmasqu)*z − beta1/2))
*dnorm(z)*z }
result1a=integrate(integrand1a, lower=−100, upper=100)
integrand1b = function(z)
{term=(−0.5*(1/sqrt(sigmasqu))*dnorm(sqrt(sigmasqu)*z + beta1/2)*(1-
pnorm(sqrt(sigmasqu)*z − beta1/2))*z
−0.5*(1/sqrt(sigmasqu))*pnorm(sqrt(sigmasqu)*z + beta1/2)*dnorm(sqrt(sigmasqu)*z
− beta1/2))
*dnorm(z)*z}
result1b=integrate(integrand1b, lower=−100, upper=100)
vectorh1=result1a$value+result1b$value
integrand1b = function(z)
```

```
{term=(0.5*dnorm(sqrt(sigmasqu)*z + beta1/2)*pnorm(sqrt(sigmasqu)*z − beta1/2)
− 0.5*pnorm(sqrt(sigmasqu)*z

+ beta1/2)*dnorm(sqrt(sigmasqu)*z − beta1/2))*dnorm(z)}

result2a=integrate(integrand2a, lower=−100, upper=100)

integrand2b = function(z)

{term=(−0.5*dnorm(sqrt(sigmasqu)*z + beta1/2)*(1-pnorm(sqrt(sigmasqu)*z −
beta1/2))

+ 0.5*(1-pnorm(sqrt(sigmasqu)*z + beta1/2))*dnorm(sqrt(sigmasqu)*z −
beta1/2))*dnorm(z)}

result2b=integrate(integrand2b, lower=−100, upper=100)

vectorh2=result2a$value+result2b$value

covarmat = matrix(c(varsigmasqu,0,0,varbeta1), ncol=2)

vectorh = c(vectorh1, vectorh2)

varkappam = 4*(vectorh %*% covarmat %*% vectorh)

}
```