

# A model and measure of agreement for population-based studies

By

Kerrie P. Nelson†,  
*Department of Statistics, University of South Carolina, U.S.A.*

and Don Edwards,  
*Department of Statistics, University of South Carolina, U.S.A.*

May 4, 2007

†*Address for correspondence:*

Department of Statistics,  
University of South Carolina, 1523 Greene Street  
Columbia,  
SC 29208, U.S.A.

E-mail: kerrie@stat.sc.edu

Phone: (803) 777-7800

Fax: (803) 777-4048

# Abstract

Agreement between physicians in their classification of items such as mammograms for the presence of disease is an important tool in assessing the reliability of a diagnostic procedure, and the modeling of agreement data is a popular topic in the biomedical and social sciences. Interest often lies in assessing agreement in the underlying diagnostic procedure and making inferences for the populations of raters and items typically involved in the rating process. However, the majority of methods currently available are limited to inference for the specific groups of raters and items selected for study, and most do not apply when many raters are involved. In this paper we describe the use of generalized linear mixed models with crossed random effects to model agreement between many raters and items over the long-run for classifications made on a binary scale. These models flexibly allow for missing and unbalanced data, many raters and items, the inclusion of covariates that may influence the agreement process and most importantly, provides inference regarding the underlying diagnostic process and the populations of the typical raters and items involved in such classifications. To provide an overall measure of agreement we propose a summary model-based statistic which is easily interpretable in a manner similar to Cohen's kappa statistic, while avoiding some of the biases that arise in Cohen's kappa usage. The proposed agreement measure can also be used to describe agreement between subgroups of raters and items by utilizing available covariate information. Simulation studies demonstrate that the proposed approach provides unbiased chance-corrected estimates of agreement. The methods are applied to an agreement dataset involving the classification of mammograms for the presence/absence of breast cancer (Beam 2003).

**Key words:** agreement, model-based kappa, Cohen's kappa, generalized linear mixed model, crossed random effects.

# 1 Introduction

The best available screening or diagnostic tests in biomedical practice often involve the subjective classifications of items such as x-rays or biopsies by qualified professionals. For example, initial screening for breast cancer is commonly carried out through the visual interpretation of mammograms by a physician. The pervasive use of these subjective diagnostic procedures provides a strong motivation to develop methods to assess their reliability, which is usually measured as the level of agreement between a number of raters each classifying the same sample of items or subjects of interest. Strong agreement between raters is considered a necessary prerequisite for the effectiveness of any subjective procedure intended for diagnostic purposes, yet it is well-acknowledged that wide variability between raters is commonly observed (Yerushalmy 1956, Landis and Koch 1977, Elmore et al 1994). This has been demonstrated in several studies over the last few decades, including the assessment of slides in the detection of uterine cancer (Holmquist 1968), the agreement between diagnostic instruments in the detection of liver cancer (Henkelman et al 1967) and between physicians in their classification of mammograms (Elmore et al 1994, Beam et al 2003). Concerns about the subjective nature of classifications in breast imaging (American College of Radiology 2004) have prompted efforts to assess the levels of agreement and identify factors that may influence agreement within the process with the aim of improving the effectiveness of this procedure (Beam et al 2003, Miglioretti and Heagerty 2007).

While the study of inter-rater agreement involves examining the classification of a set of items by a group of raters, a primary interest often ultimately lies in how the results relating to agreement extend and generalize to the underlying diagnostic procedure and the populations of raters and items typically involved, rather than regarding the specific

groups of items and raters studied. In this paper, our focus is on examining agreement in the long-run situation, over many raters and items, so that inference can be made regarding the underlying diagnostic procedure. This is especially useful for commonly used procedures.

To date, methods for modeling agreement data fall into two broad categories - summary statistics and model-based approaches. Summary statistics include Cohen's kappa statistic (Cohen 1968), the intraclass correlation coefficient (Shrout and Fleiss 1979, Kraemer 1979, Bloch and Kraemer 1989), concordance correlation coefficient (Lin 1989) and others. In particular, Cohen's kappa is very popular among biomedical professionals due to its appealingly simple calculation and interpretation. While corrected for chance agreement, it has been shown to be susceptible to a number of biases, including bias and prevalence effects (Byrt et al 1993, Maclure and Willett 1987), Simpson's paradox (Thompson 2001), is also reliant upon the number of classes chosen for the classification scale (Maclure and Willett 1987), and is rarely comparable across different studies (Feinstein and Cicchetti 1990). Despite its shortcomings, Cohen's kappa is commonly applied in many settings, and has been extended in many different ways to incorporate many raters (Fleiss 1971, Light 1977, Kraemer 1980, Conger 1998), ordinal classification scales (Cohen 1968), adjustments for marginal and prevalence effects (Byrt et al 1993, Feinstein and Cicchetti 1990), and the inclusion of covariate information (Barlow 1996, Klar et al 2000, Landis and Koch 1977). Banerjee et al (1999) provide an excellent overview of the different summary statistics available for modelling agreement.

Model-based approaches provide a more complete and broader framework for examining agreement in any setting. Methods developed include log-linear models (Tanner and Young 1988, Agresti 1988, Goodman 1979, Coull and Agresti 2002, Graham 1995), latent class and trait models (Dawid and Skene 1979, Williamson and Manatunga 2003, Uebersax

and Grove 1990), logistic regression models (Coughlin 1992, Lipsitz 2003). Several of these approaches consider the raters as fixed effects, and provide inference to the specific raters and items under study, not extending to inference regarding the populations of such raters. Such methods work well when a small number of raters is involved, but may become increasingly complex and involve a large number of parameters when more than two or three raters are included. The log linear models and latent models can incorporate covariate information, however, few methods currently available are able to jointly incorporate covariate information and yield inference regarding the underlying diagnostic procedure and populations of raters and items of interest.

Our aims in this paper are to develop a model-based approach where the levels of agreement between any number of raters classifying any number of items can be examined, and inference made regarding the populations of raters and items and underlying diagnostic process. We propose a model-based measure of agreement which is simple to interpret in a similar manner to Cohen's kappa, while avoiding many of its weaknesses. To the authors' knowledge, there are no model-based summary statistics currently available that are comparable to Cohen's kappa. Our focus is on classifications made on a binary scale, where, for example, an item is classified as diseased or not diseased. The inclusion of covariates that may impact the agreement process within the model is discussed and can provide valuable insight into how relevant factors might influence the classifications made and the agreement between the raters. The class of generalized linear mixed models provides an appropriate and flexible framework to achieve these goals using a conditional approach, flexibly incorporating many raters and items, unbalanced data, and covariates considered important in the assessment of agreement. These models have the added advantage that inferences about the specific raters and items selected for study can also be examined.

Williamson and Manatunga (1997) consider a flexible latent variable model approach for

modelling ordinal agreement data where many raters and items and covariate information can be included. Their method differs from ours in that their inference is based upon a semi-marginal approach described by Qu et al (1995) which involves averaging over the raters and items in the study, while including a small random component to explain any additional variability. As each rater is included as an individual fixed effect, with a rater chosen arbitrarily as the reference rater, Williamson and Manatunga’s method is best-suited for a small number of raters. The inclusion of covariates in their model is mentioned in passing besides a comparison of two methods used in the classifications by the raters. Nelson and Pepe (2000) comment that latent model approaches tend to be based upon strict latent model assumptions which can be difficult to verify, and in Williamson and Manatunga’s model, this includes an assumption that the total variability sums to one. Results from latent models can also be difficult to interpret.

The remainder of the paper is as follows: Section 2 describes the class of proposed models with the inclusion of covariates, while Section 3 develops the proposed model-based measure of agreement. In Section 4, simulation studies are conducted to examine the properties of the model and measure of agreement. An application to a breast cancer dataset where mammograms were classified by a number of physicians is described in Section 5, and in Section 6 concluding remarks and discussion are made.

## **2 Models and measures of agreement**

### **2.1 A generalized linear mixed model for agreement**

A natural choice for modeling agreement data when the underlying diagnostic procedure is 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 item is rated as positive, or a “success” (for example, diseased) by the  $j$ th rater, and  $y_{ij} = 0$  otherwise. It is assumed that the raters and items selected for study are random samples from their respective populations. The outcomes can be modeled using a generalized linear mixed model with either a probit or logit link function, as follows:

$$(a) \text{ Probit: } \Phi^{-1}(p_{ij}) = \eta + u_i + v_j \quad (b) \text{ Logit: } \log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \eta + u_i + v_j, \quad (1)$$

for  $i = 1, \dots, I$  and  $j = 1, \dots, J$ . The quantity  $p_{ij} = pr(y_{ij} = 1)$  is the probability of the  $i$ th item being classified as a success by the  $j$ th rater, and the intercept constant  $\eta$  is a measure of the prevalence of successes in the data. When  $\eta$  is large, the overall frequency of successes in the data is high. The terms  $u_i$  and  $v_j$  represent random effects for the  $i$ th item and  $j$ th rater respectively with assumed  $\text{Normal}(0, \sigma_u^2)$  and  $\text{Normal}(0, \sigma_v^2)$  distributions. A positive (negative) value for  $u_i$  suggests that the  $i$ th item is more (less) likely than other such items to be classified as a success over many raters. A positive (negative) value for  $v_j$  reflects a rater who is more liberal (cautious) in classifying an item as a success over many items. A large value of  $\sigma_u^2$  is indicative of items which are easy to distinguish from one another. While the logit link function is commonly employed in practice when modelling dichotomous outcomes, we will focus on the probit link function for ease of mathematics. Nelson et al (2007) demonstrate that nearly identical results are yielded for the logit link generalized linear mixed model.

## 2.2 A generalized linear mixed model for agreement with covariates

In any rating process, various factors are likely to influence the classification of items made by raters and in the agreement between the raters. For example, in the classification of

mammograms, the age of the woman on whom the mammogram is taken is an important factor as it is well established that the prevalence of breast cancer increases with a woman's age (Feuer et al 1999), thus the mammogram of an older woman is more likely to display breast cancer than that of a younger woman.

Many agreement models, including loglinear and logistic regression models, have an agreement index as the response, for example, 0 = two raters disagree, 1 = two raters agree in their classification of a single item (Agresti 1996, Tanner and Young 1985, Coughlin 1992, Lipsitz 2003). However, in the generalized linear mixed model setting considered here, and in the model described by Williamson and Manatunga (1997), the response variable  $y_{ij}$  is the classification made on the  $i$ th item by the  $j$ th rater. Thus when a covariate is to be included into the model, careful consideration has to be given as to whether it is likely to directly impact the prevalence so that it can be included as a fixed effect, or if it is more likely to influence the agreement between the raters and can then be included as a random component of the model. Covariates that influence both the prevalence and agreement can be included as both a fixed effect and random component.

Incorporating covariates into the model as fixed effects accounts for their influence on the prevalence of positive classifications or "successes" such as diseased mammograms. Previous studies (Miglioretti and Heagerty 2007, Allsbrook et al 2001, Linver et al 2002) have demonstrated that covariates 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 can impact how a rater classifies an item, and thus may also influence the agreement between raters. A patient with a clinical history of associated risk factors may present with a more severely diseased mammogram thus making it easier for all the raters to distinguish the presence of disease; raters are more likely to agree on a severely diseased item. Covariates that are likely to impact agreement

can be included as random effects in the model. The general agreement model then takes the form:

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

where  $\mathbf{x}_{ij}$  is the vector of covariates associated with the  $i$ th item classified by the  $j$ th rater. The vectors  $\mathbf{d}_i$  and  $\mathbf{d}_j$  represent the design vectors of the random effects for the  $i$ th item and  $j$ th rater respectively, and the vectors  $\mathbf{u}_i$  and  $\mathbf{v}_j$  contain the associated random effects for the items and raters. The random effects are assumed to follow multivariate normal distributions, i.e.

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

where the covariance matrices  $\boldsymbol{\Sigma}_u$  and  $\boldsymbol{\Sigma}_v$  are of dimensions  $p \times p$  and  $q \times q$  respectively. Correlation can be assumed between the item random effects, and between the rater random effects. More complex random effect structures can be utilized if necessary. Estimated values of the random effect vectors  $\mathbf{u}_i$  and  $\mathbf{v}_j$  can be used to ascertain the individual effects of the  $j$ th rater and  $i$ th item included in the study.

### 3 Measures of agreement

#### 3.1 Population-based measures of agreement

In any setting where the underlying diagnostic procedure is the primary focus, the true value of the observed or “raw” agreement rate, denoted as  $p_0$ , is the probability that two randomly selected raters agree in their classification of a randomly selected item (i.e. both raters rate the item as a 1 or as a 0). In this longrun setting over many raters and items and without assuming any specific model, the observed agreement rate cannot take values

less than 0.5 (Bloch and Kraemer 1989, Nelson et al 2007). This minimum value of 0.5 is achieved in the worst-case scenario when the randomly selected items are completely indistinguishable from each other, and the raters classify each item at random with a fifty-fifty chance as a success or failure. As described in Nelson et al (2007), the form of the observed probability of agreement for raters  $j$  and  $j'$  ( $j \neq j'$ ) and item  $i$  is:

$$\begin{aligned}
p_0 &= pr\{(y_{ij} = 1) \cap (y_{ij'} = 1)\} + pr\{(y_{ij} = 0) \cap (y_{ij'} = 0)\} \\
&= 1 - 2pr\{(y_{ij} = 1) \cap (y_{ij'} = 0)\} \quad (\text{in any long-run agreement setting}) \\
&= 1 - 2 \int_{-\infty}^{\infty} \Phi\left(\frac{z\sqrt{\rho} + \eta^*}{\sqrt{1-\rho}}\right) \left[1 - \Phi\left(\frac{z\sqrt{\rho} + \eta^*}{\sqrt{1-\rho}}\right)\right] \phi(z) dz \quad (3)
\end{aligned}$$

under the probit model, where  $\rho = \sigma_u^2/\sigma_u^2 + \sigma_v^2 + \sigma_w^2$  where  $\sigma_w^2 = 1$  is the variance of the probit function and  $\eta^* = \eta/\sqrt{\sigma_u^2 + \sigma_v^2 + \sigma_w^2}$ ;  $\phi$  and  $\Phi$  are the standard normal pdf and cdf respectively. Since raters  $j$  and  $j'$  are randomly selected from the population of such raters, it is irrelevant as to which rater is assigned the label  $j$  and which is assigned  $j'$ , hence the symmetry argument in the second line of (5) holds (Bloch and Kraemer 1989, Banerjee et al 1999). The observed probability of agreement  $p_0$  depends only on the parameters  $\eta^*$  and  $\rho$ . It can also be shown (Nelson et al 2007) that  $p_0$  is an increasing function of  $|\eta^*|$  and  $\rho$ , and a symmetric function of  $\eta^*$ , as is displayed in Figure 1. These plots suggest that as the items become more distinguishable from one another (as  $\sigma_u^2$  increases) relative to the overall variability in the data, that the observed probability of agreement between the raters increases and at a faster rate if the prevalence of success is low.

Chance agreement, denoted as  $p_c$ , is the probability that two randomly selected raters  $j$  and  $j'$  classify two randomly selected items  $i$  and  $i'$ , ( $i \neq i'$ ) in the same way. The true value of chance agreement based upon any populations of raters and items (in any

longrun setting with no assumed model) is:

$$\begin{aligned}
p_c &= pr\{(y_{ij} = 1) \cap (y_{i'j'} = 1)\} + pr\{(y_{ij} = 0) \cap (y_{i'j'} = 0)\} \\
&= 1 - 2pr\{(y_{ij} = 1) \cap (y_{i'j'} = 0)\}, && \text{(in any longrun setting)} \\
&= 1 - 2\Phi(\eta^*)[1 - \Phi(\eta^*)] && \text{(under the probit model),}
\end{aligned}$$

so that chance agreement is based solely on the standardized prevalence  $\eta^*$  when the probit link function is assumed.

### 3.2 A model-based measure of agreement

The quantity  $\rho = \sigma_u^2 / (\sigma_u^2 + \sigma_v^2 + \sigma_w^2)$  described in Section 3.1 is itself appealing as a measure of agreement, and is similar in spirit to the intra-class correlation coefficient which, for example, provides a measure of the variability between the items relative to the total variability present in the data (Shrout and Fleiss 1979, Williamson and Manatunga 1997). Large variability between items  $\sigma_u^2$  relative to the variability  $\sigma_v^2$  between raters suggests that the items are clearly distinguishable from one another, and will lead to strong agreement between the raters and consequently a large value of  $\rho$ .

As mentioned earlier, Cohen's kappa is a very popular summary measure of agreement. Due to the current widespread use of this statistic, we propose an alternative chance-corrected agreement measure based upon the class of generalized linear mixed model that will be comparable in ease of use and interpretability as Cohen's kappa, while avoiding some of the issues surrounding the Cohen's kappa, and in addition, allows for inference to be made regarding the underlying diagnostic procedure.

Denoted as  $\kappa_m$ , the proposed model-based kappa statistic is based upon the probability of observed agreement  $p_0$ , adjusted for chance agreement, and scaled to lie on the same scale

as Cohen's kappa. The quantity  $p_0$  is derived from the parameters including  $\sigma_u^2$  and  $\sigma_v^2$  in the generalized linear mixed model in equation (1). Chance agreement is accounted for by setting the prevalence term  $\eta^*$  to zero, which minimizes the effects of chance agreement on the probability of observed agreement  $p_0$ . This effect occurs since as  $p_0$  increases in value towards one, higher values of  $|\eta^*|$  (i.e. how many 'successes', or ones, are assigned by the raters in the long run) lead to raters being more likely to classify an item as a success, thus naturally inflating the probability of observed agreement  $p_0$ . The proposed model-based measure of agreement in its simplest form without the inclusion of covariates is:

$$\begin{aligned}
\kappa_m &= 2p_0(\eta = 0, \sigma_u^2, \sigma_v^2) - 1 \\
&= 2 \left[ 1 - 2 \int_{-\infty}^{\infty} \Phi \left( \frac{z\sqrt{\rho}}{\sqrt{1-\rho}} \right) \left[ 1 - \Phi \left( \frac{z\sqrt{\rho}}{\sqrt{1-\rho}} \right) \right] \phi(z) dz \right] - 1 \\
&= 1 - 4 \int_{-\infty}^{\infty} \Phi \left( \frac{z\sqrt{\rho}}{\sqrt{1-\rho}} \right) \left[ 1 - \Phi \left( \frac{z\sqrt{\rho}}{\sqrt{1-\rho}} \right) \right] \phi(z) dz, \tag{4}
\end{aligned}$$

under the probit link model, where  $0 \leq \kappa_m \leq 1$ . The corresponding maximum likelihood estimator,  $\hat{\kappa}_m = 2p_0(0, \hat{\sigma}_u^2, \hat{\sigma}_v^2) - 1$ , is based upon the corresponding maximum likelihood estimates of  $\sigma_u^2$  and  $\sigma_v^2$  obtained from fitting the generalized linear mixed model in (1). The interpretation of  $\kappa_m$  is very simple and can be summarized in Table 1.

Applying the multivariate delta theorem, the asymptotic variance of the proposed statistic is:

$$\begin{aligned}
var(\hat{\kappa}_m) &= \frac{16}{IJ} \left[ \left\{ \int_{-\infty}^{\infty} \left( \frac{1}{2\hat{\rho}(1-\hat{\rho})} \left( \frac{z\sqrt{\hat{\rho}}}{\sqrt{1-\hat{\rho}}} \right) \phi \left( \frac{z\sqrt{\hat{\rho}}}{\sqrt{1-\hat{\rho}}} \right) \left[ 1 - 2\Phi \left( \frac{z\sqrt{\hat{\rho}}}{\sqrt{1-\hat{\rho}}} \right) \right] \right) \phi(z) dz \right\}^2 \times \right. \\
&\quad \left. \left( \frac{2\hat{\sigma}_u^4}{(\hat{\sigma}_u^2 + \hat{\sigma}_v^2 + \hat{\sigma}_w^2)^4} [(\hat{\sigma}_v^2 + \hat{\sigma}_w^2)^2 + \hat{\sigma}_v^4] \right) \right].
\end{aligned}$$

where  $I$  and  $J$  are the numbers of items and raters respectively randomly selected for

study.

### 3.3 A model-based measure of agreement with covariates

The proposed model-based measure of agreement  $\kappa_m$  can also incorporate covariate information. The inclusion of covariates as fixed and random effects effects allows for a more in depth assessment of agreement for subgroups of raters and items of interest. When covariates are included as fixed and random components, in the generalized linear mixed model described in equation (2), the model-based kappa statistic is based upon the observed probability of agreement  $p_0(\eta^* = 0, \boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_v, \boldsymbol{\beta}, \mathbf{x}_{ij}, \mathbf{d}_i, \mathbf{d}_j, \mathbf{d}_{j'})$ , conditional on the specified values of the covariates and variance components, and takes the form:

$$\begin{aligned} \kappa_m &= 1 - 4 \int_{-\infty}^{\infty} \Phi \left( \frac{\beta' x_{ij} + z \sigma_{d_i}}{(1 + \sigma_{d_j}^2)^{\frac{1}{2}}} \right) \left[ 1 - \Phi \left( \frac{\beta' x_{ij'} + z \sigma_{d_i}}{(1 + \sigma_{d_{j'}}^2)^{\frac{1}{2}}} \right) \right] \phi(z) dz, \\ &= 1 - 4 \int_{-\infty}^{\infty} \Phi \left( \frac{z \sqrt{\rho_{ij}} + \frac{\beta' x_{ij}}{\sqrt{\sigma_{T_{ij}}^2}}}{(1 - \rho_{ij})^{\frac{1}{2}}} \right) \left[ 1 - \Phi \left( \frac{z \sqrt{\rho_{ij}} + \frac{\beta' x_{ij'}}{\sqrt{\sigma_{T_{ij'}}^2}}}{(1 - \rho_{ij'})^{\frac{1}{2}}} \right) \right] \phi(z) dz \quad (5) \end{aligned}$$

where  $z$  is a standard normal random variable, and the terms  $\sigma_{d_i}^2$  and  $\sigma_{d_j}^2$  represent the variances of the sums of the normally distributed random effect components  $\mathbf{d}'_i \mathbf{u}_i$  and  $\mathbf{d}'_j \mathbf{v}_j$  for the  $i$ th item and  $j$ th rater respectively. The quantity  $\sigma_{T_{ij}}^2 = \sigma_{d_i}^2 + \sigma_{d_j}^2 + 1$  is a measure of the total variability present in the model, given the covariate values of the  $i$ th item and  $j$ th rater. The term  $\rho_{ij} = \sigma_{d_i}^2 / (\sigma_{d_i}^2 + \sigma_{d_j}^2 + 1)$  is, in a similar manner for the simpler model, a measure of the item distinguishability relative to the variability between two raters with the same covariate information, given the covariate information associated with the  $i$ th item. The exact form of the variance of  $\kappa_m$  is dependent upon the random effect vectors  $\mathbf{d}_i$  and  $\mathbf{d}_j$  and the assumed correlation structures of the random effects. The model-based kappa is a function of the coefficient  $\rho_{ij}$ , i.e.  $\kappa_m = f(\rho_{ij})$ , where in turn  $\rho_{ij}$  is a function

of  $\mathbf{d}_i, \mathbf{d}_j, \sigma_{d_i}^2, \sigma_{d_j}^2$ . The multivariate delta theorem can then be applied to yield:

$$\begin{aligned} \text{var}(\kappa_m) &= 16 \text{var} \left( \int_{-\infty}^{\infty} \Phi \left( \frac{z\sqrt{\rho_{ij}} + \frac{\beta' x_{ij}}{\sqrt{\sigma_{T_{ij}}^2}}}{(1 - \rho_{ij})^{\frac{1}{2}}} \right) \left[ 1 - \Phi \left( \frac{z\sqrt{\rho_{ij}} + \frac{\beta' x_{ij'}}{\sqrt{\sigma_{T_{ij'}}^2}}}{(1 - \rho_{ij'})^{\frac{1}{2}}} \right) \right] \phi(z) dz \right) \\ &= \frac{16}{IJ} \mathbf{g}(\mathbf{h} \boldsymbol{\Sigma}_\sigma^{-1} \mathbf{h}') \mathbf{g}, \end{aligned}$$

where

$$\mathbf{h} = \left( \frac{\delta \rho_{ij}}{\delta \boldsymbol{\Sigma}_{u_1, \dots, p}}, \frac{\delta \rho_{ij}}{\delta \boldsymbol{\Sigma}_{v_1, \dots, q}} \right), \quad g = \frac{\delta \hat{\kappa}_m}{\delta \rho_{ij}} \quad \text{and} \quad \boldsymbol{\Sigma}_\sigma = -\frac{1}{IJ} E \left[ \frac{\delta^2 \log L}{\delta \theta_{p+q} \delta \theta_{p+q}} \right],$$

and  $\boldsymbol{\theta}$  is the vector of all variance component terms contained in  $\boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_v$  matrices. For very complex random effect structures, bootstrapping may provide an attractive alternative for obtaining an estimated variance for  $\hat{\kappa}_m$ .

Including covariates allows for the assessment of agreement between subgroups of items and/or raters. For example,  $\kappa_m$  can be calculated to measure agreement between very experienced raters, and another  $\kappa_m$  can be calculated to assess the agreement between less experienced raters. One advantage of the model-based kappa statistic is that all the data has been utilized to estimate the parameters, even when agreement between sub-groups are of interest. An overall summary measure of agreement can be obtained by fitting the simplest form of the generalized linear mixed model presented in equation (1).

## 4 Simulation Studies

Simulation studies were carried out to investigate the properties of the proposed model and measure of agreement. The simulations were based upon three probit generalized

linear mixed models (equation (6)). The response variable  $y_{ij}$  represents the classification made by the  $j$ th rater on the  $i$ th item, and equals 0 for an item classified as not diseased and 1 otherwise. These models are:

$$\text{Model (a): } \Phi^{-1}(p_{ij}) = \eta + u_{0i} + v_{0j}, \quad u_0 \sim N(0, \sigma_u^2), v_0 \sim N(0, \sigma_v^2)$$

$$\text{Model (b): } \Phi^{-1}(p_{ij}) = \eta + \beta x_i + u_{0i} + v_{0j} \quad u_0 \sim N(0, \sigma_u^2), v_0 \sim N(0, \sigma_v^2)$$

$$\text{Model (c): } \Phi^{-1}(p_{ij}) = \eta + \beta x_i + u_{0i} + v_{0j} + d_{1j}v_{1j},$$

$$\begin{pmatrix} u_{0i} \\ v_{0j} \\ v_{1j} \end{pmatrix} \sim \text{MVN} \left( \mathbf{0}, \begin{bmatrix} \sigma_{u_0}^2 & 0 & 0 \\ 0 & \sigma_{v_0}^2 & \rho_{v_{01}}\sigma_{v_0}\sigma_{v_1} \\ 0 & \rho_{v_{01}}\sigma_{v_0}\sigma_{v_1} & \sigma_{v_1}^2 \end{bmatrix} \right), \quad (6)$$

where  $i = 1, \dots, I$ ,  $j = 1, \dots, J$  and  $\eta$  is the prevalence or intercept term. The random effects  $u_{0i}$  and  $v_{0j}$  relate to the variability observed between the items and raters respectively and are assumed normally distributed. The additional random effect in model (c) reflects a factor that is likely to influence agreement between the raters, such as a rater's level of experience ( $d_{ij} = 1$  for a very experienced rater, and 0 otherwise). The term  $v_{1j}$  is randomly generated from a Bin(1,0.5) distribution,  $j = 1, \dots, J$ . In models (b) and (c),  $x_i$  represents a fixed covariate value for the  $i$ th item, (for example, the age of the  $i$ th subject). In the simulations,  $x_i$  is randomly generated from a Bin(1,0.5) distribution,  $i = 1, \dots, I$ . Fifty raters and items were included in the simulations such that  $I = 50$  and  $J = 50$ . Two different values for the measure of prevalence,  $\eta = 1$  and 3 were used to assess the influence of prevalence on the estimation of the model parameter estimates and measure of agreement. The fixed regression coefficient  $\beta = 0.5$ . Different values of the variance components were also included to assess the effects of increasing variability: for all three models,  $\sigma_{u_0}^2, \sigma_{v_0}^2$  and  $\sigma_{v_1}^2$  were all set at 1 and then 5, and for model (c),

$\rho_{v_{01}}$  was set at 0.25. One hundred datasets were randomly generated according to each simulation scenario.

The datasets were individually fitted using a Monte-Carlo expectation-maximization algorithm (MCEM) developed by McCulloch (1997) to obtain almost-exact maximum likelihood estimates of the parameters in the generalized linear mixed model. A brief description of this algorithm is as follows. A Metropolis-Hastings step is used to sample sets of the (unobserved) random effects by drawing vectors  $\mathbf{u}_0 = (u_{01}, u_{02}, \dots, u_{0I})$  and  $\mathbf{v}_0 = (v_{01}, v_{02}, \dots, v_{0J})$  from their conditional distributions  $f(\mathbf{u}|\mathbf{y})$  and  $f(\mathbf{v}|\mathbf{y})$ . At each iteration,  $m$  sets of random effects are drawn, and the value of  $E(l(\boldsymbol{\theta}^t)|\mathbf{y})$  is estimated through the use of Monte-Carlo averaging, where  $l(\boldsymbol{\theta}^t)|\mathbf{y}$  is the log-likelihood function of the complete data  $(\mathbf{y}, \mathbf{u}, \mathbf{v})$ . The value of  $m$  ranged from 1000 for earlier iterations up to 500,000 for the final iterations. The parameter vector  $\boldsymbol{\theta}$  was updated at each iteration using the one-step expectation-maximization algorithm. Convergence of the parameters was considered successful when the difference between the previous estimates and the updated estimates took a maximum value of 0.00001. More details on this algorithm can be found in McCulloch (1997). This algorithm was implemented using a program written in C programming language.

Starting values of the parameters were set at  $\eta = 0.5, \beta_1 = 0.05$  and the random effect parameters,  $\sigma_{u_0}^2, \sigma_{u_1}^2, \sigma_{v_0}^2$  and  $\rho_{u_{01}}$  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  $k = 1, \dots, K$  is the number of parameters to be estimated, and  $\boldsymbol{\theta}$  contains the parameters to be estimated.

The model-based kappa statistic was calculated for each individual dataset, 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 (Fleiss 1971) was calculated.

Tables 2 and 3 display the mean estimates of the parameters and their associated standard errors from the simulation studies described. We observe that the mean parameter estimates of  $\eta$  and  $\beta$  are nearly unbiased in each of the 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  $\rho$  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  $\eta = 3$  which indicates the tendency for Cohen’s kappa to overcorrect for chance agreement and thus underestimate the true level of agreement between raters.

## **5 Application to breast cancer and mammogram classification data**

An agreement study was carried out by Beam et al (2003) where 148 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, and data on a number of covariates, including the subject’s age, the number of mammograms read in the previous year by the individual rater, and the number of years of experience of the raters was collected. The subjects’ ages ranged from 40 to 85 years.

The original cancer classifications were made using the BIRADS scale (American College of Radiology 2004), 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 (2003). Data on 104 physicians was analyzed using the models and measure of agreement described in Sections 2 and 3. The age of a subject 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  $d_j = 1$  if a physician had ten or more years experience of rating mammograms and 0 otherwise. Table 4 presents a summary of the pairwise agreement between all pairs of raters for each item. Note that each pair of classifications is only included once in the table.

Three generalized linear mixed models with a probit link function of increasing complexity were fitted to this dataset using McCulloch’s MCEM algorithm (previously described in Section 4 above). Table 5 presents the parameter estimation of the three models, and corresponding values of the model-based and Cohen’s kappa statistics from these analyses. We observe from that the prevalence term  $\eta$  is negativr for all three fitted models, indicating that over half of the mammograms were classified as not having cancer present. The negative beta coefficient in models (b) and (c) suggests that the odds in favor of a younger patient being classified as having a diseased mammogram is approximately 45% of that of an older patient (over 60 years old). The variability observed between items is larger than the variability observed between the raters. In this dataset, Cohen’s kappa was estimated at  $\hat{\kappa} = 0.604$ , while the model-based kappa was estimated as  $\hat{\kappa}_m = 0.529$ , suggestive of lower agreement between raters. One possible reason for a larger value for Cohen’s kappa is a bias effect which can inflate the value of Cohen’s kappa. The agreement between highly-experienced raters classifying mammograms of younger women

(where  $d_{1j} = 1$  and  $x_i = 1$ ) is estimated using equation (6) to be  $\hat{\kappa}_m = 0.5323$  while chance-corrected agreement between less experienced raters is estimated as  $\hat{\kappa}_m = 0.4993$ , suggesting a higher level of agreement between more experienced raters. Comparing the agreement between all raters classifying mammograms of younger women with that of older women, we observe that there is higher agreement for younger women ( $\hat{\kappa}_m = 0.546$  (younger women) versus  $\hat{\kappa}_m = 0.509$  (older women)). This could be due to that fact that while breast cancer is more common among the elderly, younger women can present with more aggressive forms of breast cancer, which would lead to mammograms which more clearly display the disease, and consequently lead to higher agreement between raters.

## 6 Discussion

Our emphasis in this paper lies in examining agreement in an underlying diagnostic procedure over many raters and items, where the raters and items chosen for study are assumed to be randomly selected from their respective populations. The class of generalized linear mixed models provides a flexible framework for incorporating the classifications of many raters and items, unbalanced and missing data, and the inclusion of covariate information, and have wide applicability to any agreement setting where raters are subjectively classifying items according to a pre-defined binary scale. Agreement between subgroups of raters and/or items can also be easily examined using these models. By including covariates in the agreement model, their effects on the prevalence of success and agreement between the raters can be assessed and accounted for. We propose a summary chance-corrected measure of agreement that is model-based and interpretable in a simple manner akin to Cohen's kappa. We observe that the model-based kappa statistic more appropriately corrects for chance agreement in the population-based setting, while the use of Cohen's kappa can lead to biased estimates of agreement, particularly when the true

prevalence is close to 0 or 1. This would be a common scenario when raters are classifying items for a rare disease.

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 (McCulloch 1997), or an equivalent (Kuk and Cheng 1997). At present, software packages do not have the capacity to obtain almost-exact maximum likelihood estimates for models with crossed random effects structures. The programs used here to fit the models of agreement are available from the first author on request.

The class of generalized linear mixed models can also be used when the agreement between a fixed number of raters is of interest, rather than inference regarding the population of typical raters. One such situation where this may be of interest is when the raters are different diagnostic instruments, such as in Henkelman et al (1990). The generalized linear mixed model can easily incorporate the raters (instruments) as individual fixed effects, and the randomly selected items as random effects. Since the random effects structure is no longer crossed, the model may be easily implemented in available software packages such as SAS or R.

As is the case for any class of models where the outcomes of interest are binary, larger datasets will be required when more covariates are included into the model, requiring more parameters to be estimated. This is to ensure successful convergence of the MCEM algorithm in the parameter estimation process when using generalized linear mixed models.

A further general extension to the model might 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 effects of non-normal random effect distributions for the raters and items.

## **Acknowledgements**

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

## REFERENCES

- Agresti, A. (1988). A model for agreement between ratings on an ordinal scale. *Biometrics*, **44**, 539-548.
- Agresti, A. (1996). An introduction to categorical data analysis. John Wiley and Sons, Inc. New York.
- Allsbrook, W.C., Mangold, K.A., Johnson, M.H., Lane, R.B., Lane, C.G., Epstein, J.I. (2001). Interobserver reproducibility of Gleason grading of prostatic carcinoma. *Human Pathology*, **32** (1), 81-88.
- American College of Radiology. (1993). Breast imaging reporting and data system (BI-RADS). Reston, VA.
- Banerjee, M., Capozzoli, M., McSweeney, L. and Sinha, D. (1999). Beyond kappa: a review of interrater agreement measures. *The Canadian Journal of Statistics*, **27** (1), 3-23.
- Barlow, W. (1996). Measurement of interrater agreement with adjustment for covariates. *Biometrics*, **52**, 695-702.
- Beam, C.A., Conant, E.F. and Sickles, E.A. (2003). Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. *Journal of the National Cancer Institute*, **95** (4), 282-290.
- Bloch, D.A. and Kraemer, H.C. (1989).  $2 \times 2$  kappa coefficients: Measurements of agreement and association. *Biometrics*, **45**, 269-287.
- Byrt, T., Bishop, J., and Carlin, J.B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, **46** (5), 423-429.

- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37-46.
- Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213-220.
- Conger, A.J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, **88**, 322-328.
- Coughlin, S.S., Pickle, L.W. et al. (1992). The logistic modeling of interobserver agreement. *Journal of Clinical Epidemiology*, **45 (11)**, 1237-1241.
- Coull, B.A. and Agresti, A. (2003). Generalized log-linear models with random effects, with application to smoothing contingency tables. *Statistical Modelling*, **3**, 251-271.
- Dawid, A.P. and Skene, A.M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, **28**, 20-28.
- Nelson, K.P., Edwards, D., Gamishev T., and Kozarev, R. (2007). Population-based measures of agreement. Technical Report, Department of Statistics, University of South Carolina.
- Elmore, J.G., Wells, C.K., Lee, C.H., Howard, D.H. and Feinstein, A.R. (1994). Variability in radiologists' interpretations of mammograms. *The New England Journal of Medicine*, **331 (22)**, 1493-1499.
- Feinstein, A.R. and Cicchetti, D.V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, **43 (6)**, 543-549.
- Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76**, 378-382.

- Goodman, L.A. and Kruskal, W.H. (1954). Measures of association for cross classification. *Journal of the American Statistical Association*, **49**, 732-764.
- Graham, P. (1995). Modeling covariate effects in observer agreement studies: the case of nominal scale agreement. *Statistics in Medicine*, **14**, 299-310.
- Henkelman, R.M., Kay, I. et al. (1967). Receiver Operator Characteristic (ROC) analysis without truth. *Medical Decision Making*, **10(1)**, 24-29.
- Holmquist, N.D., McMahan, C.A. et al. (1967). Variability in classification of carcinoma in situ of the uterine cervix. *Archives of Pathology*, **84**, 334-345.
- Klar, N, Lipsitz, S.R. and Ibrahim, J.G. (2000). An estimating equations approach for modelling kappa. *Biometrical Journal*, **1**, 45-58.
- Kraemer, H.C. (1979). Ramifications of a population model for  $\kappa$  as a coefficient of reliability. *Psychometrika*, **44 (4)**, 461-472.
- Kraemer, H.C. (1980). Extension of the kappa coefficient. *Biometrics*, **36**, 207-216.
- Kuk, A.Y.C. and Cheng, Y.W. (1997). The Monte Carlo Newton-Raphson algorithm. *Journal of Statistical Computing and Simulation*, **59**, 233-250.
- Landis, J.R. and Koch, G.G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, **33**, 363-374.
- Landis, J.R. and Koch, G.G. (1977). A one-way components of variance model for categorical data. *Biometrics*, **33**, 671-679.
- Landis, J.R. and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159-174.

- Light, R.J. (1971). Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological Bulletin*, **76**, 365-377.
- Lin, L.I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, **45**, 255-268.
- Linver, M.N., Paster, S.B., Rosenberg, R.D., Key, C.R., Stidley, C.A., King, W.V. (1992). Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,633 cases. *Radiology*, **184**, 39-43.
- Lipsitz, S.R., Parzen, M et al. (2003). A two-stage logistic regression model for analyzing inter-rater agreement. *Psychometrika*, **68** (2), 289-298.
- Maclure, M. and Willett, W.C. (1987). Misinterpretation and misuse of the Kappa statistic. *American Journal of Epidemiology*, **126** (2), 161-169.
- Miglioretti, D.L. and Heagerty, P.J. (2007). Marginal modeling of nonnested multilevel data using standard software. *American Journal of Epidemiology*, **165**, 453-463.
- Nelson, J.C. and Pepe, M.S. (2000). Statistical description of interrater variability in ordinal ratings. *Statistical Methods in Medical Research*, **9**, 475-496.
- Qu, Y.S., Piedmonte, M.R., and Medendorp, S.V. (1995). Latent variable models for clustered ordinal data. *Biometrics*, **51**, 268-275.
- Shrout, P.E. and Fleiss, J.L. Intraclass Correlations: Uses in Assessing Rater Reliability. (1979). *Psychological Bulletin*, **2**, 420-428.
- Tanner, M.A. and Young, M.A. (1985). Modeling agreement among raters. *Journal of the American Statistical Association*, **80** (389), 175-180.

- Thompson, J.R. (2001). Estimating equations for kappa statistics. *Statistics in Medicine*, **20**, 2895-2906.
- Uebersax, J.S. and Grove, W.M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine*, **9**, 559-572.
- Williamson, J.M. and Manatunga, A.K. (1997). Assessing interrater agreement from dependent data. *Biometrics*, **54**, 707-714.
- Yerushalmy, J. (1956). The importance of observer error in the interpretation of photofluorograms and the value of multiple readings. *Radiology and Mass Radiography*, , 110-124.

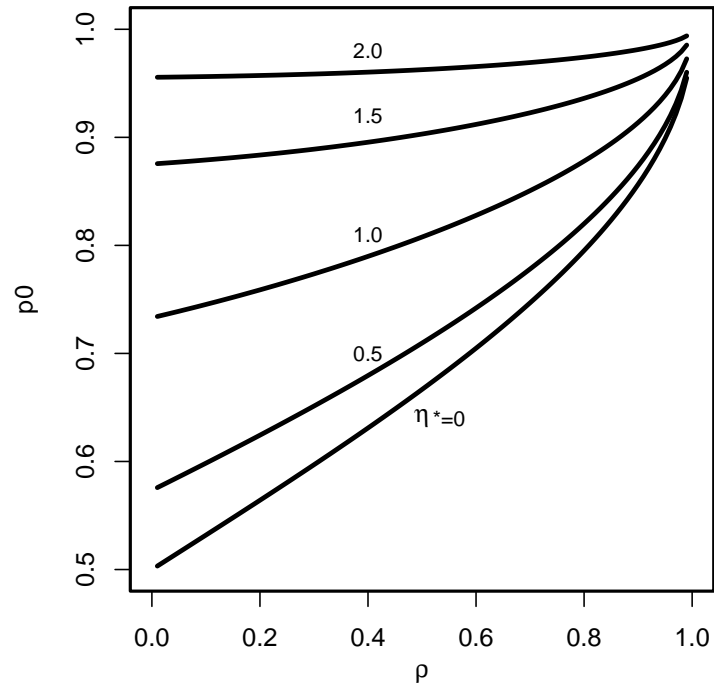


Figure 1: Plot of observed agreement rate  $p_0$  against  $\rho$  for different values of the prevalence term  $\eta$ . Note that  $\rho = \sigma_u^2 / (\sigma_u^2 + \sigma_v^2 + 1)$ .

Table 1: Interpretation of the proposed model-based kappa  $\kappa_m$  and Cohen's kappa  $\kappa$ .

Kappa $\kappa_m$	Interpretation
0 – 0.19	poor agreement
0.20 – 0.39	fair agreement
0.40 – 0.59	moderate agreement
0.60 – 0.79	substantial agreement
0.80 – 1.00	almost-to perfect agreement

Table 2: Simulation results for the probit generalized linear mixed model and model-based kappa statistic  $\kappa_m$  based upon 100 datasets for  $I = 50$  and  $J = 50$  using two different sets of parameter values  $\theta = (\eta, \beta, \sigma_{u_0}^2, \sigma_{v_0}^2, \sigma_{v_1}^2, \rho_{v_{01}})$ . The three models fitted are: (a)  $\Phi^{-1}(p_{ij}) = \eta + u_i + v_j$  (b)  $\Phi^{-1}(p_{ij}) = \eta + \beta x_i + u_i + v_j$  and (c)  $\Phi^{-1}(p_{ij}) = \eta + \beta x_i + u_{0i} + v_{0j} + d_{ij}v_{1j}$ ;  $x_i \sim \text{Bin}(1, 0.5)$ ;  $d_{1i} \sim \text{Bin}(1, 0.5)$ . Cohen's kappa =  $\kappa$  and model-based kappa =  $\kappa_m$ . Mean parameter estimates are presented with associated standard errors in parentheses.

(i) $\eta = 1, \sigma_{u_0}^2 = \sigma_{v_0}^2 = \sigma_{v_1}^2 = 1$				
Parameter	True value	Model (a)	Model (b)	Model (c)
$\eta$	1	0.949 (0.1654)	0.994 (0.240)	1.0160 (0.3640)
$\beta$	0.5	—	0.452 (0.274)	0.4938 (0.4361)
$\sigma_{u_0}^2$	1	0.965 (0.2002)	0.977 (0.240)	0.9465 (0.1866)
$\sigma_{v_0}^2$	1	1.015 (0.2286)	1.061 (0.303)	0.954 (0.3326)
$\sigma_{v_1}^2$	1	—	—	1.343 (0.4786)
$\rho_{v_{01}}$	0.25	—	—	-0.0004 (0.0007)
$\kappa$		0.1906(0.0357)		
$\kappa_m$		0.2091(0.0315)		
(i) $\eta = 1, \sigma_{u_0}^2 = \sigma_{v_0}^2 = \sigma_{v_1}^2 = 5$				
Parameter	True value	Model (a)	Model (b)	Model (c)
$\eta$	1	0.894 (0.2927)	0.714 (0.481)	1.0491 (0.5027)
$\beta$	0.5	—	0.342 (0.650)	0.4094 (0.6991)
$\sigma_{u_0}^2$	5	4.897 (0.843)	4.651 (0.975)	4.6576 (0.9168)
$\sigma_{v_0}^2$	5	5.260 (1.262)	5.074 (1.344)	4.276 (0.8240)
$\sigma_{v_1}^2$	5	—	—	6.6359 (2.2393)
$\rho_{v_{01}}$	0.25	—	—	0.0004 (0.0046)
$\kappa$		0.2814(0.0497)		
$\kappa_m$		0.2910(0.0417)		

Table 3: Simulation results for the probit generalized linear mixed model and model-based kappa statistic  $\kappa_M$  based upon 100 datasets for  $I = 50$  and  $J = 50$  using two different sets of parameter values  $\theta = (\eta, \beta, \sigma_{u_0}^2, \sigma_{v_0}^2, \sigma_{v_1}^2, \rho_{v_{01}})$ . The three models fitted are: (a)  $\Phi^{-1}(p_{ij}) = \eta + u_i + v_j$  (b)  $\Phi^{-1}(p_{ij}) = \eta + \beta x_i + u_i + v_j$  and (c)  $\Phi^{-1}(p_{ij}) = \eta + \beta x_i + u_{0i} + v_{0j} + d_{ij} v_{1j}$ ;  $x_i \sim \text{Bin}(1, 0.5)$ ;  $d_{1i} \sim \text{Bin}(1, 0.5)$ . Mean parameter estimates are presented with associated standard errors in parentheses.

(ii) $\eta = 3, \sigma_{u_0}^2 = \sigma_{v_0}^2 = \sigma_{v_1}^2 = 1$				
Parameter	True value	Model (a)	Model (b)	Model (c)
$\eta$	3	2.950 (0.2488)	3.0348 (0.3854)	3.047 (0.4236)
$\beta$	0.5	–	0.5005 (0.4202)	0.411 (0.1144)
$\sigma_{u_0}^2$	1	1.000 (0.3713)	1.0348 (0.2994)	0.733 (0.2842)
$\sigma_{v_0}^2$	1	1.085 (0.3814)	1.1589 (0.3619)	1.246 (0.5001)
$\sigma_{v_1}^2$	1	–	–	1.275 (0.9846)
$\rho_{v_{01}}$	0.25	–	–	–0.005 (0.0013)
$\kappa$		0.0887(0.0392)		
$\kappa_m$		0.2073(0.0580)		
(ii) $\eta = 3, \sigma_{u_0}^2 = \sigma_{v_0}^2 = \sigma_{v_1}^2 = 5$				
Parameter	True value	Model (a)	Model (b)	Model (c)
$\eta$	3	2.925 (0.330)	2.8962 (0.4988)	2.7937 (0.2828)
$\beta$	0.5	–	0.4696 (0.5947)	0.5467 (0.2913)
$\sigma_{u_0}^2$	5	4.999 (1.1238)	4.4429 (1.0353)	3.9862 (0.5649)
$\sigma_{v_0}^2$	5	5.387 (1.5137)	4.3796 (0.5757)	4.4650 (1.7706)
$\sigma_{v_1}^2$	5	–	–	5.3444 (3.5664)
$\rho_{v_{01}}$	0.25	–	–	0.0020 (0.0095)
$\kappa$		0.2414(0.0584)		
$\kappa_m$		0.2905(0.0428)		

Table 4: 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 (Beam et al 2003).

		Physician B		
Category		Non-diseased	Diseased	Total
Physician A	Non-diseased	460951	64531	525482
	Diseased	74467	192739	267206
Total		535418	257270	792688

Table 5: Results for the breast cancer dataset (Beam 2003). The three models fitted are: (a)  $\Phi^{-1}(p_{ij}) = \eta + u_i + v_j$  (b)  $\Phi^{-1}(p_{ij}) = \eta + \beta x_i + u_i + v_j$  and (c)  $\Phi^{-1}(p_{ij}) = \eta + \beta x_i + u_{0i} + v_{0j} + d_{1j}v_{1j}$ ;  $x_i$  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  $d_{1j} = 0$  for an inexperienced rater, and 1 for an experienced rater. Parameter estimates are presented with standard errors in parentheses.

Parameter	Model (a)	Model (b)	Model (c)
$\eta$	-0.829 (0.1494)	-0.369 (0.1684)	-0.125 (0.0232)
$\beta$	—	-0.802 (0.3928)	-0.368 (0.0316)
$\sigma_{u_0}^2$	3.540 (0.4540)	3.166 (0.4414)	3.453 (0.4001)
$\sigma_{v_0}^2$	0.250 (0.0391)	0.247 (0.0390)	0.248 (0.0346)
$\sigma_{v_1}^2$	—	—	0.244 (0.0034)
$\rho_{v_{01}}$	—	—	0.001 (0.0098)
Cohen's kappa $\kappa$	0.604		
Model-based kappa $\kappa_m$	0.529 (0.0121)		
$\kappa_{m(x_i=0)}$	—	0.5093 (0.0561)	
$\kappa_{m(x_i=1)}$	—	0.5456 (0.0640)	
$\kappa_{m(d_{ij}=0, x_i=1)}$	—	—	0.4993 (0.0565)
$\kappa_{m(d_{ij}=1, x_i=1)}$	—	—	0.5323 (0.0443)