Published in final edited form as:

Stat Methods Med Res. 2018 March ; 27(3): 812–831. doi:10.1177/0962280216643347.

A measure of association for ordered categorical data in population-based studies

Kerrie P Nelson[1] and **Don Edwards**[2]

[1]Department of Biostatistics, Boston University, Boston, USA

[2]Department of Statistics, University of South Carolina, South Carolina, USA

Abstract

Ordinal classification scales are commonly used to define a patient's disease status in screening and diagnostic tests such as mammography. Challenges arise in agreement studies when evaluating the association between many raters' classifications of patients' disease or health status when an ordered categorical scale is used. In this paper, we describe a population-based approach and chance-corrected measure of association to evaluate the strength of relationship between multiple raters' ordinal classifications where any number of raters can be accommodated. In contrast to Shrout and Fleiss' intraclass correlation coefficient, the proposed measure of association is invariant with respect to changes in disease prevalence. We demonstrate how unique characteristics of individual raters can be explored using random effects. Simulation studies are conducted to demonstrate the properties of the proposed method under varying assumptions. The methods are applied to two large-scale agreement studies of breast cancer screening and prostate cancer severity.

Keywords

Agreement; association; crossed random effects; generalized linear mixed model; ordinal classifications; weighted kappa

1 Introduction

Ordered categorical scales are commonly utilized in screening and diagnostic tests such as mammography to assess a patient's disease status or health outcome. Some examples include the Kellgren/Lawrence five-category scale which grades radiographic changes of osteoarthritis,[1] the Dermatology Index of Skin Disease Severity (DIDS) scale used to classify severity of inflammatory skin disease in patients with psoriasis and dermatitis,[2,3] and the Gustilo and Anderson scale for grading severity of open fractures.[4] In the cancer setting, breast cancer status is classified from mammograms using the BI-RADS scale,[5] and

Corresponding author: Kerrie P Nelson, Department of Biostatistics, Boston University, 801 Massachusetts Avenue, Boston, MA 02118, USA. kerrie@bu.edu.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

the severity of prostate cancer using the Gleason grading scale.[6,7] Usually some degree of subjectivity in interpretation is required on the part of the rater, which often leads to substantial inconsistencies between raters' classifications of the same subject, as has been demonstrated in widely used testing procedures including mammography.[8–11] Discrepancies may be due to either a different interpretation of the ordered categorical scale, or if the rater perceives the subject's disease status to be more or less severe than another rater based upon the X-ray, mammogram, or biopsy. These inconsistencies have motivated several large-scale studies to assess accuracy and agreement between raters' classifications and factors that may influence these properties, including cancer,[9–12] rheumatology, and bone fractures.[1,4,13]

Measures of association are frequently reported in inter-rater agreement studies in conjunction with measures of agreement when an ordered categorical scale is used to assess a patient's disease status or health outcome. Both measures provide a useful summary of the strength of relationship between raters' classifications.[14–16] However, due to the ordinal nature of the data and limited availability of statistical methods, it can be challenging to evaluate levels of association and agreement between raters, and especially so when multiple raters (more than two or three) are participating in the study.

Whereas agreement measures focus on quantifying levels of exact agreement between raters, measures of association incorporate additional valuable information about the extent of disagreement between raters' classifications. For example, two raters' classifications of a patient's test result two or three categories apart implies stronger disagreement than if their classifications were only one category apart. Values of association measures can differ substantially from measures of agreement in the same setting since strong association and yet poor agreement between raters' classifications may occur.[16,17] In this paper, we focus on developing a population-based modeling approach and measure of association that easily incorporate the ordinal classifications of any number of raters (at least three) and patients' test results, where missing classifications can be accommodated.[18] When raters and test results in the study are random samples from their respective populations, inferences can be made regarding the underlying populations under study.

Existing approaches for evaluating levels of association between a single pair of raters' ordinal classifications include Cohen's weighted kappa,[19] a nonparametric rank-invariant approach which treats classifications as ranks,[17,20] an odds ratio smoothing procedure,[21] latent trait, and log-linear models.[22–25]

Methods that can be used to assess association between multiple raters include a modeling approach[3] using generalized estimating equations with a weighted kappa measure formulated in a similar manner to Cohen's kappa.[19] Nelson and Pepe[27] describe an exploratory graphical approach to examine the variability between raters' ordinal ratings. However, these approaches generally do not extend easily to assessing association between more than a few raters, where Gonin et al.'s approach includes a fixed term for each rater with increasing complexity as the number of raters grows larger. Shrout and Fleiss's intraclass correlation coefficient (ICC)[2,1] is a commonly used summary statistic to assess reliability between multiple raters' ordinal classifications and Mielke et al. discuss Cohen's weighted kappa for multiple raters.[28–30] Extensions of Cohen's kappa statistic tend to be

sensitive to the same flaws as Cohen's original kappa statistic.[31,32] Other researchers have explored Bayesian approaches using generalized linear mixed models (GLMMs) with nested random effects to assess agreement for ordinal classifications and binary classifications.[33–35] Many of these approaches are not easily implemented in standard statistical software packages. Our proposed approach flexibly includes the ordinal ratings of any number of raters without increasing complexity and accommodates missing data. We provide freely available functions in the R software package for the implementation of the methods.[36]

In section 2, the proposed model and framework for assessing association between multiple raters' ordinal classifications are defined. The proposed model-based measure of association is developed in section 3 with a brief overview of existing approaches. Simulation studies are conducted in section 4 to investigate the properties of the proposed approach. Applications to two large-scale medical studies are presented in section 5, and a description of how to assess unique traits of individual raters and test results in an agreement study in this population-based setting is presented in section 6. A brief discussion follows in section 7.

2 An ordinal model of association

2.1 Introduction

We assume that a subject's true disease or health outcome can be modeled as a continuous unobserved latent trait variable.[33,37,38] In our setting, each of J raters independently grades the same sample of I subjects' test results by assigning classifications $Y_{ij} = c$ ($i = 1,\ldots, I$; $j = 1,\ldots, J$; $c = 1,\ldots, C$) according to an ordered categorical scale with C categories based upon their personal assessment of the subject's true underlying continuous disease status W_{ij}. The latent variable W_{ij} can be written in the form of a linear model as $W_{ij} = \beta_0 + u_i + v_j + \varepsilon_{ij}$ with intercept β_0 and a crossed random effects structure with subject random effects u_i, ($i = 1,\ldots, I$) and rater random effects v_j, ($j = 1,\ldots, J$), assumed mutually independent with $\mathrm{N}(0, \sigma_u^2)$ and $\mathrm{N}(0, \sigma_v^2)$ distributions respectively, and errors ε_{ij} distributed as N(0, σ^2). The classifications $Y_{ij} = c$ are equivalent to $\alpha_{c-1} \leq W_{ij} \leq \alpha_c$ where the set of strictly monotonically increasing thresholds $\alpha_0,\ldots,\alpha_C$ divides the underlying continuous latent variables W_{ij} into $C + 1$ intervals with $\alpha_0 = -\infty$ and $\alpha_C = +\infty$.[37]

The ordinal GLMM provides an ideal framework for modeling the ordinal classifications of multiple raters.[37,39,40] It flexibly incorporates missing data since every rater may not classify every subject in the sample. A crossed random effect structure of raters and subjects' test results appropriately accounts for the dependency between classifications due to all raters grading the same sample of subjects. An issue arises with ordinal data where the absolute location β_0 and scale σ of the latent variable are not identifiable. This is dealt with here wlog by setting $\beta_0 = 0$ and $\sigma = 1$.[37] A variety of link functions can be used as part of the ordinal GLMM framework. In our setting, the probit link function is especially appealing due to the continuous latent disease status assumption underlying the model and for the ease of mathematics and is our choice of link function. It has been previously demonstrated that nearly identical results are obtained when a logistic link is used in the GLMM.[41,42]

The ordinal GLMM with a probit link function models the cumulative probability that a subject's test result is classified into category c or lower ($c = 1,\ldots,C$)

$$\Phi^{-1}\left(Pr\left(Y_{ij} \leq c \middle| u_i, v_j\right)\right) = \alpha_c - (u_i + v_j) \quad (1)$$

This can also be rewritten as the probability of a subject's test result being classified into any particular category $Pr\left(Y_{ij}=c \middle| u_i, v_j\right) = \Phi\left(\alpha_c - (u_i+v_j)\right) - \Phi\left(\alpha_{c-1} - (u_i+v_j)\right)$, where Φ is the cumulative distribution function (cdf) of the standard normal distribution.

Rater random effects v_j ($j = 1,\ldots, J$) account for the uniqueness of each rater's classifications, where a large rater random effects variance component σ_v^2 indicates a more heterogeneous group of raters. Similarly, a large variance component σ_u^2 for the subject random effects u_i ($i = 1,\ldots, I$) suggests a set of test results displaying a broad range of clarity of disease status. In section 6, we show how random effects can be estimated for raters and subjects included in a study, which can provide useful information and feedback for training purposes of individual raters.

To obtain estimates of the parameter vector $\boldsymbol{\theta} = \left(\alpha_1, \ldots, , \alpha_{C-1}, \sigma_u^2, \sigma_v^2\right)$ for the ordinal GLMM in equation (1) we fit the GLMM model using an approximate maximum likelihood approach. The marginal likelihood function takes the form

$$L(\boldsymbol{\theta};\boldsymbol{Y}) = \int_{\mathbf{u},\mathbf{v}} L(\boldsymbol{\theta};\mathbf{u},\mathbf{v},\mathbf{y})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} = \int_{\mathbf{u},\mathbf{v}} f_{Y\left|_{u,v}\right.}(\mathbf{y};\mathbf{u},\mathbf{v})\mathbf{f}_{\mathbf{u}}(\mathbf{u};\sigma_u^2) f_v(\mathbf{v};\sigma_v^2)\mathrm{d}\boldsymbol{u}\mathrm{d}\mathbf{v}$$

$$= \int_{\mathbf{u}}\int_{\mathbf{v}} \left[\prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{c=1}^{C} \left[\Phi(\alpha_c - (u_i+v_j)) - \Phi(\alpha_{c-1} - (u_i+v_j))\right]^{d_{ijc}} \right] \left[\prod_{i=1}^{I} \frac{1}{\sqrt{2\pi\sigma_u^2}} e^{\frac{-u_i^2}{2\sigma_u^2}} \right] \left[\prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{\frac{-v_j^2}{2\sigma_v^2}} \right] \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v}$$

with indicator function $d_{ijc} = 1$ if $y_{ij} = c$ and 0 otherwise. Due to the high dimensionality of the crossed random effects, no closed-form solution for maximizing the likelihood function is available. However, multivariate Laplacian approximation provides an attractive and viable solution to obtaining approximate maximum likelihood estimates $\hat{\boldsymbol{\theta}}$.[43] Large-sample approximate standard errors are estimated by taking the square-roots of the diagonals of matrix H at convergence $se(\hat{\boldsymbol{\theta}}) = \sqrt{\mathrm{diag}\left[-\left\{\mathbf{H}(\hat{\boldsymbol{\theta}})\right\}^{-1}\right]}$, where $\boldsymbol{H} = \frac{\partial^2 l(\boldsymbol{\theta};\mathbf{u},\mathbf{v},\mathbf{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^t}$ is the second-order derivative of the log-likelihood function $l(\boldsymbol{\theta}; \mathbf{u}, \mathbf{v}, \mathbf{y})$ evaluated at the approximate maximum likelihood estimates of $\boldsymbol{\theta}$ and is generated during the model-fitting process. This fitting approach is available in the *ordinal* package in R for fitting ordinal GLMM models with crossed random effects, one of the few statistical software packages currently able to do so. Also explored were adaptive quadrature methods for fitting the ordinal GLMM model, but these were not feasible for our ordinal GLMM model due to the large number of random effects.[38,44,45] Simulation studies presented in section 4 demonstrate that reasonably unbiased estimates are obtained using the *ordinal* package under a wide range of varying conditions.

In the following section, we develop a chance-corrected model-based measure of association which is based upon the ordinal GLMM parameters $\boldsymbol{\theta}=\left(\alpha_1 \ldots,, \alpha_{C-1}, \sigma_u^2, \sigma_v^2\right), \alpha_0=-\infty$ and $\alpha_C=+\infty$ in equation (1).

3 A measure of association in the population-based setting

Measures of association are a popular choice for comparing raters' ordinal classifications, incorporating information about agreement and disagreement into a comprehensive summary measure. While measures of exact agreement can be used for both nominal (unordered) and ordinal classifications, measures of association are appropriate only for classifications based upon an ordered categorical scale. Here, we develop a chance-corrected measure of association based upon the ordinal GLMM in equation (1). We first define two probabilities which are instrumental in the development of a chance-corrected measure of association—observed and chance association.

3.1 Observed and chance association

Observed association, p_{0a} is the proportion of time raters j and j' $(j \neq j')$ classify the same patient's test result into the rth and sth categories respectively ($r, s = 1,\ldots, C$), weighted by how many categories apart they are. While any weighting scheme can be applied, two conventional choices are: linear (absolute error) weights $w_{rs}=1-|r-s|/(C-1)$; and quadratic (squared-error) weights $w_{rs}=1-(r-s)^2/(C-1)^2$ for pairs of classifications in the rth and sth categories respectively ($r, s = 1,\ldots, C$) by two raters j and j' $(j \neq j')$. Based upon the ordinal GLMM in the population-based setting p_{0a} takes the form (derivation in Appendix 1)

$$\begin{aligned} p_{0a} &= \sum_{r=1}^{C}\sum_{s=1}^{C} w_{rs}\left[Pr(Y_{ij}=r \cap Y_{ij'}=s)\right] \qquad \text{in the general population} - \text{based setting} \\ &= \int_{-\infty}^{+\infty} \sum_{r=1}^{C}\sum_{s=1}^{C} w_{rs}\left[\Phi\left(\frac{\alpha_r^*-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)-\Phi\left(\frac{\alpha_{r-1}^*-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\right]\times\left[\Phi\left(\frac{\alpha_s^*-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)-\Phi\left(\frac{\alpha_{s-1}^*-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\right] \\ &\qquad \times\phi(z)\,\mathrm{d}z \quad 0.5 \leq p_{0a} \leq 1, \end{aligned} \tag{2}$$

where $\alpha_c^*=\alpha_c/\sqrt{\sigma_u^2+\sigma_v^2+1}$ and $\rho=\sigma_u^2/(\sigma_u^2+\sigma_v^2+1)$, which itself is a natural measure comparing the variability amongst subjects' test results, σ_u^2, relative to the overall variability present between classifications. Large variability between test results relative to the variability between the raters will yield a value of ρ close to 1. In this population-based setting over many raters, observed association p_{0a} takes values between 0.5 and 1 (proof in Appendix 3).

Chance association p_{ca} is the proportion of time rater j classifies subject i into the rth category and rater j' $(j \neq j')$ classifies subject i' $(i \neq i')$ into the sth category ($r, s = 1,\ldots, C$) simply due to coincidence, weighted according to how many categories apart the ratings are. For the ordinal GLMM p_{ca} takes the form (derivation in Appendix 2)

$$p_{ca}=\sum_{r=1}^{C}\sum_{s=1}^{C}w_{rs}[\Pr(Y_{ij}=r)\times\Pr(Y_{i'j'}=s)]$$
$$=\sum_{r=1}^{C}\sum_{s=1}^{C}w_{rs}[\Phi(\alpha_r^*)-\Phi(\alpha_{r-1}^*)]\times[\Phi(\alpha_s^*)-\Phi(\alpha_{s-1}^*)],\ 0.5\le p_{ca}\le 1 \quad (3)$$

It can be shown in this population-based setting that $p_{oa} \geq p_{ca} \geq 0{:}5$ (see Appendix 3 for proof). Estimates $\hat{p}_{0a}$ and $\hat{p}_{ca}$ can be obtained from fitting the corresponding ordinal GLMM in equation (1) as outlined in section 2.

3.2 A proposed population-based measure of association

The proposed model-based measure of association κ_{ma} is a linear function of observed association p_{0a} in equation (2). Two adjustments that we make to this linear function to derive the proposed measure κ_{ma} are to minimize the effects of chance association on κ_{ma} so that the measure is chance-corrected, and to ensure that κ_{ma} is scaled to take values between 0 and 1 so that it is easily interpretable in a similar manner to Cohen's weighted kappa statistic.[19] First, we minimize the effects of chance association on κ_{ma} by finding the values of the fixed threshold terms $\left(\alpha_1^*,\ldots,\alpha_{C-1}^*\right)$ with $\alpha_0^*=-\infty$ and $\alpha_C^*=+\infty$ which minimize the expression for chance association in equation (3). The expression for p_{ca} takes a minimum value of 0.5 when the monontonically increasing threshold values denoted as $(\alpha_{\min,1}^*,\alpha_{\min,2}^*,\ldots,\alpha_{\min,C-1}^*)$ take the values (0.00001, 0.000002,…) (see Appendix 3). We then substitute these threshold values into the expression for κ_{ma}. Finally, we scale κ_{ma} to lie between 0 and 1 for similar interpretability to Cohen's weighted kappa statistic. Since $0.5 \leq p_{0a} \leq 1$, multiplying the expression for κ_{ma} by 2 and subtracting 1 scales κ_{ma} so that $0 \leq \kappa_{ma} \leq 1$. The form of κ_{ma} (4) is thus

$$\kappa_{ma}=2*p_{0a}\left(\left(-\infty,\alpha_{\min,1}^*,\ldots,,\alpha_{\min,C-1}^*,+\infty\right),\rho\right)-1$$
$$=2*\int_{-\infty}^{+\infty}\sum_{r=1}^{C}\sum_{s=1}^{C}w_{rs}\left[\Phi\left(\frac{\alpha_{\min,r}^*-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)-\Phi\left(\frac{\alpha_{\min,r-1}^*-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\right]\times\left[\Phi\left(\frac{\alpha_{\min,s}^*-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)-\Phi\left(\frac{\alpha_{\min,s-1}^*-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\right]$$
$$\times\phi(z)\mathrm{d}z-1 \qquad 0\le\kappa_{ma}\le 1 \quad (4)$$

A value of κ_{ma} near 0 indicates poor chance-corrected association between raters, while a value closer to 1 suggests very strong chance-corrected association between raters. We demonstrate in section 4 that, in contrast to Cohen's weighted kappa, κ_{ma} is unaffected by changes in the underlying disease prevalence.

Estimation of κ_{ma} for a dataset involves first fitting the GLMM in equation (1) and obtaining estimates $\hat{\sigma}_u^2$ and $\hat{\sigma}_v^2$. These values are used to calculate the coefficient $\hat{\rho}=\hat{\sigma}_u^2/(\hat{\sigma}_u^2+\hat{\sigma}_v^2+1)$ which in turn is incorporated into the estimate $\hat{\kappa}_{ma}$.

The variance var $(\hat{\kappa}_{ma})$ is derived using the multivariate delta method as a function of the rater and subject random effects variance components (assumed independent) where $var(\hat{\sigma}_u^2)=2(\sigma_u^2)^2/I$ and $var(\hat{\sigma}_v^2)=2(\sigma_v^2)^2/J$ for large *I* and *J*. The variance of $\hat{\rho}$ is calculated as

$$var(\hat{\rho})=\frac{2(\sigma_u^2)^2(\sigma_v^2+1)^2}{I(\sigma_u^2+\sigma_v^2+1)^4}+\frac{2(\sigma_v^2)^2(\sigma_u^2)^2}{J(\sigma_u^2+\sigma_v^2+1)^4}$$

Since κ_{ma} is a function of ρ, the delta method is again applied

$$\begin{aligned}var(\hat{\kappa}_{ma})=4\times var(\hat{\rho})\times\Bigg[\int_{-\infty}^{+\infty}\sum_{r=1}^{C}\sum_{s=1}^{C}w_{rs}\times\Bigg(\Bigg[\Phi\left(\frac{\alpha^*_{\min,r}-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)-\Phi\left(\frac{\alpha^*_{\min,r-1}-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\Bigg]\\ \times\Bigg[\phi\left(\frac{\alpha^*_{\min,s}-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\left(\frac{-z}{2\sqrt{\rho(1-\rho)}}+\frac{\alpha^*_{\min,s}-z\sqrt{\rho}}{2(1-\rho)^{3/2}}\right)-\phi\left(\frac{\alpha^*_{\min,s-1}-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\\ \times\left(\frac{-z}{2\sqrt{\rho(1-\rho)}}+\frac{\alpha^*_{\min,s-1}-z\sqrt{\rho}}{2(1-\rho)^{3/2}}\right)\Bigg]+\Bigg[\Phi\left(\frac{\alpha^*_{\min,s}-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)-\Phi\left(\frac{\alpha^*_{\min,s-1}-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\\ \times\Bigg[\phi\left(\frac{\alpha^*_{\min,r}-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\left(\frac{-z}{2\sqrt{\rho(1-\rho)}}+\frac{\alpha^*_{\min,r}-z\sqrt{\rho}}{2(1-\rho)^{3/2}}\right)-\phi\left(\frac{\alpha^*_{\min,r-1}-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\\ \times\left(\frac{-z}{2\sqrt{\rho(1-\rho)}}+\frac{\alpha^*_{\min,r-1}-z\sqrt{\rho}}{2(1-\rho)^{3/2}}\right)\Bigg]\Bigg)\phi(z)\mathrm{d}z\Bigg]^2\end{aligned}$$

For practical purposes, functions in R to fit the ordinal GLMM in equation (1) and to estimate κ_{ma} and var $(\hat{\kappa}_{ma})$ for an inter-rater agreement dataset are available from the first author and in supplemental material on the journal's website. Figure 1 demonstrates the effects of the rater random effect variance σ_v^2 on the proposed measure of association κ_{ma} as the subject random effects variance σ_u^2 increases. The strongest association is observed for small values of the rater variance σ_v^2. The association measure κ_{ma} increases with σ_u^2; this is due to more clearly defined disease status observed in a more heterogeneous group of subject test results.

3.3 Cohen's weighted kappa with model-based parameters

Cohen's weighted kappa statistic is a chance-corrected statistic for assessing association between two raters, based upon the observed weighted proportion of pairs in agreement and chance weighted proportion of pairs in agreement expected under a statistical model of independence.[16,19] Here, we generate a population-based Cohen's weighted kappa statistic for multiple raters incorporating ordinal GLMM probabilities of observed and chance association p_{0a} and p_{ca} (defined in section 3.1) for comparison with our proposed measure of association κ_{ma} (4). This statistic will be referred to as $\kappa_{GLMM,a}$ and takes the following form, with choice of weights described in section 3.1

$$\kappa_{\mathrm{GLMM},a}=\frac{p_{0a}-p_{ca}}{1-p_{ca}}\ \text{estimated as}\ \hat{\kappa}_{\mathrm{GLMM},a}=\frac{\hat{p}_{0a}-\hat{p}_{ca}}{1-\hat{p}_{ca}}\ \text{with}\ 0\leq\kappa_{\mathrm{GLMM},a}\leq 1$$

3.4 Shrout and Fleiss' ICC[2,1]

Shrout and Fleiss' [2,1] statistic is derived from a two-way ANOVA model and is commonly used as a measure of association to assess reliability between raters.[28] While six forms of the ICC are described in their paper, the ICC[2,1] is an appropriate statistic when all subjects are graded by the same set of raters who are assumed to be a random subset of all possible raters. This statistic has been demonstrated to be equivalent to Cohen's weighted kappa with quadratic weights when comparing two raters' classifications.[46]

4 Simulation studies

Simulation studies were conducted under a varying range of scenarios as presented in Table 1 to investigate the behavior of the proposed measure of association κ_{ma} and the parameters of the ordinal GLMM in equation (1). Simulation scenarios included increasing rater and subject-level random effect variances and numbers of raters and items (sample size) and assessing their impact on estimation of the parameter vector $\boldsymbol{\theta} = \left(\alpha_1, \ldots, , \alpha_{C-1}, \sigma_u^2, \sigma_v^2\right)$ and κ_{ma} when estimated using the *ordinal* package in R. Effects of non-normally distributed random effects parameter estimation were also explored.

Sets of 1000 simulated datasets based upon the ordinal GLMM in equation (1) were generated for each simulation scenario in the following manner. A parameter vector containing true values $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \sigma_u^2, \sigma_v^2)$ for $C = 5$ and the number of raters J and subjects I was specified for each set of simulations according to Table 1 (every combination). Random subject effects u_i ($i = 1,\ldots, I$) and rater effects v_j ($j = 1,\ldots, J$) were generated using R functions *rnorm, rexp*, and *runif* depending on the scenario and centered and scaled after choosing parameters λ and a to achieve the specified σ_u^2 and σ_v^2. A sample of $n = IJ$ ordinal classifications $Y_{ij} = c$ ($c = 1,\ldots, C$) was then randomly generated from a multinomial distribution using the R function *rmultinom* according to the probability mass function

$$\prod_{c=1}^{C} [\Pr(Y_{ij}=c|u_i, v_j]^{d_{ijc}} = \prod_{c=1}^{C} [\Phi(\alpha_c - (u_i+v_j)) - \Phi(\alpha_{c-1} - (u_i+v_j))]^{d_{ijc}}$$

where $d_{ijc} = 1$ if $Y_{ij} = c$ and 0 otherwise. The *clmm* function in the *ordinal* package in R was used to fit the ordinal GLMM (1) and obtain parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\sigma}_u^2, \hat{\sigma}_v^2)$ with their estimated standard errors for each simulated dataset.

Table 2((a) and (b)) displays simulation results for the proposed chance-corrected measure of association, κ_{ma}. The true value of κ_{ma} is presented for each simulation scenario along with the mean of the estimates $\hat{\kappa}_{ma}$ from the one thousand simulated datasets. The standard error (S.E.) is presented as the average of the one thousand standard error estimates $se(\hat{\kappa}_{ma})$. Slight bias is observed in the estimation of the proposed association measure κ_{ma} for the smaller sample size ($I = 100$, $J = 10$) especially when the rater random effect variance σ_v^2 is large. This bias diminished at the larger sample size ($I = 250$, $J = 100$). Slightly increased

levels of bias in the estimation of κ_{ma} were observed in both small and large sample sizes when the random effects were not normally distributed, though corresponding standard errors were similar to those for normally distributed random effects. These results indicate the proposed association measure is estimated in a reasonably unbiased manner for large and smaller sample sizes, varying random effects variances, and certain departures from the distributional assumptions such as normality of the random effects' distributions.

Tables 3 and 4 present results for a selected range of the simulation studies, with further sets of simulations (Tables 3(c) and 3(d)) presented in the Supplemental Material online. Parameter estimates $\hat{\boldsymbol{\theta}}=(\hat{\alpha}_1,\hat{\alpha}_2,\hat{\alpha}_3,\hat{\alpha}_4,\hat{\sigma}_u^2,\hat{\sigma}_v^2)$, a model-based measure of agreement $\hat{\kappa}_{ma}$ and proposed measure of association $\hat{\kappa}_{ma}$ are presented in these tables.[42] The standard deviation of the observed 1000 estimates is presented for each parameter, with standard error reported as the mean of the 1000 standard error estimates.

Results demonstrate that variance components $\sigma_u^2$and σ_v^2 were estimated with little or no bias at small values $\left(\sigma_u^2 \text{ and } \sigma_v^2=1 \text{ or } 5\right)$ for small and large sample sizes (*I, J*) and for both normally and non-normally distributed random effects. Low to moderate levels of bias in the estimation of σ_u^2 and σ_v^2 was observed for larger values of the rater variance component $\left(\sigma_v^2=10 \text{ or } 20\right)$, especially for smaller sample sizes (I = 100, J = 10) and for non-normal random effects. For larger sample sizes (I = 250, J = 100) only minimal bias in the estimates of σ_u^2 and σ_v^2 was observed when the random effects were normally distributed; however, low to moderate bias remained in the estimates of σ_u^2 and σ_v^2 when non-normally distributed random effects were included.

Thresholds $\alpha_1,\ldots,\alpha_{C-1}$ $(\alpha_0=-\infty,\alpha_C=+\infty)$ were generally estimated with no or minimal bias. When the variability between raters' classifications was large $\left(\sigma_v^2=10 \text{ or } 20\right)$, some bias was noted. Some slight bias was observed in the estimates of $\alpha_1,\ldots,\alpha_{C-1}$ for the simulation studies with non-normal random effects at smaller sample sizes, which receded at the larger sample size.

Coefficient ρ was estimated with slight bias at smaller sample sizes with the bias was more evident when the rater variability was large $\left(\sigma_v^2=10 \text{ or } 20\right)$. This is likely due to ρ being a function of σ_u^2 and σ_v^2 which also exhibited moderate bias for large rater variability. This bias receded at larger sample sizes for normally distributed random effects.

In summary, the simulation results suggest the parameters of the ordinal GLMM and the proposed measure of association are estimated with very little or no bias using the *ordinal* package in R especially when the rater random effect variance are moderately low $\left(\sigma_v^2=1 \text{ or } 5\right)$, which is common in real-life inter-rater agreement studies. Even at the smaller sample size (I = 100, J = 10), observed biases in parameter estimates were generally small, though some bias was observed in large and small sample sizes for non-normally distributed random effects.

Figures 2(a) and (b) displays the effects of varying disease prevalence on the different summary measures of association, κ_{ma}, ICC[2,1] and Cohen's weighted $\kappa_{GLMM,a}$ with linear weights as described in section 3.1 for an ordinal classification scale with five categories ($C = 5$). Cohen's weighted $\kappa_{GLMM,a}$ with quadratic weights generated very similar estimates to the ICC[2,1] coefficient and is not presented separately here. Table 5 presents the percent of classifications in each of the five categories for the various disease prevalences used in Figure 2(a) and (b). True parameter values were used in the plots with the exception of Shrout and Fleiss' ICC[2,1] statistic, which was averaged over sets of 200 simulated datasets for each value of ρ. In Figure 2(a) and (b), it is seen that as ρ increases from 0 (minimum) to 1 (maximum), each measure of association increases in value and at a more extreme rate as ρ approaches 1. As shown in Figure 2(a), as disease prevalence varies from extremely high or low to being equally distributed over the five categories, κ_{ma} remains unchanged and is thus robust to changes in disease prevalence, while the ICC[2,1] statistic fluctuates in value with varying disease prevalence, and increases especially when disease prevalence is extreme. Figure 2(b) demonstrates that Cohen's linear-weighted population-based GLMM measure $\kappa_{GLMM,a}$ is also sensitive to changes in prevalence.

5 Applications to real-life studies

5.1 Breast cancer screening example

Beam et al.[9] recently conducted a large-scale study to investigate factors that potentially may influence accuracy in radiologists' interpretation of screening mammograms. A random sample of 104 radiologists independently classified screening mammograms of 148 women randomly selected via stratified sampling using a modified ordinal BI-RADS scale with five categories ranging from normal to probably malignant. Forty-three percent of the 148 sets of mammograms were from women with breast cancer. We examine levels of association and agreement between the radiologists in this dataset using our proposed methods and compare these with existing measures of association and agreement. Our population-based approach allows conclusions to be drawn regarding association between typical radiologists who interpet screening mammograms since the radiologists and patients were randomly selected from their respective populations. Table 6 presents a sample of the classifications made by individual radiologists.

The ordinal GLMM in equation (1) was fitted to the dataset consisting of $n = 15{,}392$ ($IJ = 104 \times 148$) classifications using the *clmm* function in the *ordinal* package in R. The procedure took less than 2 min to run. Parameter estimates are presented in Table 7. Based upon the dataset, estimated model-based probabilities of being classified into each of the five ordered categories were 35% (normal), 15% (benign), 17% (probably benign), 22% (possibly malignant), and 11% (probably malignant).

Estimates of the various measures of association for the Beam mammogram study are presented in Table 6 including the proposed measure of association κ_{ma}, Shrout and Fleiss' ICC[2,1] statistic, and Cohen's GLMM-based weighted kappa with quadratic weights, $\kappa_{GLMM,a}$ (section 3.3). Quadratic weights were also used for κ_{ma}. Commonly used measures of (exact) agreement are also presented, including Fleiss' kappa κ_F and Light and Conger's kappa κ_{LC}, both adaptations of Cohen's original kappa statistic, and an ordinal GLMM

model-based agreement measure κ_m.[42,47–49] Model-based observed agreement is estimated as $\hat{p}_0=0.430$, indicating low to moderate observed agreement between typical pairs of radiologists in this setting. In contrast, observed association is very strong at $\hat{p}_{0a}=0.907$, suggesting that while pairs of radiologists may not often provide identical classifications to the same patient's mammogram, raters' classifications may typically disagree by only one category on the five-category ordered categorical scale, rather than by several categories.

The proposed measure of association was estimated as $\hat{\kappa}_{ma}=0.475\ (se=0.022)$ indicating moderate chance-corrected association between radiologists who typically interpret screening mammograms, based upon the table in Landis and Koch.[50] Our proposed approach provides a chance-corrected measure of association for the study that is not affected by disease prevalence. The Shrout and Fleiss' ICC[2,1] coefficient was estimated at 0.652 (95% *c.i.* = (0.601, 0.706)) suggesting moderately strong heterogeneity between subjects' mammograms relative to the variability between raters' classifications. Cohen's GLMM-based weighted kappa with quadratic weights was estimated at $\hat{\kappa}_{\mathrm{GLMM},a}=0.611$, also a higher value than $\hat{\kappa}_{ma}$ likely due to a prevalence effect as depicted in Figure 2(a).

Each of the estimated measures of (exact) agreement indicated only low levels of chance-corrected agreement between raters, including the model-based measure of (exact) agreement $\hat{\kappa}_m=0.241\ (se=0.015)$ Fleiss' kappa $\hat{\kappa}_F=0.297$, Conger's and Light's kappa $\hat{\kappa}_{LC}=0.298$, and Cohen's GLMM-based (unweighted) kappa $\hat{\kappa}=0.257$.

Overall, there appears to be substantial discrepancies between raters' ordinal classifications for grading screening mammograms in this population, reflected in the low and moderate levels of chance-corrected agreement and association.

5.2 Gleason grading study for prostate cancer

Allsbrook et al.[51] reported on a study conducted to examine agreement and association between 41 general pathologists each classifying the same sample of 38 biopsy slides for the severity of prostate cancer. They utilized a modified earlier version of the Gleason grading scale consisting of four categories defined as: category (i) Gleason scores 2–4 (mild disease); category (ii) Gleason scores 5–6; category (iii) Gleason score 7; category (iv) Gleason scores 8–9 (severe disease). However, there were two missing observations, and since the proposed approach accommodates missing data, this did not lead to any further issues. A sample of this dataset is presented in the supplemental material online.

To assess the association in a unified approach between the ordinal classifications of the 41 raters, the ordinal GLMM with a crossed random effects structure in equation (1) was fit to the dataset using the *clmm* function in the *ordinal* package in R. The resulting parameter estimates and summary measures are presented in Table 7.

Based upon this sample of 38 slides, the probabilities of being classified into the four categories (from mild to severe disease) according to the GLMM model were 17%, 30%, 21%, and 32% respectively. Observed association between the raters was estimated to be very strong at $\hat{p}_{0a}=0.917$. The chance-corrected measure of association κ_{ma} with quadratic

weights was estimated as $\hat{\kappa}_{ma}=0.554\ (s.e.=0.043)$ indicating moderate levels of chance-corrected association between the 41 general pathologists, where a value of 1 indicates perfect association. In comparison, Shrout and Fleiss' ICC[2,1] statistic is estimated at 0.734 (95% confidence interval 0.642, 0.824) reflecting the stronger heterogeneity between patients' biopsies slides $(\hat{\sigma}_u^2=4.805)$ relative to the variability observed between raters' classifications $(\hat{\sigma}_v^2=0.480)$. Cohen's GLMM-based weighted kappa with quadratic weights was also estimated at a high value as $\hat{\kappa}_{\mathrm{GLMM},a}=0.687$. These large values of the ICC[2,1] coefficient and $\kappa_{GLMM,a}$ compared to κ_{ma} in the Gleason grading study may be attributed to the tendency of these two measures to be influenced by the prevalence of disease (high or low), and to take higher values when ρ is large, as in the Gleason Grading study, where $\hat{p}=0.765$ as demonstrated in Figure 2(a). The agreement measures κ_m, κ_F, and κ_{LC} all suggested low (exact) agreement.

Overall, these results indicate that chance-corrected (exact) agreement between the general pathologists is low, but chance-corrected association is moderate, where the proposed measure κ_{ma} does not over-estimate the strength of association between the pathologists as the other measures do due to a high value of ρ. The proposed methods used here allow classifications of all 41 pathologists to be analyzed and interpreted in one unified approach. This approach leads to easily interpretable results, in comparison to studying agreement and association between each pair of pathologists, which leads to many statistics that can be difficult to interpret. Characteristics of each pathologist and the subjects' test slides included in the study can be examined through their estimated random effect terms $\hat{v}_j$ if required, which is described for this Gleason grading study in the next section.

6 Estimation of individual rater and subject traits

The primary focus of population-based agreement studies is usually to draw conclusions about the strength of agreement and association between raters who typically classify patients' test results in the underlying population. It can also be informative to examine unique characteristics of individual raters and subjects included in the study for rater awareness and training purposes which is discussed below.

Unique characteristics of each rater's classifications are adjusted for in the ordinal GLMM in (1) via their random effect term v_j, (j = 1,…, J). In the ordinal package *clmm*, predictions of the rater estimated effects $\hat{v}_j, (j=1,\ldots,J)$ are generated as part of the modeling process as conditional modes, i.e. the modes of the distributions for the random effects given the observed data and estimated model parameters (also known as posterior Bayesian modes) using a Newton–Raphson algorithm.[36] A corresponding measure of uncertainty for each estimated effect, the conditional variance computed from the second order derivatives of the conditional distribution of the random effects, is also generated in the model-fitting process. Figure 3(a) and (b) presents plots of the rater estimated effects (J = 41) and subject estimated effects (I = 38) respectively with 95% confidence intervals using the conditional variance for the Gleason grading agreement study (section 5.2). For example, pathologist 21 has a large negative estimated effect $\hat{v}_{21}=-4.02$ indicating a rater who tends to consistently assign milder disease status to patient biopsy slides relative to the other raters. Rater 34 has a large

positive estimated effect $\hat{v}_{34}=4.12$ signaling a rater who assigns more severe disease categories to patients' slides relative to other raters.

A similar approach can be used to obtain predictions of the subject estimated effects, $\hat{u}_i,\ \ i=1,\ldots,I$. These estimates reflect the heterogeneity observed between the patients' test results. For the Gleason grading study, subject estimated effects ranged from −2.1 to 0.98. Large positive (negative) subject effects indicate test results that clearly show disease (no disease), while values close to 0 suggest a test result whose disease status is less obvious.

7 Discussion

The use of ordered classification scales is widespread in medical tests and diagnostic procedures to grade a patient's disease or health status.[8–11] However, while strong reliability between raters is an important attribute of an accurate diagnostic procedure, poor agreement and association between raters have been reported in many of these settings.[1,4,9–12] Furthermore, in agreement studies with multiple raters using an ordinal classification scale to classify patients' test results, it is challenging to assess levels of association and agreement in a unified approach since few statistical approaches are available or easy to implement.

In this paper, we have described a model-based approach to assess association between any number of raters (at least three) classifying subjects' test results according to an ordered categorical scale. Missing data can be accommodated where some raters may not classify every subject in the sample. Many agreement studies report inter-rater reliability between two raters at a time which usually leads to several summary statistics with complexities in interpretation.[52,53] The proposed model-based approach describes association between the group of raters in a unified and comprehensive approach with a single summary measure, lending itself to increasing power and efficiency and simpler interpretation. An important advantage is that the chance-corrected measure of association is not affected by the underlying disease prevalence, in contrast to other existing measures including Shrout's and Fleiss' ICC[2,1] statistic. Results can be generalized to the raters and subjects who typically undertake these procedures and tests when the study participants and raters are randomly sampled from their respective populations. The proposed approach can be fit efficiently to an agreement dataset using author-written functions in the freely available R software package, making it a viable and attractive approach to implement in practice.

Simulation studies demonstrated that estimation of the proposed measure of association appears fairly robust to varying sample sizes of raters and subjects, large and small variance components which measure the variability between the groups of raters and subjects, and non-normally distributed random effect distributions of the raters and subjects under a varying range of situations encountered in real-life studies. Some bias was noted in the estimation of $\boldsymbol{\kappa}_{ma}$ when there was extreme variability between raters. Further work is required to fully explore the impact of non-normal random effects on the estimation of the measure of association, and to assess the effects of rater and subject characteristics on agreement by incorporating covariates into the ordinal GLMM model.

Measures of association are often preferred over measures of agreement when assessing strength of relationship between ordinal classifications since they incorporate information about the extent of disagreement in addition to exact agreement. For parametric approaches such as Cohen's weighted kappa and our proposed measure of association, less "credit" is assigned in the kappa statistic to pairs of raters' classifications that disagree more and are further apart on the categorical scale by use of a weighting scheme.[16] While the choice of weights is left up to the researcher, quadratic and linear weights are common options, and use of these schemes makes for easier comparability between studies. The ICC has been shown to be equivalent to Cohen's weighted kappa statistic for pairs of raters' ordinal classifications.[46,54] Our simulation studies indicated that a modified version of Cohen's weighted kappa with quadratic weights using population measures of chance and observed association yielded similar values to Shrout and Fleiss' ICC[2,1] statistic, and was sensitive to the disease prevalence in a similar manner to Cohen's original kappa.[31]

Liu and Agresti[38] note that when the ordinal classifications are assumed to be based upon an underlying unobserved latent variable, such as disease status, the effects are invariant to the number of categories and thresholds of the categorical scale used, and that different studies employing different scales should lead to similar conclusions.

Acknowledgments

We thank Dr Allsbrook and Professor Beam for kindly providing us with their datasets. We also thank Aya Mitani for her assistance and Rune Haubo Christensen for his help in using the ordinal package in R.

Funding

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

Appendix 1 Derivation of observed association

$$p_{0a}=\sum_{r=1}^{C}\sum_{s=1}^{C} w_{rs}[Pr(Y_{ij}=r \cap Y_{ij'}=s)] \quad \text{in the general population} - \text{based setting}$$

$$=\int_{-\infty}^{+\infty}\sum_{r=1}^{C}\sum_{s=1}^{C} w_{rs}[Pr(Q \le \alpha_r - u_i - v_j) - Pr(Q \le \alpha_{r-1} - u_i - v_j)] \times [Pr(Q' \le \alpha_s - u_i - v_{j'}) - Pr(Q' \le \alpha_{s-1} - u_i -$$

$$=\int_{-\infty}^{+\infty}\sum_{r=1}^{C}\sum_{s=1}^{C} w_{rs}[Pr(Q+v_j \le \alpha_r - u_i) - Pr(Q+v_j \le \alpha_{r-1} - u_i)] \times [Pr(Q'+v_{j'} \le \alpha_s - u_i) - Pr(Q'+v_{j'} \le \alpha_{s-1} - u_i)]\phi$$

$$=\int_{-\infty}^{+\infty}\sum_{r=1}^{C}\sum_{s=1}^{C} w_{rs}\left[Pr\left(\frac{Q+v_j}{\sigma_u} \le \frac{\alpha_r}{\sigma_u} - z\right) - Pr\left(\frac{Q+v_j}{\sigma_u} \le \frac{\alpha_{r-1}}{\sigma_u} - z\right)\right] \times \left[Pr\left(\frac{Q'+v_{j'}}{\sigma_u} \le \frac{\alpha_s}{\sigma_u} - z\right) - Pr\left(\frac{Q'+v_{j'}}{\sigma_u} \le \frac{\alpha_{s-}}{\sigma_u}\right.\right.$$

$$=\int_{-\infty}^{+\infty}\sum_{r=1}^{C}\sum_{s=1}^{C} w_{rs}\left[\Phi\left(\frac{\frac{\alpha_r}{\sigma_u} - z}{\sqrt{\frac{1+\sigma_v^2}{\sigma_u^2}}}\right) - \Phi\left(\frac{\frac{\alpha_{r-1}}{\sigma_u} - z}{\sqrt{\frac{1+\sigma_v^2}{\sigma_u^2}}}\right)\right] \times \left[\Phi\left(\frac{\frac{\alpha_s}{\sigma_u} - z}{\sqrt{\frac{1+\sigma_v^2}{\sigma_u^2}}}\right) - \Phi\left(\frac{\frac{\alpha_{s-1}}{\sigma_u} - z}{\sqrt{\frac{1+\sigma_v^2}{\sigma_u^2}}}\right)\right] \phi(z)\,\mathrm{d}z,$$

$$=\int_{-\infty}^{+\infty}\sum_{r=1}^{C}\sum_{s=1}^{C} w_{rs}\left[\Phi\left(\frac{\alpha_r^* - z\sqrt{\rho}}{\sqrt{1-\rho}}\right) - \Phi\left(\frac{\alpha_{r-1}^* - z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\right] \times \left[\Phi\left(\frac{\alpha_s^{*} - z\sqrt{\rho}}{\sqrt{1-\rho}}\right) - \Phi\left(\frac{\alpha_{s-1}^{*} - z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\right] \phi(z)\,\mathrm{d}z$$

where, $\frac{Q+v_j}{\sigma_u} \sim N\left(0, \frac{1+\sigma_v^2}{\sigma_u}\right)$ and raters j and j' ($j \neq j'$) are interchangeable since from the same large population of raters and $z \sim N(0,1)$.

Appendix 2 Derivation of chance association

$$p_{ca} = \sum_{r=1}^{C}\sum_{s=1}^{C} w_{rs}[Pr(Y_{ij}=r) \times Pr(Y_{i'j'}=s)]$$

$$= \sum_{r=1}^{C}\sum_{s=1}^{C} w_{rs}[Pr(Q \leq \alpha_r - (u_i+v_j)) - Pr(Q \leq \alpha_{r-1} - (u_i+v_j))] \times [Pr(Q \leq \alpha_s - (u_{i'}+v_{j'})) - Pr($$

$$= \sum_{r=1}^{C}\sum_{s=1}^{C} w_{rs}[Pr(Q+u_i+v_j \leq \alpha_r) - Pr(Q+u_i+v_j \leq \alpha_{r-1})] \times [Pr(Q+u_{i'}+v_{j'} \leq \alpha_s) - Pr(Q+$$

$$= \sum_{r=1}^{C}\sum_{s=1}^{C} w_{rs}\left[Pr\left(\frac{Q+u_i+v_j}{\sqrt{1+\sigma_u^2+\sigma_v^2}} \leq \frac{\alpha_r}{\sqrt{1+\sigma_u^2+\sigma_v^2}}\right) - Pr\left(\frac{Q+u_i+v_j}{\sqrt{1+\sigma_u^2+\sigma_v^2}} \leq \frac{\alpha_{r-1}}{\sqrt{1+\sigma_u^2+\sigma_v^2}}\right)\right] \times \left[Pr\left(\frac{Q'+u_{i'}+v_{j'}}{\sqrt{1+\sigma_u^2+\sigma_v^2}} \leq \frac{\alpha_s}{\sqrt{1+\sigma_u^2}}\right.\right.$$

$$= \sum_{r=1}^{C}\sum_{s=1}^{C} w_{rs}[\Phi(\alpha_r^*) - \Phi(\alpha_{r-1}^*)] * [\Phi(\alpha_s^*) - \Phi(\alpha_{s-1}^*)] \quad , 0.5 \leq p_{ca} \leq 1.$$

Appendix 3

Theorem

It can be shown in the population-based setting over many raters with ordinal classifications that $p_{0a} \geq p_{ca} \geq 0.5$.

We begin with a proposition as follows:

Proposition

Under the model in equation (1), observed association is always greater or equal to chance association, i.e. $p_{0a} \geq p_{ca}$.

Proof

Choose two raters at random, and let their ordinal ratings for randomly selected items i and i' be denoted Y_{i1}, $Y_{i'2}$. We allow $i = i'$ to discuss the case where they look at the same randomly selected item. Let $D = |Y_{i1} - Y_{i'2}|$. Let f and F be the mass function and cumulative distribution function of D, respectively.

We first show that p_{0a} can be written as a weighted average of $F(d)$ values, $d = 0, 1, 2, \ldots, C-1$

$$p_{0a} = \sum_{r=1}^{C}\sum_{s=1}^{C} w_{rs} p_{rs}$$

Group terms by diagonals of the weight matrix corresponding to $D = 0$, $D = 1$, etc., and let $w_0=1\geq w_1\geq w_2\geq\ldots w_{C-1}\geq 0$ be the weights as determined by their off-diagonal locations.

$$\begin{aligned} p_{0a} &= \sum_{d=0}^{C-1} w_d f(d) \\ &= F(0) + \sum_{d=1}^{C-1} w_d [F(d) - F(d-1)] \\ &= F(0) + \sum_{d=1}^{C-1} w_d F(d) - \sum_{d=1}^{C-1} w_d F(d-1) \\ &= (1-w_1) F(0) + (w_1 - w_2) F(1) + (w_2 - w_3) F(2) + \ldots (w_{C-2} - w_{C-1}) F(C-2) + w_{C-1} F(C-1). \end{aligned}$$

And since $w_0 = 1$ $w_1\geq w_2\geq\ldots w_{C-1}\geq 0$, all coefficients of $F(d)$ terms above are nonnegative. Now, if we can show that D is stochastically smaller when $i=i'$ than when $i\neq i'$, it will follow that p_{0a}, the above expression with $p_{rs}=\Pr\{(Y_{i1}=r)\cap(Y_{i2}=s)\}$ is greater than p_{ca}, the same expression with $p_{rs}=\Pr\{(Y_{i1}=r)\cap(Y_{i'2}=s)\}$

Suppress the asterisks on $-\infty=\alpha_0^*, \alpha_1^*, \alpha_2^*,\ldots, \alpha_C^*=\infty$. Figure 4 shows the region $\{D\leq 1\}$ in terms of the underlying variables W_{i1}, W_{i2} and an irregular choice of α's (when the threshold values α_c are unequally spaced) when $C=5$. Regions where $D=0$ are squares on the diagonal line $W_{i1}=W_{i2}$. Regions where $D=1$ are rectangles, some of which are infinite in extent. Note that the region $\{D\leq 1\}$, and more generally $\{D\leq d\}$, is symmetric about the line $W_{i1}=W_{i2}$.

Let R be any region symmetric about the line $W_{i1}=W_{i2}$. Define $T_{ii'}=W_{i1}+W_{i'2}$ and $S_{ii'}=W_{i1}-W_{i'2})=(u_i-u_{i'})+(v_1-v_2)$ under our model $W_{ij}=\alpha_c-(u_i+v_j)$. Note that $T_{ii'}$, $S_{ii'}$ are independent. Let G_T denote the cdf of $T_{ii'}$ and given $T_{ii'}=t$, let the symmetric region R be defined by $-b(t)\leq S_{ii'}\leq b(t)$.

$$\begin{aligned} \overset{ii'}{Pr}\{R\} &= \int_{t=-\infty}^{\infty} Pr\{-b(t)\leq S_{ii'}\leq b(t)|T_{ii'}=t\}\partial G_T(t) \\ &= \int_{t=-\infty}^{\infty} Pr\{|S_{ii'}|\leq b(t)\}\partial G_T(t) \\ &= \int_{t=-\infty}^{\infty} Pr\{|(u_i-u_{i'})+(v_1-v_2)|\leq b(t)\}\partial G_T(t) \\ &\leq \int_{t=-\infty}^{\infty} Pr\{|(v_1-v_2)|\leq b(t)\}\partial G_T(t) \\ &= \overset{ii}{Pr}\{R\} \end{aligned}$$

The inequality follows from the fact that $|(u_i-u_{i'})+(v_1-v_2)|$ is stochastically larger than $|(v_1-v_2)|$, they are the absolute values of normal variables with different variances. Since the probability of any region symmetric about the line $W_{i1}=W_{i2}$ is larger when $i=i'$, the random variable D is stochastically smaller when $i=i'$, and the result follows.

We can then demonstrate that $p_{ca}\geq 0.5$ using the following proof by induction: in equation (3), chance association p_{ca} is written as (for the ith item and jth rater)

$$p_{ca}=\sum_{r=1}^{C}\sum_{s=1}^{C}w_{rs}[Pr(Y_{ij}=r)\times Pr(Y_{i'j'}=s)]$$
$$=\sum_{r=1}^{C}\sum_{s=1}^{C}w_{rs}[\Phi(\alpha_r^*)-\Phi(\alpha_{r-1}^*)]\times[\Phi(\alpha_s^*)-\Phi(\alpha_{s-1}^*)].$$

Simplest case

For an ordinal scale with C = 2 categories, where thresholds $\alpha_0=-\infty$ and $\alpha_{C=2}=+\infty$. To find the value of α_1^* which minimizes chance association p_{ca}, we set the first derivative to 0:

$\frac{dp_{ca}}{d\alpha_1^*}=4\Phi(\alpha_1^*)\varphi(\alpha_1^*)-2\phi(\alpha_1^*)=0\Rightarrow\alpha_1^*=0$ minimizes the expression for p_{ca} (with the second derivative > 0) and leads to the minimum value of p_{ca}=0.5 for linear and quadratic weights.

For an ordinal scale with C=3 categories, where $\alpha_0=-\infty$ and $\alpha_{C=3}=+\infty$.

To find the values of α_1^* and α_2^* which minimizes chance association p_{ca}, we can set each of the first derivatives to 0 and jointly solve:

$\frac{dp_{ca}}{d\alpha_1^*}=0$ and $\frac{dp_{ca}}{d\alpha_2^*}=0\Rightarrow\alpha_1^*,\alpha_2^*=0$ minimizes the expression for p_{ca} (with the second derivatives > 0) and leads to the minimum value of p_{ca}=0.5. Since the ordinal GLMM model for association requires monotonically increasing thresholds, we set the thresholds to be $\alpha_0^*<\alpha_1^*<\alpha_2^*<\alpha_3^*$, or −∞ < 0.00001 < 0.00002 < +∞. Including these threshold values into the expression for chance association, p_{ca} leads to a minimum value of p_{ca}=0.5.

More generally, for any number of categories C, where C > 2, where $\alpha_0=-\infty$ and $\alpha_C=+\infty$.

To find the values that minimize chance association p_{ca}, we can set each of the first derivatives to 0 and jointly solve $\Rightarrow\alpha_1^*,\alpha_2^*,\ldots,,\alpha_{C-1}^*=0$ which minimizes the expression for p_{ca} (with the second derivatives > 0) and leads to the minimum value of p_{ca}=0.5. Since the ordinal GLMM model for association requires monotonically increasing thresholds, we set the thresholds to be $\alpha_0^*<\alpha_1^*<\ldots<\alpha_{C-1}^*<\alpha_c^*$ or $\left(\alpha_{\min,1}^*=0.00001,\alpha_{\min,2}^*=0.00002,\ldots\right)$ with $\alpha_0^*=-\infty$ and $\alpha_C^*=+\infty$. By setting all the intermediate thresholds to be close to zero, this effectively turns the minimization into a two-category situation, which then leads to the minimum value of p_{ca}=0.5 as demonstrated above.

Thus, we have shown through proof by induction that the minimum value of p_{ca} is 0.5, and that this occurs when the monotonically increasing thresholds $\alpha_1^*,\alpha_2^*,\ldots,,\alpha_{C-1}^*$ take values −∞<0.00001<0.00002< … < +∞. Since we have demonstrated earlier that $p_{0a}\geq p_{ca}$, we can state that $1\geq p_{0a}\geq p_{ca}\geq 0.5$. These results hold for linear and quadratic weights.

References

1. Kallman DA, Wigley FM, Scott WW, et al. New radiographic grading scales for osteoarthritis of the hand. Reliability for determining prevalence and progression. Arthritis Rheum. 1989; 32:1584–1591. [PubMed: 2490851]
2. Faust HB, Gonin R, Chuang TY, et al. Reliability testing of the dermatology index of disease severity (DIDS) – An index for staging the severity of cutaneous inflammatory disease. Arch Dermatol. 1997; 133:1443–1448. [PubMed: 9371030]
3. Gonin R, Lipsitz SR, Fitzmaurice GM, et al. Regression modelling of weighted kappa by using generalized estimating equations. J Roy Stat Soc Ser C. 2000; 49:1–18.
4. Brumback RJ, Jones AL. Interobserver agreement in the classification of open fractures of the tibia – the results of a survey of 245 orthopedic surgeons. J Bone Joint Surg. 1994; 76A:1162–1166.
5. Sickles, EA., D'Orsi, CJ., Bassett, LW. ACR BI-RADS atlas, breast imaging reporting and data system. Reston, VA: American College of Radiology; 2013. ACR BI-RADS mammography.
6. Gleason, DF. The Veteran's Administration Cooperative Urologic Research Group: Histologic grading and clinical staging of prostatic carcinoma. In: Tannnbaum, M., editor. Urologic pathology: The prostate. 1977. p. 171-198.
7. Epstein JI, Allsbrook WC, Amin MB, et al. The 2005 International Society of Urological Pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma. Am J Surg Pathol. 2005; 29:1228–1242. [PubMed: 16096414]
8. Holmquist ND, McMahan CA, Williams OD. Variability in classification of carcinoma in situ of the uterine cervix. Arch Pathol. 1967; 84:334–345. [PubMed: 6045443]
9. Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. J Natnl Cancer Inst. 2003; 95:282–290.
10. Elmore JG, Jackson SL, Abraham L, et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology. 2009; 253:641–651. [PubMed: 19864507]
11. Onega T, Smith M, Miglioretti DL, et al. Radiologist agreement for mammographic recall by case difficulty and finding type. J Am Coll Radiol. 2012; 9:788–794. [PubMed: 23122345]
12. Miglioretti DL, Smith-Bindman R, Abraham L, et al. Radiologist characteristics associated with interpretive performance of diagnostic mammography. J Ntnl Cancer Inst. 2007; 99:1854–1863.
13. Frandsen PA, Andersen E, Madsen F, et al. Garden classification of femoral-neck fractures – an assessment of interobserver variation. J Bone Joint Surg. 1988; 70:588–590.
14. Bloch DA, Kraemer HC. 2×2 kappa-coefficients – measures of agreement or association. Biometrics. 1989; 45:269–287. [PubMed: 2655731]
15. Kraemer HC. Measurement of reliability for categorical data in medical research. Stat Meth Med Res. 1992; 1:183–199.
16. 'Graham P, Jackson R. The analysis of ordinal agreement data – beyond weighted kappa. J Clin Epidemiol. 1993; 46:1055–1062. [PubMed: 8263578]
17. Svensson E, Holm S. Separation of systematic and random differences in ordinal rating-scales. Stat Med. 1994; 13:2437–2453. [PubMed: 7701145]
18. Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: A review. Test. 2009; 18:1–43. [PubMed: 21218187]
19. Cohen J. Weighted kappa – nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968; 70:213–220. [PubMed: 19673146]
20. Svensson E. Different ranking approaches defining association and agreement measures of paired ordinal data. Stat Med. 2012; 31:3104–3117. [PubMed: 22714677]
21. Coull BA, Agresti A. Generalized log-linear models with random effects, with application to smoothing contingency tables. Stat Model. 2003; 3:251–271.
22. Uebersax JS, Grove WM. A latent trait finite mixture model for the analysis of rating agreement. Biometrics. 1993; 49:823–835. [PubMed: 10798855]
23. Tanner MA, Young MA. Modeling agreement among raters. J Am Stat Assoc. 1985; 80:175–180.

24. Agresti A. A model for agreement between ratings on an ordinal scale. Biometrics. 1988; 44:539–548.
25. Becker MP, Agresti A. Log-linear modeling of pairwise interobserver agreement on a categorical scale. Stat Med. 1992; 11:101–114. [PubMed: 1557566]
26. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960; 20:37–46.
27. Nelson JC, Pepe MS. Statistical description of interrater variability in ordinal ratings. Stat Meth Med Res. 2000; 9:475–496.
28. Shrout PE, Fleiss JL. Intraclass correlations – uses in assessing rater reliability. Psychol Bull. 1979; 86:420–428. [PubMed: 18839484]
29. Berry KJ, Johnston JE, Mielke PW. Weighted kappa for multiple raters. Percept Motor Skills. 2008; 107:837–848. [PubMed: 19235413]
30. Mielke PW, Willett WC. Unweighted and weighted kappa as measures of agreement for multiple judges. Int J Manage. 2009; 26:213–223.
31. Maclure M, Willett WC. Misinterpretation and misuse of the kappa-statistic. Am J Epidemiol. 1987; 126:161–169. [PubMed: 3300279]
32. Williamson JM, Lipsitz SR, Manatunga AK. Modeling kappa for measuring dependent categorical agreement data. Biostatistics. 2000; 1:191–202. [PubMed: 12933519]
33. Johnson VE. On Bayesian analysis of multirater ordinal data: An application to automated essay grading. J Am Stat Assoc. 1996; 91:42–51.
34. Johnson, VE., Albert, JH. Ordinal data modeling (Statistics for Social Science and Public Policy). New York: Springer; 1999.
35. Hsiao CK, Chen PC, Kao WH. Bayesian random effects for interrater and test-retest reliability with nested clinical observations. J Clin Epidemiol. 2011; 64:808–814. [PubMed: 21292442]
36. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2015. http://www.R-project.org/
37. Hedeker D, Gibbons RD. A random-effects ordinal regression-model for multilevel analysis. Biometrics. 1994; 50:933–944. [PubMed: 7787006]
38. Liu I, Agresti A. The analysis of ordered categorical data: An overview and a survey of recent developments. Test. 2005; 14:1–30.
39. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Am Stat Assoc. 1993; 88:9–25.
40. Agresti, A. Analysis of ordinal categorical data. 2nd. New York: Wiley; 2010.
41. Nelson KP, Edwards D. On population-based measures of agreement for binary classifications. Can J Stat. 2008; 36:411–426.
42. Nelson KP, Edwards D. Measures of agreement between many raters for ordinal classification. Stat Med. 2015; 34:3116–3132. [PubMed: 26095449]
43. Shun ZM, McCullagh P. Laplace approximation of high-dimensional integrals. J Roy Stat Soc Ser B. 1995; 57:749–760.
44. Gueorguieva R. Multivariate generalized linear mixed model for joint modeling of clustered outcomes in the exponential family. Stat Model. 2001; 1:177–193.
45. Capanu M, Gonen M, Begg CB. An assessment of estimation methods for generalized linear mixed models with binary outcomes. Stat Med. 2013; 32:4550–4566. [PubMed: 23839712]
46. Fleiss JL, Cohen J. Equivalence of weighted kappa and intraclass correlation coefficient as measures of reliability. Educ Psychol Meas. 1973; 33:613–619.
47. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971; 76:378–382.
48. Light RJ. Measures of response agreement for qualitative data – some generalizations and alternatives. Psychol Bull. 1971; 76:365–377.
49. Conger AJ. Integration and generalization of kappas for multiple raters. Psychol Bull. 1980; 88:322–328.
50. Landis JR, Koch GG. Measurement of observer agreement for categorical data. Biometrics. 1977; 33:159–174. [PubMed: 843571]

51. Allsbrook WC, Mangold KA, Johnson MH, et al. Interobserver reproducibility of Gleason grading of prostatic carcinoma: General pathologist. Human Pathol. 2001; 32:81–88. [PubMed: 11172299]

52. Allsbrook WC, Mangold KA, Johnson MH, et al. Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. Human Pathol. 2001; 32:74–80. [PubMed: 11172298]

53. Tagliafico A, Tagliafico G, Tosto S, et al. Mammographic density estimation: Comparison among BI-RADS categories, a semi-automated software and a fully automated one. Breast. 2009; 18:35–40. [PubMed: 19010678]

54. Banerjee M. Beyond kappa: A review of interrater agreement measures. Can J Stat. 1999; 27:3–23.

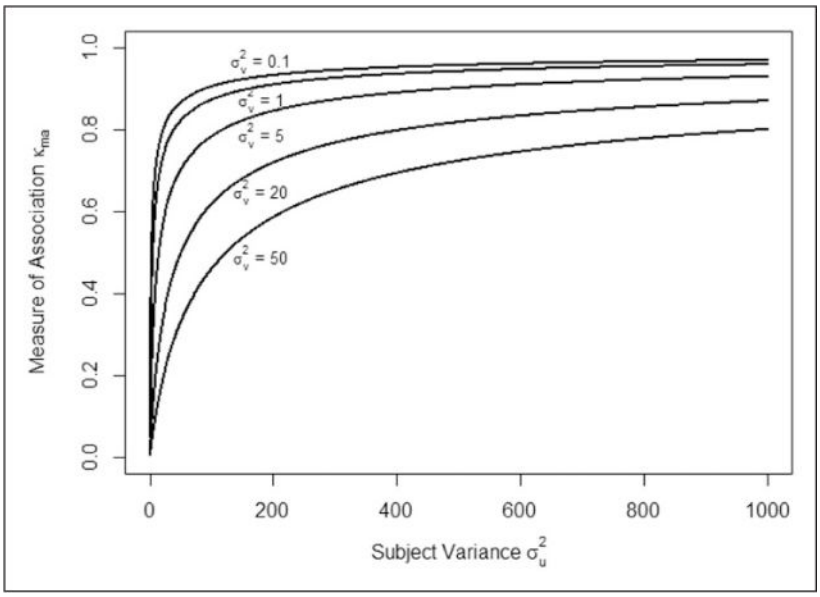

Figure 1.

The effects of varying rater random effects variance σ_v^2 and subject random effects variances σ_u^2 on the proposed measure of association κ_{ma} with quadratic weights.

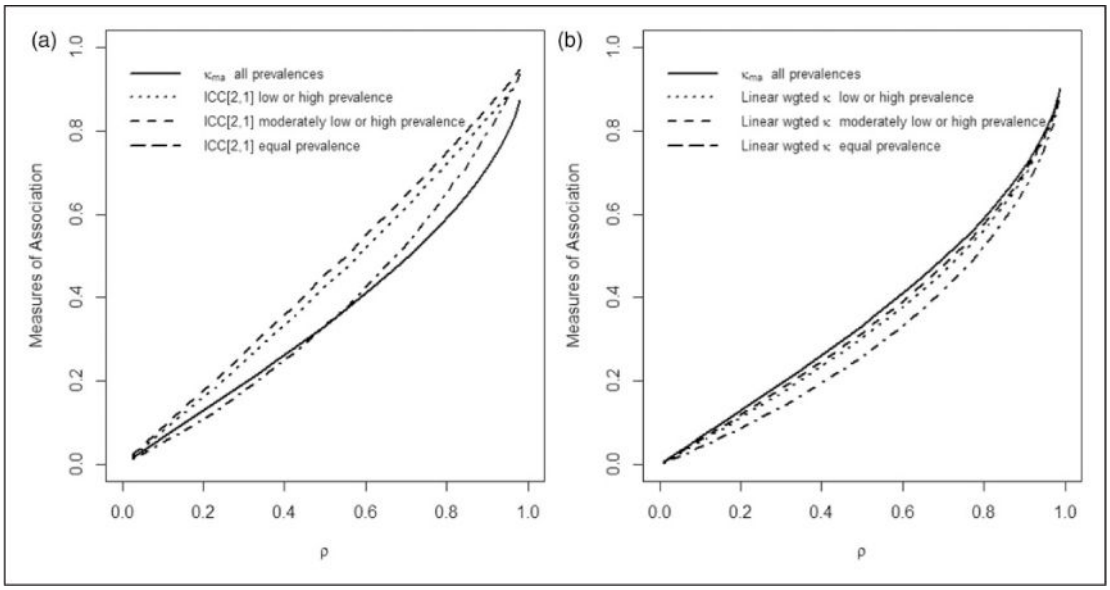

Figure 2.

Plots of measures of association versus ρ at different prevalences (a) Shrout and Fleiss' ICC[2,1] and κ_{ma}; (b) κ_{ma} and Cohen's GLMM-based weighted kappa with linear weights $\kappa_{GLMM,a}$. The prevalence is varied (extreme low or high; moderately high or low; equal in each category) with the percent of observations falling into each of the $C = 5$ categories for each prevalence case given in Table 5.

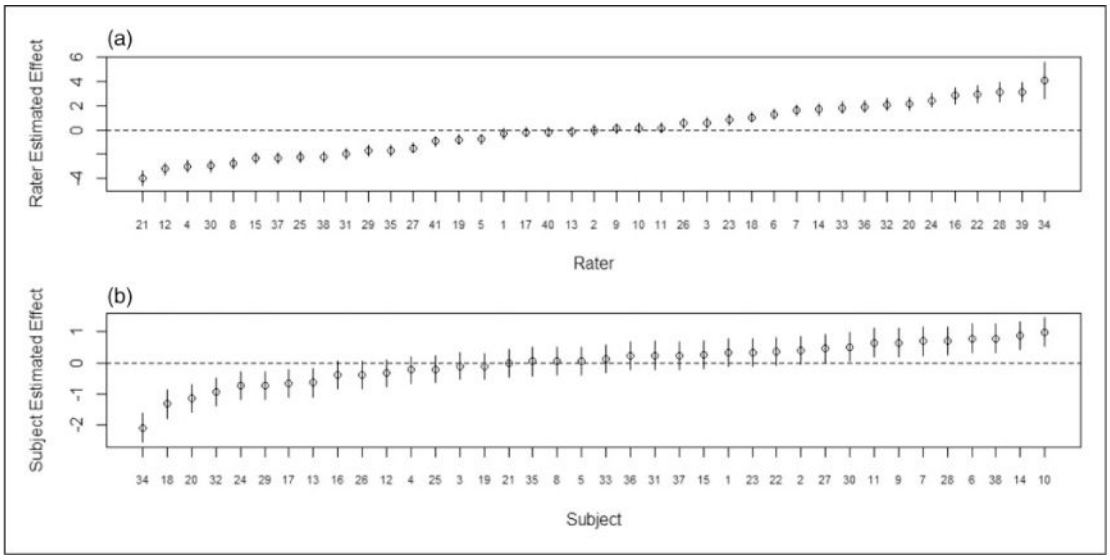

Figure 3.

(a) Rater and (b) subject effects estimated as conditional modes for the Gleason grading study.[51] Figure 3(a) shows $J = 41$ rater estimated effects and Figure 3(b) shows $I = 38$ subject estimated effects respectively, with 95% confidence intervals based upon the conditional variance.

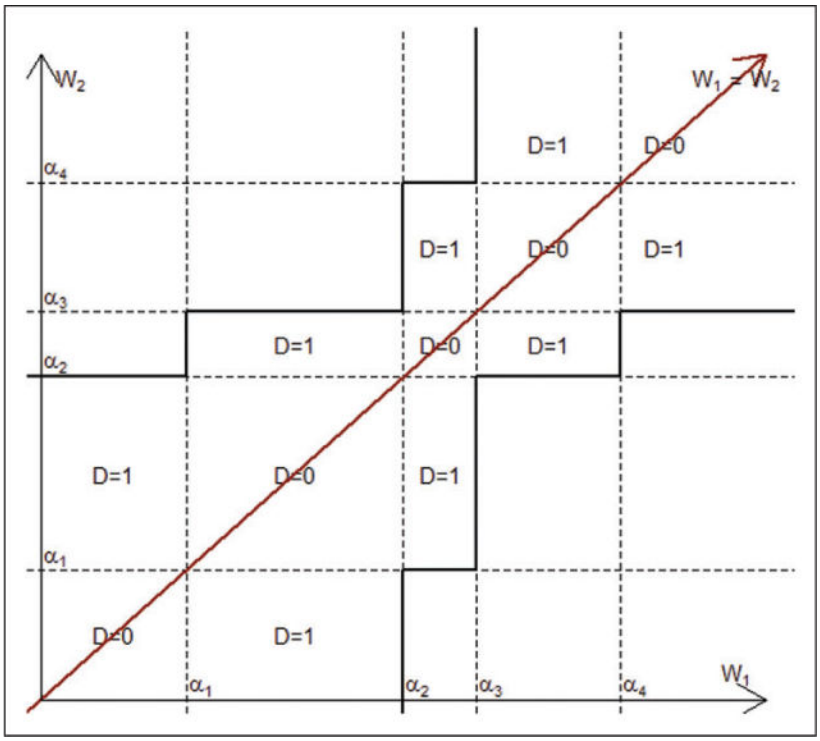

Figure 4.

The region $\{D \leq 1\}$ in terms of the underlying variables W_{i1}, W_{i2} and an irregular choice of α's (when the threshold values α_c are unequally spaced) when C=5.

Table 1

Parameter values for the simulation scenarios examined.

Variance components $\left(\sigma_u^2, \sigma_v^2\right)$	**Random effects distributions**	**Number of items *I*, Number of raters *J***
(1, 5)		
(5, 1)	$u_i \sim$ Exp(λ) and $v_j \sim$ Unif(a, $-a$)	(I=100, J=10)
(5, 20)	$u_i \sim \mathrm{N}\left(0, \sigma_u^2\right)$ and $v_j \sim \mathrm{N}\left(0, \sigma_v^2\right)$	(I=250, J=100)
(20, 5)		
(10, 10)		

Table 2

True values and mean estimates (mean standard error) of the proposed measure of association κ_{ma} for each of the 20 simulation studies. Each set of simulations is based upon 1000 simulated datasets with C=5 categories.

(a) Normally distributed random effects

		I=100, J=10	I=250, J=100
$\left(\sigma_u^2, \sigma_v^2\right)$	True κ_{ma}	Mean $\hat{\kappa}_{ma}$ (S.E.)	Mean $\hat{\kappa}_{ma}$ (S.E.)
(1, 5)	0.091	0.110 (0.033)	0.094 (0.012)
(5, 20)	0.123	0.153 (0.050)	0.127 (0.017)
(10, 10)	0.316	0.347 (0.075)	0.320 (0.028)
(5, 1)	0.506	0.508 (0.046)	0.503 (0.021)
(20, 5)	0.559	0.560 (0.066)	0.551 (0.026)

(b) Non-normally distributed random effects

		I=100, J=10	I=250, J=100
$\left(\sigma_u^2, \sigma_v^2\right)$	True κ_{ma}	Mean $\hat{\kappa}_{ma}$ (S.E.)	Mean $\hat{\kappa}_{ma}$ (S.E.)
(1, 5)	0.091	0.104 (0.033)	0.091 (0.012)
(5, 20)	0.123	0.128 (0.044)	0.118 (0.016)
(10, 10)	0.316	0.312 (0.073)	0.293 (0.027)
(5, 1)	0.506	0.499 (0.047)	0.476 (0.021)
(20, 5)	0.559	0.540 (0.068)	0.504 (0.027)

Table 3

Results from five simulation studies. Each is based upon 1000 datasets simulated from an ordinal GLMM with five categories (C=5), with thresholds $\alpha_0 = -\infty$ and $\alpha_5 = +\infty$ and I= 100 items and J= 10 raters. Random effects $u_i \sim \mathrm{N}(0, \sigma_u^2)$, and $v_j \sim \mathrm{N}(0, \sigma_v^2)$.

***C* = 5 Categories, *I*= 100 items, *J* = 10 raters**															
	Simulation set #1			**Simulation set #2**			**Simulation set #3**			**Simulation set #4**			**Simulation set #5**		
Parameter	**Truth**	**Est. mean**	**Est. S.E. (obs.)**	**Truth**	**Est. mean**	**Est. S.E. (obs.)**	**Truth**	**Est. mean**	**Est. S.E. (obs.)**	**Truth**	**Est. mean**	**Est. S.E. (obs.)**	**Truth**	**Est. mean**	**Est. S.E. (obs.)**
σ_u^2	1	1.008	0.195 (0.202)	5	4.798	0.878 (0.867)	5	5.019	0.971 (1.063)	20	17.465	3.467 (3.324)	10	9.778	1.816 (1.866)
σ_v^2	5	4.555	2.221 (2.375)	1	0.910	0.423 (0.441)	20	18.654	9.631 (12.617)	5	4.442	2.059 (2.234)	10	9.057	4.243 (4.723)
α_1	0	−0.006	0.673 (0.745)	0	−0.010	0.378 (0.405)	0	0.026	1.360 (1.647)	0	−0.046	0.795 (0.855)	0	−0.026	0.984 (1.083)
α_2	1	0.998	0.675 (0.748)	1	0.989	0.380 (0.409)	1	1.030	1.362 (1.651)	1	0.941	0.797 (0.857)	1	0.973	0.985 (1.083)
α_3	2	2.000	0.678 (0.754)	2	1.988	0.385 (0.410)	2	2.034	1.365 (1.649)	2	1.936	0.801 (0.858)	2	1.973	0.989 (1.088)
α_4	3	3.005	0.684 (0.755)	3	2.988	0.394 (0.421)	3	3.036	1.370 (1.654)	3	2.931	0.807 (0.864)	3	2.982	0.994 (1.093)
ρ	0.143	0.171	0.051 (0.063)	0.714	0.714	0.051 (0.05)	0.192	0.236	0.075 (0.094)	0.769	0.766	0.067 (0.069)	0.476	0.514	0.099 (0.111)
var(ρ)	0.002	0.004		0.003	0.003		0.0049	0.0088		0.005	0.005		0.0115	0.0123	
κ_m	0.035	0.043	0.014 (0.017)	0.264	0.267	0.032 (0.038)	0.048	0.062	0.023 (0.029)	0.306	0.312	0.050 (0.059)	0.141	0.162	0.042 (0.050)
var(κ_m)	0.0002	0.0002 (0.0003)		0.001	0.001 (0.001)		0.0004	0.0006 (0.0008)		0.0029	0.0025 (0.0035)		0.0018	0.0018 (0.0025)	
κ_{ma}	0.091	0.110	0.033 (0.041)	0.506	0.508	0.046 (0.050)	0.123	0.153	0.050 (0.062)	0.559	0.560	0.066 (0.068)	0.316	0.347	0.075 (0.084)
var(κ_{ma})	0.001	0.0011 (0.0017)		0.002	0.002 (0.002)		0.002	0.003 (0.004)		0.005	0.004 (0.005)		0.0060	0.006 (0.007)	

Table 4

Results from five simulation studies. Each is based upon 1000 datasets simulated from an ordinal GLMM with five categories ($C = 5$), with thresholds $\alpha_0 = -\infty$ and $\alpha_5 = +\infty$ and ($I = 250$ items and $J = 100$ raters. Random effects $u_i \sim \mathrm{N}(0, \sigma_u^2)$, and $v_j \sim \mathrm{N}(0, \sigma_v^2)$.

***C* = 5 Categories, *I* = 250 items, *J* = 100 raters**															
	Simulation set #1			**Simulation set #2**			**Simulation set #3**			**Simulation set #4**			**Simulation Set #5**		
Parameter	**Truth**	**Est. mean**	**Est. S.E. (obs.)**	**Truth**	**Est. mean**	**Est. S.E. (obs.)**	**Truth**	**Est. mean**	**Est. S.E. (obs.)**	**Truth**	**Est. mean**	**Est. S.E. (obs.)**	**Truth**	**Est. mean**	**Est. S.E. (obs.)**
σ_u^2	1	1.001	0.093 (0.093)	5	4.942	0.473 (0.454)	5	5.002	0.460 (0.444)	20	19.226	1.896 (1.785)	10	9.997	0.925 (0.934)
σ_v^2	5	4.879	0.721 (0.713)	1	1.006	0.144 (0.142)	20	19.566	2.958 (2.922)	5	5.016	0.718 (0.721)	10	9.844	4.414 (4.436)
α_1	0	−0.003	0.231 (0.234)	0	−0.013	0.174 (0.175)	0	−0.009	0.467 (0.480)	0	−0.017	0.359 (0.373)	0	−0.01 1	0.372 (0.373)
α_2	1	0.997	0.231 (0.234)	1	0.986	0.174 (0.175)	1	0.991	0.467 (0.481)	1	0.982	0.360 (0.373)	1	0.990	0.372 (0.373)
α_3	2	2.000	0.231 (0.234)	2	1.987	0.174 (0.176)	2	1.991	0.468 (0.481)	2	1.982	0.360 (0.374)	2	1.988	0.372 (0.374)
α_4	3	2.998	0.231 (0.233)	3	2.988	0.175 (0.177)	3	2.990	0.468 (0.481)	3	2.982	0.360 (0.373)	3	2.988	0.373 (0.373)
ρ	0.143	0.147	0.018 (0.019)	0.714	0.710	0.023 (0.023)	0.192	0.198	0.026 (0.026)	0.769	0.761	0.027 (0.026)	0.476	0.481	0.039 (0.039)
var(ρ)	0.0003	0.0004		0.0005	0.0005		0.0006	0.0007		0.0007	0.0007		0.0015	0.0016	
κ_m	0.035	0.036	0.005 (0.005)	0.264	0.262	0.015 (0.016)	0.048	0.049	0.007 (0.007)	0.306	0.300	0.020 (0.021)	0.141	0.144	0.015 (0.016)
var(κ_m)	0.00002	0.00002 (0.00003)		0.0002	0.0002 (0.0002)		0.00005	0.00005 (0.00005)		0.0004	0.0004 (0.0005)		0.0002	0.0002 (0.0003)	
κ_{ma}	0.091	0.094	0.012 (0.012)	0.506	0.503	0.021 (0.021)	0.123	0.127	0.017(0.017)	0.559	0.551	0.026 (0.026)	0.316	0.320	0.028 (0.029)
var(κ_{ma})	0.0001	0.0001 (0.0002)		0.0004	0.0004 (0.0004)		0.0003	0.0003 (0.0003)		0.0007	0.0007 (0.0007)		0.0008	0.0008 (0.0008)	

Table 5

Varying disease prevalence examined in Figure 2(a) and (b) based upon an ordinal classification scale with C=5 categories.

	Percentage (%) of classifications in each category					
Disease prevalence	**Category 1**	**Category 2**	**Category 3**	**Category 4**	**Category 5**	**Total**
Very low	80	10	3.4	3.3	3.3	100
Moderately low	50	26	16	6	2	100
Equal	20	20	20	20	20	100
Moderately high	2	6	16	26	50	100
Very high	3.3	3.3	3.4	10	80	100

Table 6

Table of classifications by individual raters for the beam mammogram study.[9] Based upon an ordinal classification scale with C=5 categories; I=148 mammograms; J=104 radiologists.

	Classifications by individual radiologists (J=104)									
	Rater									
Subject	**1**	**2**	**3**	**4**	**…**	**100**	**101**	**102**	**103**	**104**
1	1	1	1	1		4	4	3	2	3
2	1	4	3	4	…	3	2	1	4	2
3	4	4	4	4		4	3	4	5	4
4	5	4	5	4		4	2	3	5	3
5	2	4	3	2	…	4	5	5	2	2
⋮			⋮				⋮			⋮
144	1	4	3	1		1	1	1	1	2
145	4	1	4	4		1	4	5	5	3
146	4	3	4	2	…	5	5	3	5	3
147	5	4	5	5		4	5	5	5	4
148	5	4	4	4		5	3	3	5	4

Table 7

Results for the beam mammogram study[9] where 104 radiologists (J=104) classified mammograms of 148 patients (I = 148) using an ordered BIRADS scale with C = 5 categories (1 = normal; 2=benign; 3=probably benign; 4=possibly malignant; 5 = probably malignant).

Parameter	Symbol	Estimate	S.E.	Z-value
Ordinal GLMM:				
Thresholds: ($\alpha_0 = -\infty$, $\alpha_5 = +\infty$)				
Between categories 1 and 2	α_1	−0.897	0.135	−6.643
Between categories 2 and 3	α_2	−0.197	0.135	−1.460
Between categories 3 and 4	α_3	0.761	0.135	5.630
Between categories 4 and 5	α_4	2.539	0.137	18.574
Subject random effect variance	σ_u^2	2.442	0.427	
Rater random effect variance	σ_v^2	0.158	0.073	
Rho	ρ	0.678	0.026	
GLMM-based observed agreement	p_0	0.430		
GLMM-based observed association (quadratic weights)	p_{0a}	0.907		
Measures of agreement:				
Model-based (unweighted) kappa (Nelson and Edwards[41])	κ_m	0.241	0.015	
Fleiss' kappa (Fleiss[47])	κ_F	0.297	0.001	
Light and Conger's kappa (Light,[48] Conger[49])	κ_{LC}	0.298		
Measures of association: (with quadratic weights)				
Model-based weighted kappa	κ_{ma}	0.475	0.022	
Shrout and Fleiss' ICC[2,1] (Shrout and Fleiss[28])		0.652	95% c.i. = (0.601, 0.706)	
Cohen's GLMM-based weighted kappa	$\kappa_{GLMM,a}$	0.611		

Table 8

Parameter estimates for the Gleason grading study[51] with $J = 41$ general pathologists classifying the severity of prostate cancer of $I = 38$ patients from biopsy slides using a modified version of the Gleason grading scale with $C = 4$ categories: Category (i) Gleason scores 2–4 (mild disease); Category (ii) Gleason scores 5–6; Category (iii) Gleason score 7; Category (iv) Gleason scores 8–9 (severe disease).

Parameter	Symbol	Estimate	S.E.	Z-value
Ordinal GLMM:				
Thresholds: ($\alpha_0 = -\infty$, $\alpha_4 = +\infty$)				
Between categories 1 and 2	α_1	−2.416	0.382	−6.326
Between categories 2 and 3	α_2	−0.218	0.377	−0.578
Between categories 3 and 4	α_3	1.168	0.378	3.094
Subject random effect variance	σ_u^2	4.805	0.382	
Rater random effect variance	σ_v^2	0.480	0.368	
Rho	ρ	0.765	0.043	
GLMM-based observed agreement	p_0	0.531		
GLMM-based observed association (quadratic weights)	p_{0a}	0.917		
Measures of agreement:				
Model-based (unweighted) kappa (Nelson and Edwards[41])	κ_m	0.357	0.036	
Fleiss' kappa (Fleiss[47])	κ_F	0.404		
Light and Conger's kappa (Light,[48] Conger[49])	κ_{LC}	0.405		
Measures of association: (with quadratic weights)				
Model-based weighted kappa	κ_{ma}	0.554	0.043	
Shrout and Fleiss' ICC[2,1] (Shrout and Fleiss[28])		0.734	95% c.i. = (0.642, 0.824)	
Cohen's GLMM-based weighted kappa	$\kappa_{GLMM,a}$	0.687		