Published in final edited form as:

Stat Med. 2015 October 15; 34(23): 3116–3132. doi:10.1002/sim.6546.

Measures of Agreement Between Many Raters for Ordinal Classifications

Kerrie P. Nelson[*] and
Department of Biostatistics, Boston University, 801 Massachusetts Avenue, Boston, MA 02118.

Don Edwards
Department of Statistics, University of South Carolina, Columbia SC 29208 edwards@stat.sc.edu

Abstract

Screening and diagnostic procedures often require a physician's subjective interpretation of a patient's test result using an ordered categorical scale to define the patient's disease severity. Due to wide variability observed between physicians' ratings, many large-scale studies have been conducted to quantify agreement between multiple experts' ordinal classifications in common diagnostic procedures such as mammography. However, very few statistical approaches are available to assess agreement in these large-scale settings. Existing summary measures of agreement rely on extensions of Cohen's kappa [1 - 5]. These are prone to prevalence and marginal distribution issues, become increasingly complex for more than three experts or are not easily implemented. Here we propose a model-based approach to assess agreement in large-scale studies based upon a framework of ordinal generalized linear mixed models. A summary measure of agreement is proposed for multiple experts assessing the same sample of patients' test results according to an ordered categorical scale. This measure avoids some of the key flaws associated with Cohen's kappa and its extensions. Simulation studies are conducted to demonstrate the validity of the approach with comparison to commonly used agreement measures. The proposed methods are easily implemented using the software package R and are applied to two large-scale cancer agreement studies.

Keywords

Inter-rater agreement; Cohen's kappa; Fleiss' kappa; Generalized linear mixed model; Ordinal categorical data

1. Introduction

Ordered categorical scales with three or more ordered categories are commonly used in screening and diagnostic tests to classify a patient's disease or health status, with higher scores often linked with increasing disease severity. Examples include the Gleason grading scale for categorizing severity of a patient's prostate cancer from biopsies [6,7] and the New York Heart Association functional four-point classification scale for categorizing a patient's

[*] Contact Author: Kerrie P. Nelson, Department of Biostatistics, Boston University, 801 Massachusetts Avenue, Boston, MA 02118. kerrie@bu.edu Phone: 617-638-5866 Fax: 617-638-6484.

level of heart failure [8]. Due to imperfect screening and diagnostic test procedures, classification of a patient's test result often involves some degree of subjective interpretation by an expert, and substantial variability between experts' ratings has been observed in many widely used diagnostic and screening procedures, including mammography [9-11]. Consequently large-scale agreement studies involving many experts and ordinal classification scales are becoming increasingly common to assess levels of agreement and to investigate factors that may be linked with the observed variability between experts [9-13], providing a strong motivation to develop statistical approaches that can flexibly handle many raters. While some statistical methods exist to assess agreement between two experts using an ordinal classification scale, only a very limited number of approaches exist to assess agreement between ordinal classifications in larger-scale studies involving multiple (more than two or three) experts.

In this paper we explore a flexible approach to assess agreement between multiple experts classifying the same sample of patients' test results according to an ordered categorical scale. Since diagnostic tests have such widespread use, an important advantage of our large-scale approach is the ability to generalize findings to experts and patients who typically use these diagnostic tests, if experts and subjects are randomly sampled from their populations.

Measures of agreement for ordinal classifications focus on quantifying levels of exact agreement between experts, i.e. where experts each assign an identical category to a patient's test result. Existing summary measures for assessing agreement between multiple raters when ordinal classifications are being examined include Fleiss' kappa [3], Light's and Conger's kappa [4,5], Kraemer's kappa coefficient [14] and an AC2 statistic [15]. Many of these existing summary measures of agreement for ordinal classifications are either extensions of Cohen's kappa or are formulated as a Cohen's kappa-like statistic [3-5,16] and are prone to the same issues as the original Cohen's kappa, including sensitivity to marginal distributions of the experts and disease prevalence effects [16-18]. Model-based approaches for assessing agreement between multiple experts for ordinal classifications include log-linear models [19-21], latent class models [22], a marginal model generalized estimating equations method with Cohen's kappa-like formulated summary measures of agreement [16], a Bayesian approach with a nested random effect structure [23] and an exploratory graphical approach [24]. In practice, many of these procedures and measures are difficult to implement due to a lack of availability in statistical software packages, and/or become increasingly complex for more than three experts. Banerjee et al [25] provide a comprehensive list of agreement measures for nominal and ordinal classification data.

The proposed population-based approach utilizes the framework of ordinal generalized linear mixed models and generates a model-based summary measure of agreement for ordinal classifications based upon variance components for unobservable variables. Unlike most other available summary measures, our proposed agreement measure appropriately corrects for chance agreement in a different formulation from Cohen's kappa, and consequently is not influenced by the prevalence of the disease. Earlier work has demonstrated use of variance components in agreement studies for binary (for example, diseased versus not-diseased) classifications [26,27], and ordinal classifications [23,28,29], where the ordered nature of the classifications brings a unique set of challenges to the estimation and modeling process beyond those of binary classifications. Our methods can be extended to incorporate characteristics of experts and subjects that may potentially influence agreement. Unbalanced observations are allowed, where not every expert classifies every subject in the sample. Unique features of individual experts in the study can be examined, and the population-based nature of the approach ensures that conclusions can be drawn about agreement between typical experts and subjects from their respective underlying populations, not only about experts and subjects included in the study.

The paper is structured as follows. General concepts of agreement for ordered classifications in the population-based setting are defined in the next section. Section 3 describes the proposed model of agreement with details on estimation and fitting, and the proposed measure of agreement is presented in Section 4. Simulation studies demonstrating validity of the proposed approach are included in Section 5, with methods applied to two large-scale cancer agreement studies in Section 6. Section 7 presents a comparison to a logistic generalized linear mixed model while we conclude with a brief discussion in Section 8.

2. A Measure of Agreement in the Population-Based Setting

The primary goal of a population-based agreement study is to draw inference regarding levels of agreement between typical experts and patients in a specified setting such as a diagnostic or screening test. When experts and patients are randomly sampled from their respective populations, a well-defined measure of agreement describes how well one expert's classification of a subject agrees with what other experts would have reported (inter-rater reliability), after appropriately correcting for chance agreement [30]. For a single test result, agreement between two experts is defined where both experts classify the subject into the same category using an ordinal classification scale. In this setting, a natural measure of observed agreement, p_0, is the proportion of time two experts j and j' ($j \neq j'$) assign the ith subject to the same category (1). Since the two experts are randomly selected, classifications made by the jth and j'th experts on a subject are interchangeable, and thus any pair of ratings has a distribution that is invariant under permutations of the experts [25]. The raw observed agreement rate is the proportion of the total number of pairs of ratings which place subjects into the same category.

Chance agreement is the proportion of time experts agree in their classifications simply due to coincidence. The true measure of agreement expected by chance, p_c is the probability that an identical categorical rating is given to two randomly selected subjects i and i' ($i \neq i'$) by two randomly selected experts j and j' ($j \neq j'$) based upon an ordinal classification scale with C categories (1):

$$p_0=\sum_{c=1}^{C}\left[Pr\left(Y_{ij}=c\cap Y_{ij'}=c\right)\right] \qquad \text{and} \qquad p_c=\sum_{c=1}^{C}\left[pr\left(Y_{ij}=c\right)\times pr\left(Y_{i'j'}=c\right)\right] \quad (1)$$

These quantities of observed and chance agreement hold in general in the population-based setting and do not rely on any particular statistical model. An important link between p_0 and p_c and their minimum values in this setting is shown in Theorem One (proof in Appendix 1, Supplementary Materials):

Theorem 1

In the population-based setting, if the number of experts N in the population is not small, $p_0 \geq p_C \geq 1/C$.

The lower bound $1/C$ is likely to be achieved when the disease status of subjects' test results is not easily recognizable from the screening or diagnostic test, such that experts essentially randomly classify subject test results into one of the C categories (equivalent to the roll of a C-sided die).

3. Proposed Model of Agreement for Ordered Classifications

The class of ordinal generalized linear mixed models (GLMMs) [31-33] provides a natural and appealing framework for modeling agreement in large-scale studies between experts in the population-based setting since any number (at least three) of experts and subjects can be included without increasing the complexity of the model, in contrast to many other approaches, and each expert may rate some or all of the subjects [34]. Factors such as rater training or experience can be incorporated into the model to assess their impact on agreement between experts [16,35]. When experts and subjects included in the study are randomly sampled from their populations, results can be generalized to future users (both experts and subjects) of the diagnostic test under study.

We assume that subject i's true disease status as assigned by expert j is an unobserved continuous latent variable W_{ij} linked to the observed ordered classifications through a series of strictly monotonically increasing thresholds $\alpha_0,...,\alpha_c$ dividing the real line into $C+1$ intervals, with $\alpha_0 = -\infty$ and $\alpha_c = +\infty$ [32,33,36]. The latent variable W_{ij} depends on unobserved subject and expert random effects and can be modeled using a linear random effects regression model. In its simplest form $W_{ij} = \beta_0 + u_i + v_j + \varepsilon_{ij}$ where errors ε_{ij} are assumed $N(0, \sigma^2)$, β_0 is the intercept, and u_i and v_j are random effects for the ith subject and jth expert respectively, assumed to come from $N(0, \sigma_u^2)$ and $N(0, \sigma_v^2)$ distributions. The observed categorical classification $Y_{ij} = c$ occurs if and only if $\alpha_{c-1} \leq W_{ij} < \alpha_c$, $c=1,....,C$. The absolute location β_0 and scale σ of the latent variable are not identifiable; to overcome this identifiability issue, wlog set $\beta_0 = 0$ and $\sigma=1$ [32]. Our set-up assumes that each of J experts ($j=1,..., J$) independently classifies each of I subjects' test results ($i=1,..., I$) according to an ordered classification scale with C categories, $c=1,...,C$, yielding classifications $Y_{ij} = c$. The probability $\Pr(Y_{ij} = c)$ can be estimated using an ordinal generalized linear mixed model with a crossed random effects structure, appropriately accounting for dependency between classifications caused by the fact that the same sample of subjects is being classified by each expert. The ordinal GLMM with a probit link calculates the cumulative probability of a subject's test result being classified into category c or lower:

$$\Phi^{-1}\left(Pr\left(Y_{ij} \leq c|u_j, v_j\right)\right)=\alpha_c-\left(u_i+v_j\right) \quad \text{or} \quad Pr\left(Y_{ij} \leq c|u_i, v_j\right)=Pr\left(W_{ij}<\alpha_c\right)=\Phi\left(\alpha_c-\left(u_i+v_j\right)\right) \tag{2}$$

such that $Pr\left(Y_{ij}=c|u_i, v_j\right)=\Phi\left(\alpha_c-\left(u_i+v_j\right)\right)-\Phi\left(\alpha_{c-1}-\left(u_i+v_j\right)\right)$, with W_{ij} distributed as $N(0,1)$ and Φ the cumulative distribution function (cdf) of the standard normal distribution. While a choice of link functions are available for the ordinal GLMM, the probit link function is a natural and appealing choice due to the continuous latent disease status assumption underlying the model [36], and for ease of mathematics. In Section Seven we demonstrate that similar results are obtained using a logit link function, another common choice of link function for ordered GLMMs.

Natural heterogeneity amongst subjects' test results is reflected in the subject random effect variance $\sigma^2{}_u$. Subjects with clearly defined disease will have a higher positive random effect term u_i and be classified more often by experts into a higher disease category with stronger agreement between experts; more modest values of u_i are observed in test results with less clearly defined disease status. The experts' random effect variance $\sigma^2{}_v$ is higher for a more heterogeneous group of experts; an expert who liberally (sparingly) assigns high disease categories has a higher positive (negative) value of v_j; experts who are not overly liberal or conservative in their ratings will have more modest values of v_j. The proposed model also allows for possible interactions between expert and subject random effect terms, as can be seen from an alternative derivation of the model (Appendix 2, Supplementary Materials).

Parameters of the ordered GLMM (2), $\boldsymbol{\theta}=\left(\alpha_{\mathbf{1}},\ldots,\alpha_{\mathbf{C-1}},\sigma_{\mathbf{u}}^{\mathbf{2}},\sigma_{\mathbf{v}}^{\mathbf{2}}\right)$, $\alpha_o = -\infty$ and $\alpha_C = +\infty$ provide valuable information about the agreement process under study, and are incorporated into the summary measure of agreement described in Section Four. Estimation of the parameter vector θ requires the marginal likelihood function of the corresponding ordinal GLMM, L(θ; y). Given the random effects u_i and v_j, and defining $d_{ijc} = 1$ if $y_{ij} = c$ and 0 otherwise, the ordered classification of the ith subject by the jth expert y_{ij} is a multinomial variable with probability mass function:

$$\prod_{c=1}^{C}\left[pr\left(Y_{ij}=c\right)\right]^{d_{ijc}}=\prod_{c=1}^{C}\left[\Phi\left(\alpha_c-\left(u_i+v_j\right)\right)-\Phi\left(\alpha_{c-1}-\left(u_i+v_j\right)\right)\right]^{d_{ijc}}.$$

The marginal likelihood function is:

$$L\left(\boldsymbol{\theta};\mathbf{Y}\right)=\int_{u,v}L\left(\boldsymbol{\theta};\mathbf{u},\mathbf{v},\mathbf{y}\right)du\quad dv=\int_{u,v}f_{Y|u,v}\left(y;u,v\right)f_u\left(u;\sigma_u^2\right)f_v\left(v;\sigma_v^2\right)du\quad dv$$
$$=\int_u\int_v\left[\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{c=1}^{C}\left[\Phi\left(\alpha_c-\left(u_i+v_j\right)\right)-\Phi\left(\alpha_{c-1}-\left(u_i+v_j\right)\right)\right]^{d_{ijc}}\right]\left[\prod_{i=1}^{I}\frac{1}{\sqrt{2\pi\sigma_u^2}}e^{\frac{-u_i^2}{2\sigma_u^2}}\right]\left[\prod_{j=1}^{J}\frac{1}{\sqrt{2\pi\sigma_v^2}}e^{\frac{-v_j^2}{2\sigma_v^2}}\right]du\quad dv$$

This likelihood does not have a closed form due to the high-dimensionality of the crossed random effects structure, also restricting use of adaptive quadrature as a fitting procedure which becomes computationally infeasible as the number of random effects increases [36-38]. However, multivariate Laplacian maximum likelihood approximation [39] is an efficient and valid approach for estimating the parameter vector θ, and is implemented in the R package *ordinal* used here for estimation purposes.

The covariance matrix Σ of the parameter estimates $\hat{\theta}$ is obtained as the inverse of the negative of the Hessian matrix H, where $H=\frac{\partial^2 l(\boldsymbol{\theta};\mathbf{u},\mathbf{v},\mathbf{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathbf{t}}}$ is the second-order derivative of the log-likelihood function $l(\theta; u, v, y)$ evaluated at the approximate maximum likelihood estimates of θ and is generated during the algorithmic multivariate Laplacian model-fitting process. The standard errors of $\hat{\theta}$ are obtained by taking the square-roots of the diagonals of H at convergence as $se\left(\hat{\theta}\right)=\sqrt{diag\left[-\left\{\mathrm{H}\left(\hat{\theta}\right)\right\}^{-1}\right]}$. In Section Five we demonstrate using simulation studies that this approach leads to unbiased estimators of the parameters of the ordered GLMM. In the next section, a summary measure of agreement is derived based upon this GLMM model.

4. Proposed Summary Measure of Agreement

4.1. A Model-Based Measure of Agreement

We propose a population-based summary measure of agreement κ_m based upon the ordinal GLMM model of agreement in Section Three. The proposed measure κ_m provides an overall assessment of chance-corrected agreement between many experts classifying the same sample of subjects' test results using an ordered categorical scale. It is a linear transformation of observed (exact) agreement p_0 (4), corrected for chance agreement p_c (4) and scaled to lie between 0 and 1 to allow for easy interpretation in a manner similar to Cohen's kappa [40] and Scott's pi [41]:

$$\kappa_m=\left(\frac{C}{C-1}\right)*\int_{-\infty}^{+\infty}\left\{\sum_{c=1}^{C}\left[\Phi\left(\frac{\Phi^{-1}\left(\frac{c}{C}\right)-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)-\Phi\left(\frac{\Phi^{-1}\left(\frac{c-1}{C}\right)-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\right]^{2}\right\}\phi\left(z\right)\quad dz-\frac{1}{C-1},\quad 0\leq\kappa_m\leq 1 \qquad (3)$$

where $\rho=\sigma_u^2/\left(\sigma_u^2+\sigma_v^2+1\right)$ and z is a N(0,1) variable. Within the summation expression in (3) for category c = 1 the second term in brackets is set to 0, and the first term in brackets for category c = C is set to 1. A value of κ_m close to 0 is interpreted as little or no agreement, a value around 0.5 suggests a moderate amount of agreement, and a value of 1 as perfect agreement between the multiple experts, after correcting for chance agreement [40]. The forms of observed agreement p_0 (proof in Appendix 3, Supplementary Materials) and chance agreement p_c are:

$$p_0=\int_{-\infty}^{+\infty}\left[\sum_{c=1}^{C}\left[\Phi\left(\frac{\alpha_c^*-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)-\Phi\left(\frac{\alpha_{c-1}^*-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\right]^{2}\right]\phi\left(z\right)\quad dz\quad \text{and}\quad p_c=\sum_{c=1}^{C}\left[\Phi\left(\alpha_c^*\right)-\Phi\left(\alpha_{c-1}^*\right)\right]^{2} \qquad (4)$$

with standardized thresholds $\alpha_c^*=\alpha_c/\sqrt{\sigma_u^2+\sigma_v^2+1}$. As shown in Theorem 1, the minimum value that chance agreement can take in this population-based setting is p_c = 1/C and is obtained when the standardized thresholds $(\alpha_0^*,\alpha_1^*\ldots,\alpha_c^*)$ take the values $\left(-\infty,\Phi^{-1}\left(\frac{1}{C}\right),\Phi^{-1}\left(\frac{2}{C}\right),\ldots,\Phi^{-1}\left(\frac{C-1}{C}\right),+\infty\right)$. These values divide the real line into C segments, each having equal probability under the standard normal curve, minimizing differences within the squared brackets in the right-hand side expression in (4). The values are then incorporated in the expression for κ_m to appropriately minimize the effects of chance agreement on the statistic; consequently the value of κ_m is not influenced by prevalence of the disease, which here is indicated by the percent of observations falling into each of the C categories.

For any particular dataset, the estimate $\hat{\kappa}_m$ is obtained from (3) using estimated random effects variances $\hat{\sigma}_v^2$ and $\hat{\sigma}_u^2$ from the fitted ordinal GLMM in (2). Estimates $\hat{p_0}$ and $\hat{p_c}$ of observed and chance agreement respectively can be obtained from (4) using the estimated parameters from the fitted ordinal GLMM in (2). We describe κ_m in (3) as a function of parameter ρ ($0 \leq \rho \leq 1$), which itself is a natural measure of the variability amongst subjects' test results, σ_u^2 relative to the overall variability present between classifications in a similar manner to the ICC (2) for multiple experts [16]. Values of ρ close to 0 suggest that variability between experts is greater relative to variability between the test results, while values of ρ closer to 1 suggest variability between test results is the dominating factor.

The variance of $\hat{\kappa}_m$ is derived using the multivariate delta method based upon parameters in the GLMM model with crossed random effects, where expert and subject random effects are assumed independent, with $var\left(\hat{\sigma}_u^2\right) = 2\left(\sigma_u^2\right)^2/I$ and $var\left(\hat{\sigma}_v^2\right) = 2\left(\sigma_v^2\right)^2/J$. The variance of $\hat{\rho}$ as a function of σ_v^2 and σ_u^2 is $var\left(\hat{\rho}\right) = \frac{2\left(\sigma_u^2\right)^2\left(\sigma_v^2+1\right)^2}{I\left(\sigma_u^2+\sigma_v^2+1\right)^4} + \frac{2\left(\sigma_v^2\right)^2\left(\sigma_u^2\right)^2}{J\left(\sigma_u^2+\sigma_v^2+1\right)^4}$. Since κ_m is a function of ρ, the delta method can be further applied leading to:

$$var\left(\hat{\kappa}_m\right) = \left(\frac{C}{C-1}\right)^2 * var\left(\hat{\rho}\right) * \left[\int_{-\infty}^{+\infty} \sum_{c=1}^{C} 2 * \left[\Phi\left(\frac{\Phi^{-1}\left(\frac{c}{C}\right) - z\sqrt{\rho}}{\sqrt{1-\rho}}\right) - \Phi\left(\frac{\Phi^{-1}\left(\frac{c-1}{C}\right) - z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\right] * \left[\phi\left(\frac{\Phi^{-1}\left(\frac{c}{C}\right) - z\sqrt{\rho}}{\sqrt{1-\rho}}\right) - \phi\left(\frac{\Phi^{-1}\left(\frac{c-1}{C}\right) - z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\left(\frac{-z}{2\sqrt{\rho(1-\rho)}} + \frac{\Phi^{-1}\left(\frac{c-1}{C}\right) - z\sqrt{\rho}}{2(1-\rho)^{3/2}}\right)\right] \phi(z)\, dz\right]^2.$$

Functions in R to calculate the proposed measure $\hat{\kappa}_m$ and variance $var\left(\hat{\kappa}_m\right)$ for a dataset are available upon request from the first author {or in supplementary materials**}.

4.2. Cohen's Kappa with Model-Based Parameters

Many of the existing summary measures of agreement for multiple raters classifying subjects according to an ordinal scale are either extensions of Cohen's kappa [1-5] or take the form of Cohen's kappa $\kappa = (p_0 - p_c)/(1 - p_c)$ obtaining terms p_0 and p_c using a model-based technique [16]. For comparison with our proposed summary measure of agreement, it is informative for us to also calculate a model-based kappa formulated as a Cohen's kappa-like statistic using our ordinal GLMM quantities of observed and chance agreement p_0 and p_c. This measure will be referred to as κ_{GLMM} and can be estimated using the estimated parameters $\hat{p_0}$ and $\hat{p_c}$ from the ordinal GLMM in (2):

$$\kappa_{GLMM} = \frac{p_0 - p_c}{1 - p_c}.$$

4.3. Other Existing Summary Measures for Agreement

Existing measures that can be used to assess agreement between multiple raters in the ordinal classification setting include Fleiss' kappa κ_F [3], Light's and Conger's kappa κ_{LC} [4,5] and Mielke et al's kappa κ_{MB} [42]. All of these measures are derivations or extensions of Cohen's kappa. These existing measures will be compared with the proposed measure κ_m in the following simulation studies and examples.

5. Simulation Studies

Performance of the ordinal GLMM (2) and proposed summary agreement measure (3) was investigated via extensive simulation studies. Sets of one thousand datasets were randomly generated based upon an ordinal GLMM with a crossed random effect structure with *C*=5 categories with a set of true parameter values $\boldsymbol{\theta}=\left(\boldsymbol{\alpha_1, \ldots, \alpha_{C-1}, \sigma_u^2, \sigma_v^2}\right)$, a_o = −∞ and a_C = +∞. For each dataset *I* subject random effects u_i and *J* expert random effects v_j were randomly generated from $N\left(0, \sigma_u^2\right)$ and $N\left(0, \sigma_v^2\right)$ respectively. The *rmultinom* function in R was used to randomly generate *n* = *I***J* (*I* ratings per expert) observations Y_{ij} based upon multivariate normal probabilities $Pr\left(Y_{ij}=c|u_i, v_j\right)=\Phi\left(\alpha_c-\left(u_i+v_j\right)\right)-\Phi\left(\alpha_{c-1}-\left(u_i+v_j\right)\right)$. The *clmm* function in the R package *ordinal* [43] was used to fit the GLMM model for each dataset to obtain parameter estimates, and the proposed summary measures and their variances estimated. Four sets of simulations were conducted to investigate effects of varying rater and subject variances, numbers of subjects and experts included, and extreme or moderate disease prevalence (based upon probabilities of being classified in high or low disease categories). A fifth set of simulations was conducted to explore the robustness of the proposed approach and measure κ_m to the assumption of normally distributed random effects. In this fifth set of simulations, one thousand datasets were generated with non-normal subject random effects u_i, *i* = 1,....,*I* randomly sampled from a scaled and centered chi-squared distribution [44] with five degrees of freedom $\left(\chi^2-5\right)/\sqrt{2{*}5/10}$ which has a mean of 0 and the same variance, 10, as in two of the other sets of simulations; the expert random effects v_j, *j*=1,...,*J* were randomly sampled from a Uniform($-\sqrt{3}, \sqrt{3}$) distribution yielding a mean of 0 and variance of 1, the same mean and variance as for all other sets of simulations. Conditional on the random effects u_i and v_j, the classifications Y_{ij} were assumed independent and generated according to probabilities based upon a multivariate normal distribution [44] in a similar manner to the first four sets of simulations. Simulation results, including the estimated means of the parameters and their standard errors (based upon the averages of the one thousand estimates) for the five simulation studies are presented in Table 1. Observed standard errors (calculated by taking the square-root of the variance of one thousand parameter estimates) are also presented.

The simulation studies demonstrate that the *ordinal* package in R yielded essentially unbiased parameter GLMM estimates of the threshold parameters ($\hat{\alpha}_1, \ldots, \hat{\alpha}_4$) and variance components ($\hat{\sigma}_u^2, \hat{\sigma}_v^2$) in both large and small sample sizes, and large ($\sigma_u^2=10$) and small ($\sigma_u^2=1$) random effects variances. In simulation set #2, the scenario with a large subject random effects variance (σ_u^2) and smaller total sample size (*n* = *I***J* = 500), the observed variability of the one thousand estimates of σ_u^2 is slightly larger than expected, but this issue was not observed at the larger sample size of *n* = 5000. The proposed measure $\hat{\kappa}_m$ and its variance were estimated essentially in an unbiased manner in all simulation settings. In the fifth set of simulations with non-normal random effects (simulation set #5), the ordinal GLMM thresholds $\hat{\alpha}_1, \ldots, \hat{\alpha}_4$ were estimated with more bias than in the sets of simulations with normally distributed random effects, although variance parameters were estimated on average with little or no bias. The proposed measure κ_m which is not reliant upon the estimated threshold values and its variance was estimated with essentially no bias. These results suggest that the proposed approach may be fairly robust to the normal random effects assumption and confirm results in earlier studies examining the impact of non-normal random effects in GLMMs [44,45], though further work is required to confirm these findings.

The behavior of the new measure of agreement κ_m was explored for a range of different settings including varying disease prevalence and ρ for an ordinal classification scale with five categories (C=5), with results presented in Figure 1. Comparisons were made with existing measures of agreement (Section 4.2) and a Cohen's kappa based upon GLMM parameters of agreement (Section 4.3). Figure 1 presents plots of the measures of agreement for increasing ρ and varying prevalence (extreme low or high, moderate, equal in each category as presented in Table 2). Fleiss' kappa, Light and Conger's kappa and Cohen's kappa based upon GLMM parameters all yielded virtually identical average values so only Fleiss' kappa κ_F is presented on the plots. True parameter values were used in the plots with the exception of κ_F, which was averaged over sets of 1000 simulated datasets. All agreement measures increased in value as ρ increased, and at a steeper rate as ρ approached 1. Thus, experts agree more often when there is a wider spread of test results (larger σ_u^2) relative to the variability between experts. All agreement measures took very similar values when disease prevalence (as indicated by the percent of observations in each category) was distributed evenly over the five categories. However, as disease prevalence became more extreme (either high or low), the proposed new measure of agreement κ_m remained unaffected, while the Cohen's kappa-based agreement measures (κ_{GLMM}, κ_F, and κ_{LC}), which are prone to prevalence effects, all increased in value.

6. Examples: Two Cancer Agreement Studies

6.1. Prostate Cancer Agreement Study

Allsbrook et al [46] conducted a study of agreement between 10 urologic pathologists each independently interpreting the severity of prostate cancer of 46 patients' biopsies using a condensed version of the Gleason Grading scale [7]. The scale had 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). Table 3(a) displays a subset of the ordinal classifications of the 46 patients' test results by each of the 10 urologists. Tables 3(b) and (c) present the observed classifications of selected pairs of urologists. An ordinal GLMM (2) was then fitted to the full dataset incorporating the dependency between the classifications of the experts via the crossed random effects structure using the *ordinal* package in R, taking less than 1 minute of computational time. Parameter estimates and summary measures are presented in Table 4. As an indicator of disease prevalence, the estimated probabilities of being classified into each of the four categories over the ten experts are: (from mild to severe cancer) 6%, 31%, 28% and 36%

respectively, thus there was a high probability of being categorized with a moderate to high degree of prostate cancer. Observed agreement (3) based upon the ordinal GLMM in (2) was estimated to be $\hat{p_0}$ = 0.669. The proposed new agreement measure $\hat{\kappa}_m=0.484$ (*se* = 0.035) indicated a moderate level of chance-corrected agreement between the ten urologists. This was substantially lower in value than Fleiss' kappa $\hat{\kappa}_F=0.569$, Conger's and Light's kappa, $\hat{\kappa}_{LC}=0.570$, and Cohen's GLMM-based kappa estimated at $\hat{\kappa}_{GLMM}=0.526$ (described in Sections 4.2 and 4.3), which may be attributed to the effects of high disease prevalence that Cohen's kappa measures are prone to as observed in Figure 1, while the proposed kappa is unaffected by disease prevalence. Mielke et al's kappa [42,47] which is based upon the probability that every one of the ten urologists classify a subject's test result into exactly the same category is estimated as 0.196; this kappa reflects the low chance that such classifications would occur very often in practice, and less so as the number of experts increases.

The original analysis of agreement presented in Allsbrook et al's paper [46] compared Cohen's kappas (unweighted) for each pair of urologists, yielding 45 pairwise Cohen's kappas ranging between 0.31 – 0.79. This approach leads to complexities in interpretation with limited overall conclusions about the agreement between the ten urologists. In contrast, our proposed approach allows classifications of all ten urologists to be analyzed and interpreted in one unified approach. Characteristics of individual urologists can also be examined through their estimated random effect terms $\hat{v_j}$, j=1,..., J if required.

6.2. Cervical Cancer Agreement Study

An early study was conducted by Holmquist et al [11] to examine the variability in the classifications of cervical cancer from histological slides and to assess the level of overall agreement among pathologists. Seven pathologists evaluated and classified 118 slides according to a five-category ordinal scale: (1) negative; (2) atypical squamous hyperplasia; (3) carcinoma in situ; (4) squamous carcinoma with early stromal invasion; and (5) invasive carcinoma. A table of individual classifications made by each pathologist is presented in Landis and Koch [40]. The original analysis presented in Holmquist et al [11] focused on examining the number of slides between each pair of pathologists that were rated as higher by one pathologist in the pair. Table Five presents the results from fitting the proposed ordinal GLMM model and summary measures and existing summary measures. Based upon all seven pathologists' classifications, the observed probabilities of being rated in categories 1 (negative) to 5 (most severe cancer) were 27%, 17%, 52%, 3% and 1.7% respectively, indicating only a small proportion of slides indicated more severe disease. The variability in classifications between the patient slides ($\sigma_u^2=4.13$) is large relative to the variability observed amongst the pathologists ($\sigma_v^2=0.627$) yielding a moderately high value of ρ = 0.717. The model-based kappa $\hat{\kappa}_m=0.266$ (*se* = 0.032), indicating only a low level of chance-corrected agreement between the seven pathologists. Each of the Cohen's kappa-based summary measures, Fleiss' kappa, Conger's and Light's kappa and Cohen's GLMM-based kappa yielded slightly higher estimated chance-corrected agreement ($\hat{\kappa}_F=0.354$, $\hat{\kappa}_{LC}=0.361$ and $\hat{\kappa}_{GLMM}=0.296$ respectively). This may be attributed, in a similar manner to the Gleason Grading example in Section 6.1, to prevalence effects, where severe disease (categories 4 and 5) is very rarely observed for this sample of slides. Mielke et al's kappa [42,47] is very small at 0.127, demonstrating the unlikely scenario that all seven pathologists would assign an identical classification to any particular slide.

7. A Logistic Generalized Linear Mixed Model for Modeling Agreement

The generalized linear mixed model with a logistic link function is often a popular choice when modeling ordinal classifications. In this section we demonstrate that almost identical results are obtained for observed agreement p_0 whether a probit or logit link function is employed in the ordinal GLMM framework for modeling agreement.

The logistic function $\Psi\left(\alpha_c-\left(u_i+v_j\right)\right)=1/\left(1+e^{-\left(\alpha_c-\left(u_i+v_j\right)\right)}\right)$ replaces the probit function $\Phi(\cdot)$ in the logistic ordinal GLMM. Wlog we set $\beta_0$to 0 for identifiability purposes, with variance of the logistic distribution $\pi^2/3$, and the logistic ordinal GLMM in terms of cumulative logits is:

$$\Psi^{-1}\left(Pr\left(Y_{ij}\le c|u_i,v_j\right)\right)=log\left(\frac{Pr\left(Y_{ij}\le c|u_i,v_j\right)}{1-Pr\left(Y_{ij}\le c|u_i,v_j\right)}\right)=\alpha_c-\left(u_i+v_j\right),\quad c=1,\ldots,C$$

with the probability of the *i*th subject's test result being classified by expert *j* into category *c* as $Pr\left(Y_{ij}=c|u_i,v_j\right)=\Psi\left(\alpha_c-\left(u_i+v_j\right)\right)-\Psi\left(\alpha_{c-1}-\left(u_i+v_j\right)\right)$. It is helpful to define a random variable Q with a logistic distribution with mean 0 and variance π^2 / 3, and density function $f_Q\left(q\right)=e^{-q}/\left(1+e^{-q}\right)^2$. Observed agreement p_0 is derived for the logistic ordinal GLMM as:

$$p_0=\sum_{c=1}^{C}Pr\left[\left(Y_{ij}=c\right)\cap\left(Y_{ij'}=c\right)\right]=\int_{-\infty}^{+\infty}\left\{\sum_{c=1}^{C}\left[F_{Q^*}\left(\frac{\alpha_c}{\sigma_u}-z\right)-F_{Q^*}\left(\frac{\alpha_{c-1}}{\sigma_u}-z\right)\right]^2\right\}\phi\left(z\right)\,dz$$

where $Q^*=\left(Q+v_j\right)/\sigma_u$. The cdf $F_{Q^*}\left(q^*\right)$ takes the form [48]:

$$F_{Q^*}\left(q^*\right)=Pr\left(Q^*\le q^*\right)=\int_{-\infty}^{+\infty}e^{-q}\left(1+e^{-q}\right)^{-2}\quad\Phi\left(\frac{q^*-q/\sigma_u}{\sigma_v/\sigma_u}\right)dq,\quad \text{leading to:}$$

$$p_0=\int_{-\infty}^{+\infty}\left\{\sum_{c=1}^{C}\left[\int_{-\infty}^{+\infty}e^{-q}\left(1-e^{-q}\right)^{-2}\quad\Phi\left(\frac{\alpha_c-\sigma_u z-q}{\sigma_v}\right)dq-\int_{-\infty}^{+\infty}e^{-q}\left(1-e^{-q}\right)^{-2}\quad\Phi\left(\frac{\alpha_{c-1}-\sigma_u z-q}{\sigma_v}\right)dq\right]^2\right\}\varphi\left(z\right)\,dz$$

This expression is evaluated using the *integrate* function in R. Figure 2 demonstrates that almost identical values of observed agreement p_0 are obtained under the ordinal GLMM framework irrespective of choice of probit or logit link function. It is also interesting to note that for both choices of link function, as disease prevalence becomes more extreme (high or low), model-based observed agreement increases; and as subjects' test results become more distinguishable from each other relative to the variability between experts (increasing ρ), observed agreement between experts also increases, though at a faster rate for prevalences that are more moderately spread over the *C* categories.

8. Discussion

In this paper we describe a comprehensive framework and propose a chance-corrected summary measure in a population-based setting to examine agreement between multiple experts classifying a sample of subjects' test results according to an ordered categorical scale. The proposed approach flexibly can accommodate large numbers of experts and subject test results without increasing complexity as the number of experts increases, while allowing for missing data. The proposed methods are easily implemented using the freely available software package R. Initial findings suggest the approach is fairly robust to non-normal random effects, though further work is needed in this area.

Limited alternative approaches exist to examine agreement in this large-scale setting with ordinal classifications, and many existing measures of agreement are prone to prevalence effects of the disease or condition under study. Due to a lack of available methods for studies involving ordinal classifications made by many experts, many agreement studies in the medical literature have instead elected to use approaches intended to assess agreement between just two experts at a time. For example, several pairwise Cohen's kappa statistics between all possible pairs of experts are often calculated, leading to complexities in interpretation and limited overall conclusions regarding the group of experts as a whole. Our approach allows agreement between all the experts to be assessed in one unified approach, lending power and efficiency to the study of agreement between the multiple experts when assumptions are met, and a simpler interpretation of results. If the experts and patients included in the study are randomly sampled from their respective populations, the results are generalizable to these populations of experts and patients, which is especially advantageous for widely used screening procedures. There is also the opportunity to examine the rating characteristics of individual experts through exploration of individual random effect components. In addition, our approach is not vulnerable to some of flaws observed when using Cohen's kappa, such as prevalence effects. Examining the effects of factors (such as rater training) that may play an influential role in the levels of agreement between experts, such as the levels of training and volume of tests read annually will be examined in a future paper.

In addition to measures of agreement, measures of association are also commonly used to compare the ordered categorical classifications of two or more experts. These measures incorporate valuable information regarding the extent of disagreement between experts, where disagreement arises when two experts assign a different classification to the same subject's test result. For example, stronger disagreement is implied when two experts' classifications are three categories apart rather than one or two categories apart. A common measure of association is Cohen's weighted kappa [2,14,49] which employs a weighting scheme which assigns partial "credit" for classifications not in exact agreement, with larger weights ("credit") assigned to pairs of classifications closer together. Extensions of Cohen's weighted kappa to more than two experts [50] has been described as problematic [42]. Gonin et al [51] describe a model-based approach using generalized estimating equations that generates a weighted kappa coefficient, including a fixed term for each expert, thus best-suited to a smaller number of experts. Extending the proposed population-based approach described in this paper for multiple experts to incorporate information about disagreement using measures of association is a topic for future research.

Due to the categorical nature of ordinal classifications, non-parametric rank-invariant approaches are preferred by some researchers [52-55], where rank invariant methods are not influenced by a relabeling of the ordinal classification scale [52-56]. Svensson et al [52,53] describe a non-parametric rank-invariant approach for evaluating the various components of disagreement in paired rank-invariant data. These approaches are currently restricted to assessing association and agreement between pairs of experts, so are better suited to studies with a smaller number of experts. Liu and Agresti [36] note that for parametric approaches, when the latent variable model holds, for example, where disease status is considered an unobserved latent variable, the estimated effects are invariant to the number of categories of the classification scale and their cutpoints, and when the model fits well, different studies employing different scales for the classifications should lead to similar conclusions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The authors are grateful for the support provided by grant 1R01CA17246301-A1 from the United States National Institutes of Health. We thank Dr Allsbrook for kindly providing us with his dataset. We appreciate the comments and suggestions provided by the reviewers of this manuscript.

REFERENCES

1. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960; 20:37–46.
2. Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin. 1968; 70(4):213–220. [PubMed: 19673146]
3. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971; 76:378–382.
4. Light RJ. Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychological Bulletin. 1971; 76:365–377.
5. Conger AJ. Integration and generalization of kappa for multiple raters. Psychological Bulletin. 1980; 88:322–328.
6. Gleason, DF. The Veteran's Administration Cooperative Urologic Research Group: histologic grading and clinical staging of prostatic carcinoma”.. In: Tannenbaum, M., editor. Urologic Pathology: The Prostate. Lea and Febiger; Philadelphia: 1977. p. 171-198.
7. Epstein JI, Allsbrook WC Jr, Amin MB, Egevad LL, ISUP Grading Committee. The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason grading of prostatic carcinoma. American Journal of Surgical Pathology 2005. 29(9):1228–42.
8. The Criteria Committee of the New York Heart Association. Nomenclature and Criteria for Diagnosis of Diseases of the Heart and Great Vessels. 9th edition. Little, Brown & Co.; Boston: 1994. p. 253-256.
9. Elmore JG, Jackson SL, Abraham L, Miglioretti DL, Carney PA, Geller BM, et al. Variability in interpretive performance at screening mammography and radiologists’ characteristics associated with accuracy. Radiology. 2009; 253(3):641–651. [PubMed: 19864507]

10. Miglioretti DL, Smith-Bindman R, Abraham L, et al. Radiologist characteristics associated with interpretive performance of diagnostic mammography. Journal of the National Cancer Institute. 2007; 99(24):1854–1863. [PubMed: 18073379]
11. Holmquist ND, McMahan CA, et al. Variability in classification of carcinoma in situ of the uterine cervix. Archives of Pathology. 1967; 84:334–345. [PubMed: 6045443]
12. Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. Journal of the National Cancer Institute. 2003; 95(4):282–290. [PubMed: 12591984]
13. Onega T, Smith M, Miglioretti DL, et al. Radiologist agreement for mammographic recall by case difficulty and finding type. Journal of the American College of Radiology. 2012; 9:788–794. [PubMed: 23122345]
14. Kraemer HC. Measurement of reliability for categorical data in medical research. Statistical Methods in Medical Research. 1992; 1(2):183–199. [PubMed: 1341657]
15. Gwet, KL. Advanced Analytics. Third Edition.. LLC; Maryland: 2012. Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among multiple raters..
16. Williamson JM, Manatunga AK, Lipsitz SR. Modeling kappa for measuring dependent categorical agreement data. Biostatistics. 2000; 1(2):191–202. [PubMed: 12933519]
17. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. American Journal of Epidemiology. 1987; 126:161–169. [PubMed: 3300279]
18. Mielke PW, Berry KJ, Johnston JE. The exact variance of weighted kappa with multiple raters. Psychological Reports. 2007; 101:655–660. [PubMed: 18175509]
19. Tanner MA, Young MA. Modeling agreement between raters. Journal of the American Statistical Association. 1985; 80:175–180.
20. Agresti A. A model for agreement between ratings on an ordinal scale. Biometrics. 1988; 44:539–548.
21. Agresti A. Modelling patterns of agreement and disagreement. Statistical Methods in Medical Research. 1992; 1:201–218. [PubMed: 1341658]
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. Johnson, VE.; Albert, JH. Ordinal data modeling. Statistics for Social Science and Public Policy. Springer-Verlag; New York Inc: 1999.
24. Nelson JC, Pepe MS. Statistical description of interrater variability in ordinal ratings. Statistical methods in medical research. 2000; 9:475–496. [PubMed: 11191261]
25. Banerjee M, Capozzoli M, McSweeney L, et al. Beyond kappa: a review of interrater agreement measures. The Canadian Journal of Statistics. 1999; 27(1):3–23.
26. Hsaio CK, Chen P-C, Kao W-H. Bayesian random effects for interrater and test-retest reliability with nested clinical observations. Journal of Clinical Epidemiology. 2011; 64:808–814. [PubMed: 21292442]
27. Nelson KP, Edwards D. On population-based measures of agreement for binary classifications. Canadian Journal of Statistics. 2008; 36(3):411–426.
28. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational Psychological Measures. 1973; 33:613–619.
29. Coull B, Agresti A. Generalized log-linear models with random effects, with application to smoothing contingency tables. Statistical Modelling. 2003; 0:1–21.
30. Bloch DA, Kraemer HC. 2 x 2 kappa coefficients: measurements of agreement or association. Biometrics. 1989; 45(1):269–287. [PubMed: 2655731]
31. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association. 1993; 88:9–25.
32. Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. Biometrics. 1994; 50:933–944. [PubMed: 7787006]
33. Agresti, A. Analysis of ordinal categorical data. 2nd Edition.. John Wiley & Sons; 2010.
34. Ibrahim JG, Molenberghs G. Missing data issues in longitudinal studies: a review. Test (Madr). 2009; 18(1):1–43. [PubMed: 21218187]

35. Nelson KP, Edwards D. Improving the reliability of diagnostic tests in population- based agreement studies. Statistics in Medicine. 2010; 29:617–626. [PubMed: 20128018]

36. Liu I, Agresti A. The analysis of ordered categorical data: an overview and a survey of recent developments. Test. 2005; 14(1):1–73.

37. Gueorguieva R. A multivariate generalized linear mixed model for joint modeling of clustered outcomes in the exponential family. Statistical Modelling. 2001; 1:177–193.

38. Capanu M, Gönen M, Begg CB. An assessment of estimation methods for generalized linear mixed models with binary outcomes. Statistics in Medicine. 2013; 32(26):4550–4566. [PubMed: 23839712]

39. Shun Z, McCullagh P. Laplace approximation of high dimensional integrals. Journal of the Royal Statistical Society Series B. 1995; 57:749–760.

40. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159–174. [PubMed: 843571]

41. Scott WA. Reliability of content analysis: the case of nominal scale coding. Public Opinion Quarterly. 1955; 19:321–325.

42. Mielke PW, Berry KJ, Johnston JE. Unweighted and weighted kappa as measures of agreement for multiple judges. International Journal of Management. 2009; 26(2):213–223.

43. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2014.

44. Papageorgiou G, Hinde J. Multivariate generalized linear mixed models with semi- nonparametric and smooth nonparametric random effects densities. Statistical Computing. 2012; 22:79–92.

45. Chen J, Zhang D, Davidian M. A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. Biostatistics. 2002; 3(3):347–360. [PubMed: 12933602]

46. Allsbrook WC, Mangold KA, Johnson MH, Lane RB, Lane CG, Amin MB, et al. Interobserver reproducibility of Gleason grading of prostatic carcinoma: Urologic pathologists. Human Pathology. 2001; 32(1):74–80. [PubMed: 11172298]

47. Berry KJ, Johnston JE, Mielke PW. Weighted kappa for multiple raters. Perceptual and Motor Skills. 2008; 107:837–858. [PubMed: 19235413]

48. Nadarajah S. Linear combination, product and ratio of normal and logistic random variables. Kybernetika. 2005; 41(6):787–798.

49. Graham P, Jackson R. The analysis of ordinal agreement data: beyond weighted kappa. Journal of Clinical Epidemiology. 1993; 46(9):1055–1062. [PubMed: 8263578]

50. Warrens MJ. Inequalities between multi-rater kappa. Advances in Data Analysis and Classification. 2010; 4:271–286.

51. Gonin R, Lipsitz SR, Fitzmaurice GM, et al. Regression modelling of weighted kappa by using generalized estimating equations. Journal of the Royal Statistical Society Series C - Applied Statistics. 2000; 49:1–18.

52. Svensson E, Holm S. Separation of systematic and random differences in ordinal rating scales. Statistics in Medicine. 1994; 13:2437–2453. [PubMed: 7701145]

53. Svensson E. Different ranking approaches defining association and agreement measures of paired ordinal data. Statistics in Medicine. 2012; 31:3104–3117. [PubMed: 22714677]

54. Ledenius K, Svensson E, et al. A method to analyze observer disagreement in visual grading studies: example of assessed image quality in paediatric cerebral multidetector CT images. The British Journal of Radiology. 2010; 83:604–611. [PubMed: 20335429]

55. Hand DJ. Statistics and the theory of measurement. Journal of the Royal Statistical Society: Series A. 1996; 159:445–492.

56. Kuruppumullage P, Sopriyarachchi R. Log-linear models for ordinal multidimensional categorical data. Journal of the National Science Foundation Sri Lanka. 2007; 35(1):29–40.

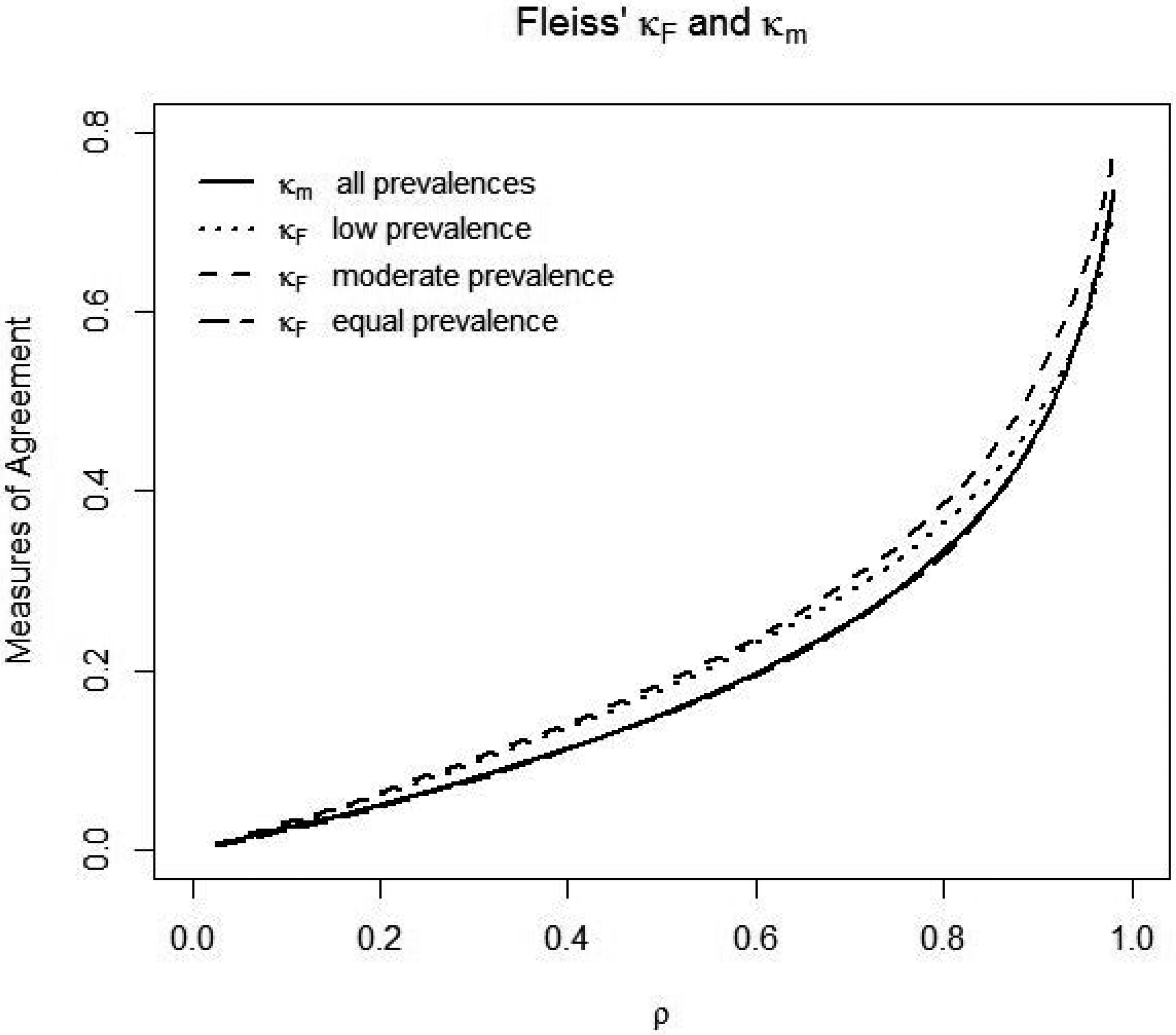
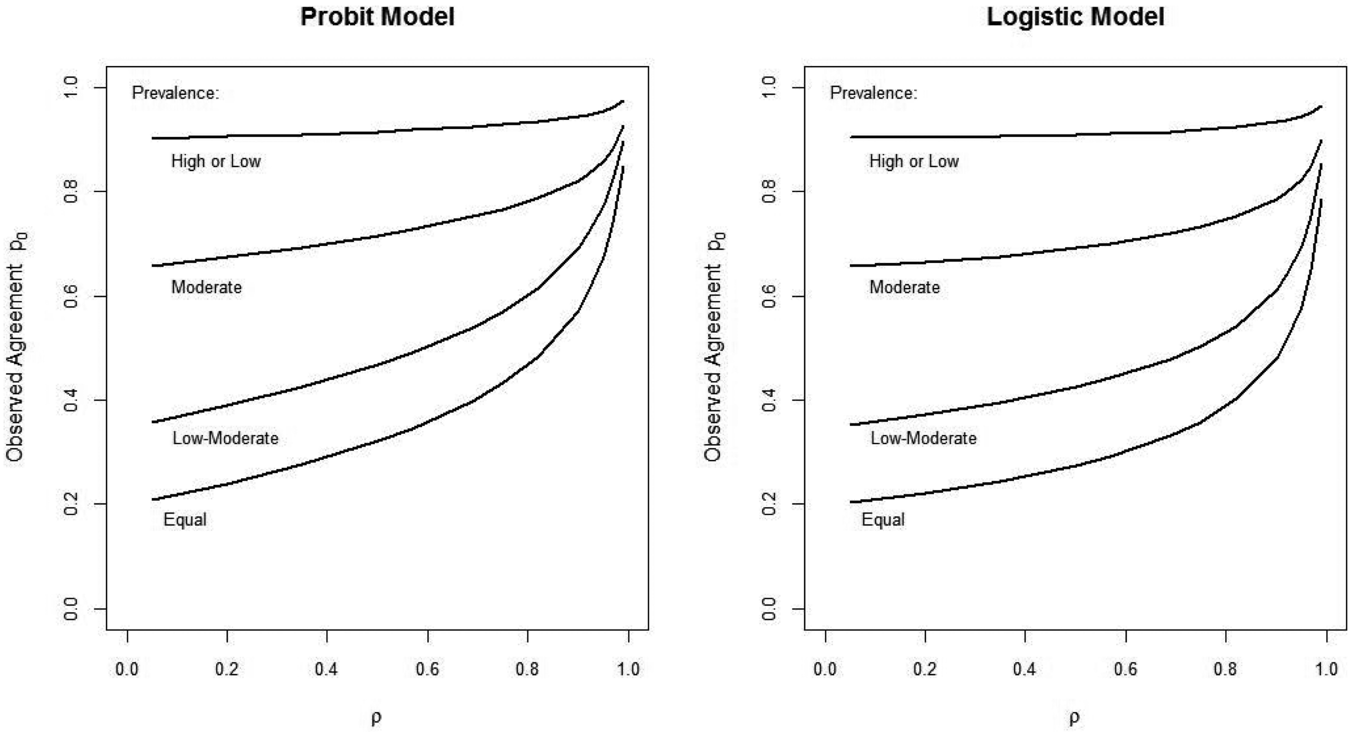


Figure (i).

Plots of agreement measures, proposed κm and κF versus ρ for varying prevalence (extreme low or high, moderate, equal in each category; the percent of observations falling into each of the Ci categories, i=1,...,5 in each prevalence case are presented in Table 2) with σ2v set to 1 and σ2u increasing in value.



Figure(2).

Model-based observed agreement p0 for probit and logistic ordinal GLMMs versus increasing ρ and varying prevalence of disease for an ordinal classification scale with five categories (C=5).

Table (i)

Results from sets of five simulation studies, each based upon 1000 datasets simulated from an ordinal GLMM with five classification categories (C=5), with thresholds $\alpha_0 = -\infty$ and $\alpha_5 = +\infty$.

	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.)
n	500			500			5000			5000			5000		
I	50			50			100			100			100		
J	10			10			50			50			50		
C	5			5			5			5			5		
α_1	0	−0.001	0.340 (0.354)	0	0.022	0.565 (0.657)	0	0.007	0.173 (0.175)	0	0.018	0.349 (0.353)	0	0.239	0.352 (0.347)
α_2	1	0.995	0.344 (0.358)	1	1.020	0.570 (0.667)	1	1.009	0.174 (0.175)	1	1.018	0.349 (0.356)	1	1.243	0.353 (0.347)
α_3	2	2.003	0.354 (0.365)	2	2.024	0.578 (0.668)	2	2.009	0.176 (0.178)	2	2.020	0.350 (0.354)	2	2.239	0.355 (0.350)
α_4	3	3.022	0.382 (0.399)	3	3.019	0.590 (0.691)	3	3.011	0.180 (0.181)	3	3.021	0.352 (0.353)	3	3.234	0.357 (0.351)
Random effects u_i	$N(0, \sigma_u^2)$			$N(0, \sigma_u^2)$			$N(0, \sigma_u^2)$			$N(0, \sigma_u^2)$			$(\chi^2 - 5)$		
σ_u^2	1	1.017	0.269 (0.279)	10	9.330	2.602 (3.936)	1	0.999	0.151 (0.141)	10	9.469	1.571 (1.485)	10	9.701	1.610 (1.692)
Random effects v_j	$N(0, \sigma_v^2)$			$N(0, \sigma_v^2)$			$N(0, \sigma_v^2)$			$N(0, \sigma_v^2)$			$Unif(-\sqrt{3}, \sqrt{3})$		
σ_v^2	1	0.939	0.450 (0.461)	1	0.985	0.487 (0.504)	1	0.985	0.204 (0.208)	1	0.993	0.208 (0.212)	1	1.000	0.210 (0.146)
ρ	0.333	0.346	0.063 (0.075)	0.833	0.816	0.044 (0.050)	0.333	0.335	0.038 (0.041)	0.833	0.824	0.025 (0.027)	0.833	0.826	0.025 (0.027)
var(ρ)	0.004	0.006		0.002	0.002		0.0015	0.0017		0.0006	0.0007		0.0006	0.0007	
κ_m	0.090	0.095	0.021 (0.025)	0.368	0.356	0.040 (0.055)	0.090	0.091	0.012 (0.013)	0.368	0.360	0.024 (0.029)	0.368	0.362	0.024 (0.029)
var(κ_m)	0.0005	0.0005 (0.0006)		0.002	0.002 (0.003)		0.0002	0.0002 (0.0002)		0.0006	0.0008 (0.0300)		0.0006	0.0006 (0.0009)	

Table (ii)

Varying levels of disease prevalence examined in simulation studies and figures 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**
Very low	80%	10%	3.4%	3.3%	3.3%
Moderately low	50%	26%	16%	6%	2%
Equal	20%	20%	20%	20%	20%
Moderately High	2%	6%	16%	26%	50%
Very High	3.3%	3.3%	3.4%	10%	80%

Table (iiia)

Table of classifications by individual raters for the Gleason Grading study [46]. Based upon an ordinal classification scale with C=4 categories; I=46 patient biopsies (cases); J=10 experts.

	Classifications by Individual Experts (J = 10)									
	Expert									
Case	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	2	2	2	2	2	2	2	2	2	2
2	2	3	4	3	3	3	2	3	3	3
3	4	4	4	4	4	3	4	4	3	4
4	4	4	4	4	4	4	4	3	4	4
5	2	2	2	2	2	2	1	2	2	2
6	3	3	3	3	3	3	3	3	3	3
7	2	3	3	2	3	3	2	2	2	2
8	3	3	3	3	3	4	2	3	3	3
9	3	3	4	3	4	4	4	3	3	4
10	3	3	4	3	3	4	3	3	3	3
⋮			⋮				⋮			⋮
41	3	3	4	3	3	4	4	3	3	4
42	2	2	3	2	2	2	2	2	2	2
43	2	2	2	2	2	5	2	2	2	2
44	4	4	4	4	4	4	4	4	4	4
45	3	3	3	3	3	3	4	3	3	3
46	3	3	3	3	3	3	3	3	2	3

Table (iii b and c)

Tables showing pairwise agreement between some randomly selected pairs of urologists classifying 46 slides [46] according to an ordinal scale based upon the Gleason grading scores with *C*=4 categories.

		Rater Three				
	Category	**1**	**2**	**3**	**4**	**Total**
Rater One	1	1	5	0	0	6
	2	0	6	4	4	14
	3	0	0	5	5	10
	4	0	0	0	16	16
	Total	1	11	9	25	46

		Rater Six				
	Category	**1**	**2**	**3**	**4**	**Total**
Rater Two	1	0	0	0	0	0
	2	2	13	0	0	15
	3	0	2	10	6	18
	4	0	0	3	10	13
	Total	2	15	13	16	46

Table (iv)

Results for the Gleason Grading Agreement Study [46] based upon an ordinal classification scale with *C*=4 categories; *I*=46 patient biopsies; *J*=10 urologists.

Parameter	Symbol	Estimate	S.E.	Z-value
Ordinal GLMM:				
Thresholds: ($\alpha_0 = -\infty$, $\alpha_4 = +\infty$)				
Between categories 1 and 2	α_1	−5.226	0.641	−8.154
Between categories 2 and 3	α_2	−1.258	0.522	−2.411
Between categories 3 and 4	α_3	1.549	0.521	2.971
Subject Random Effect variance	σ_u^2	9.295	2.453	
Rater Random Effect variance	σ_v^2	0.358	0.200	
Rho	ρ	0.873	0.027	
GLMM-based Observed Agreement	p_0	0.669		
Agreement Measures:				
Model-based Kappa	κ_m	0.484	0.035	
Fleiss’ kappa	κ_F	0.569	0.014	
Light and Conger's kappa	κ_{LC}	0.570		
Mielke's et al kappa		0.196		
Cohen's GLMM-based kappa	κ_{GLMM}	0.526		

Table (v)

Results for the Holmquist et al's [11] Carcinoma in situ of the uterine cervix Agreement Study based upon an ordinal classification scale with C=5 categories; I=118 patient slides; J=7 pathologists.

Parameter	Symbol	Estimate	S.E.	Z-value
<u>Ordinal GLMM:</u>				
Thresholds: ($\alpha_0 = -\infty$, $\alpha_5 = +\infty$)				
Between categories 1 and 2	α_1	−1.364	0.364	−3.747
Between categories 2 and 3	α_2	0.370	0.361	1.024
Between categories 3 and 4	α_3	2.856	0.376	7.605
Between categories 4 and 5	α_4	4.214	0.407	10.347
Subject Random Effect variance	σ_u^2	4.130	0.684	
Rater Random Effect variance	σ_v^2	0.627	0.348	
Rho	ρ	0.717	0.049	
GLMM-based Observed Agreement	p_0	0.485		
<u>Agreement Measures:</u>				
Model-based Kappa	κ_m	0.266	0.032	
Fleiss’ kappa	κ_F	0.354	0.012	
Light and Conger's kappa	κ_{LC}	0.361		
Mielke's et al kappa		0.127		
Cohen's GLMM-based kappa	κ_{GLMM}	0.296		