Published in final edited form as:

Biom J. 2020 November ; 62(7): 1687–1701. doi:10.1002/bimj.201900177.

Measuring intrarater association between correlated ordinal ratings

Kerrie P. Nelson¹, Thomas J. Zhou¹, Don Edwards²

¹Department of Biostatistics, Boston University, Boston, MA, USA

²Department of Statistics, University of South Carolina, Columbia, SC, USA

Abstract

Variability between raters' ordinal scores is commonly observed in imaging tests, leading to uncertainty in the diagnostic process. In breast cancer screening, a radiologist visually interprets mammograms and MRIs, while skin diseases, Alzheimer's disease, and psychiatric conditions are graded based on clinical judgment. Consequently, studies are often conducted in clinical settings to investigate whether a new training tool can improve the interpretive performance of raters. In such studies, a large group of experts each classify a set of patients' test results on two separate occasions, before and after some form of training with the goal of assessing the impact of training on experts' paired ratings. However, due to the correlated nature of the ordinal ratings, few statistical approaches are available to measure association between raters' paired scores. Existing measures are restricted to assessing association at just one time point for a single screening test. We propose here a novel paired kappa to provide a summary measure of association between many raters' paired ordinal assessments of patients' test results before versus after rater training. Intrarater association also provides valuable insight into the consistency of ratings when raters view a patient's test results on two occasions with no intervention undertaken between viewings. In contrast to existing correlated measures, the proposed kappa is a measure that provides an overall evaluation of the association among multiple raters' scores from two time points and is robust to the underlying disease prevalence. We implement our proposed approach in two recent breast-imaging studies and conduct extensive simulation studies to evaluate properties and performance of our summary measure of association.

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

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "Reproducible Research" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially for data confidentiality reasons.

Keywords

association; generalized linear mixed model; model-based kappa; ordinal classifications; screening test

1 | INTRODUCTION

Subjective assessment is an unavoidable element of the screening process for many medical conditions such as psoriasis (Gonin, Lipsitz, Fitzmaurice, & Molenberghs, 2000), Alzheimer's disease (Chaves et al., 2007), and cancer (Ciatto et al., 2006; Elmore et al., 2015; Gard, Aiello Bowles, Miglioretti, Taplin, & Rutter, 2015; Kerlikowske et al., 1998), often leading to excessive variability in clinical outcome measurements (Sadler, Yamamoto, Khurana, & Dallabrida, 2017). Inconsistencies are reported not only between different experts' ratings (interrater agreement) but also between an expert's own repeated ratings (reproducibility or intrarater agreement) of patients' test results. Poor intrarater and interrater agreement between raters in clinical settings can lead to added complexity when making clinical decisions and has further implications for patient care. In efforts to improve raters' interpretive performance, studies are often conducted in clinical settings to investigate the effectiveness of a new training tool, such as teaching, seminars, online, and practical instruction (Antonio & Crispi, 2010; Grassano, Pagana, Daperno, Bibbona, & Gasparini, 2018; Sadler et al., 2017).

Correlated ratings naturally arise as part of these studies, where a common choice of study design involves a group of raters scoring the same set of patients' test results twice using a binary or ordinal rating scale. When some form of rater training is undertaken between viewings, the goal is to assess whether the training intervention modifies a rater's classifications of patients' test results, and to compare levels of interrater association between raters before versus after training. If no intervention occurs between viewings, the goal is to assess the consistency between a rater's paired assessments, or intrarater agreement and reproducibility. In both settings, dependencies are induced between classifications due to the same raters assigning scores repeatedly to the same set of patients' test results. For simplicity, for the remainder of the paper, we will refer more generally to either setting as intrarater association.

A large literature exists on the use of correlated kappa statistics for assessing inter- and intrarater agreement and association. Earlier approaches focused on the correlated binary ratings of only two raters (Baker, Freedman, & Parmar, 1991; Donner, Shoukri, Klar, & Bartfay, 2000; Gwet, 2016) and are based on Cohen's kappa (Cohen, 1960, 1968) and Scott's (1955) π . Scott's π is equivalent to the intraclass correlation coefficient (ICC; Bloch & Kraemer, 1989; Vanbelle, Mutsvari, Declerck, & Lesaffre, 2012). Barnhart and Williamson (1997) describe testing the equality of dependent intraclass kappa statistics using weighted least squares for two raters for binary and ordinal ratings. Generalized estimating equation (GEE) modeling approaches were developed by Williamson and Manatunga (1997) and Williamson, Lipsitz, and Manatunga (2000) for modeling correlated categorical data for interrater agreement studies, while Gonin et al. (2000) propose a GEE

approach with a weighted Cohen's kappa for ordinal ratings and covariates to study interrater association. These approaches can be implemented in studies with a small fixed number of raters. In practice, many of these approaches are difficult to implement due to a lack of availability in statistical software packages, or become increasingly complex for more than two or three experts, limiting their usefulness in large studies involving many experts. In practice, clinical papers often resort to the simplest approach of reporting intrarater Cohen's kappas for each individual rater (Choa & Giele, 2015), though this becomes rapidly infeasible with a large number of raters.

More recently, Hsaio, Chen, and Kao (2011) and Tsai (2012) proposed Bayesian hierarchical modeling approaches based on nested generalized linear mixed models (GLMMs; Breslow & Clayton, 1993) to flexibly assess inter- and intrarater agreement for many raters' binary ratings using correlation coefficients and ICCs as summary measures. Further, Cao, Sen, Peery, and Dellon (2016) and Grassano et al. (2018) describe approaches based on correlated Fleiss' kappas (1971) to compare interrater agreement of raters' categorical ratings between repeated viewings of patient images taking into consideration the dependencies that arise because the same sample of patients is viewed on two occasions by the same sample of raters. These approaches are prone to drawbacks of Cohen and Fleiss' kappas, including bias due to the underlying disease prevalence and overcorrection of chance agreement (Banerjee, 1999; Brennan & Prediger, 1981; Maclure & Willett, 1987; Thompson & Walter, 1988).

Our proposed paired kappa fills an important gap in the statistical literature to provide summary measures of intrarater agreement and association in large clinical studies when many raters (more than two) contribute repeated ordinal ratings on the same sample of patients' test results on two separate occasions. When raters undertake training between viewings, the impact of the intervention on a rater's paired assessments can be examined. Dependencies between the correlated ordinal ratings are flexibly modeled in a GLMM model framework, which allows for missing ratings. In contrast to earlier approaches we do not assume the use of Cohen's kappa or Fleiss' kappa. The new proposed kappa possesses many desirable properties. Our approach can flexibly incorporate many raters simultaneously and a binary or ordinal rating scale without increasing the complexity of modeling. As a population-based approach, it is required that a minimum of three raters contribute ratings when calculating the measures of intrarater agreement and association. Association (or agreement) due to chance is appropriately accounted for, and the kappa is demonstrated to be robust to the underlying prevalence of disease in the patient population. Our approach focuses on population-based inference for the effects of training and intrarater agreement and association for the underlying population of raters and patients under study when randomly sampled. Furthermore, our modeling approach allows for the scoring behavior of individual raters at each time point to be assessed.

Measures of association are usually advocated when an ordinal grading scale is used (Banerjee, 1999; Gonin et al., 2000; Graham & Jackson, 1993). Although a measure of agreement focuses only on pairs of ratings assigned identical categories, measures of association including Cohen's (1968) weighted kappa provides credit for pairs of observations in "near agreement," with diminishing credit given for pairs of ratings that are further apart. Due to these differences, measures of agreement and association can take

different values for the same dataset, even yielding poor agreement and strong association (Graham & Jackson, 1993). Here we describe our proposed paired kappa for intrarater association, and also its usage as a measure of (exact) intrarater agreement.

We proceed by first introducing the GLMM model and notation that forms the basis of the proposed paired kappa measure. In Section 3, we derive the paired kappa for association, while an evaluation of the proposed kappa via simulation studies and properties of the kappa are presented in Section 4 and checking the goodness of fit is described in Section 5. Applications to two recent breast-imaging studies including a breast-screening MRI longitudinal study conducted at three time points and a large-scale breast density study are presented in Sections 6 and 7. We conclude with discussion in Section 8.

2 | A GENERALIZED LINEAR MIXED MODEL FOR PAIRED ORDINAL RATINGS

An ordinal GLMM model with a crossed random effects structure flexibly accommodates the dependent ordinal ratings of many raters to assess intrarater association and the effects of rater training. In our setting, we assume that the true disease severity of a patient is an unobserved continuous latent variable. The ordinal GLMM model uses a probit link function to reflect the underlying continuous disease, and creates a categorized version of the latent variable, with ordinal categories of increasing disease severity (Agresti, 2010; Liu & Agresti, 2005). Cutoff points along the real line are estimated from the data to define the boundaries of each ordinal category for each viewing. Further, our use of the probit link function leads to smoother mathematical tractability when deriving the measure of association (Nelson & Edwards, 2015, 2018). We assume that J raters ($j = 1, \dots, J$) each independently assess a sample of I items ($i = 1, \dots, I$) using an ordinal grading scale with C categories ($c = 1, \dots, C$) at two fixed time points ($k = 1, 2$) yielding $I \times J \times 2$ ordinal ratings $Y_{ijk} = c$. The ordinal GLMM model takes the form

$$P(Y_{ijk} \leq c \mid u_i, v_{kj}) = \Phi(\alpha_c - (\beta x_k + u_i + v_{kj})) \quad k = 1\text{st and } 2\text{nd viewings}, \quad (1)$$

where threshold terms α_c ($c = 1, \dots, C - 1$) represent the cutoffs defining the ordinal grading scale with $\alpha_0 = -\infty$ and $\alpha_C = +\infty$. Covariate x_k takes the value 0 for an assessment at the first viewing and 1 for the second viewing. The fixed coefficient β provides for an overall adjustment in the distribution of ratings at the second viewing as needed. A Wald hypothesis test, $H_0: \beta = 0$ with test statistic $z = (\hat{\beta} - 0)/SE(\hat{\beta})$ can be used to test whether β significantly differs from 0, which would indicate a shift in the distribution of classifications between the two sets of readings. Patient random effects u_i ($i = 1, \dots, I$) are assumed independent and normally distributed with 0 mean and variance σ_u^2 . This random effect u_i describes the visibility of disease for the i th patient's test result, which is viewed at both time points. Based on the ordinal GLMM in Equation (1) a large positive value of u_i suggests that the i th patient's test result is assigned a higher probability of being scored in the higher (more severe) disease categories (e.g., $c = 3$ or 4 when a four-category ordinal scale is being used for rating) across all raters and viewings assuming constant values of other parameters. This signals a patient's test result whose disease status is more easily observed relative to other

patients' test results that have smaller values of u_i . As the same rater views each patient's test result twice, the paired rater random effects at the two time points v_{1j} and v_{2j} ($j = 1, \dots, J$) are assumed correlated. A large positive value of v_{kj} indicates that the j th rater has a higher probability of assigning a score to patients' test results at the k th viewing in the higher (more severe) disease across all items (assuming a constant random effect value u_i), relative to other raters. The distribution of the paired rater random effects for a rater is

$$\begin{pmatrix} v_{1j} \\ v_{2j} \end{pmatrix} \sim \text{Bivariate normal} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_v = \begin{bmatrix} \sigma_{v1}^2 & \rho_v \sigma_{v1} \sigma_{v2} \\ \rho_v \sigma_{v1} \sigma_{v2} & \sigma_{v2}^2 \end{bmatrix} \right). \quad (2)$$

3 | PROPOSED PAIRED KAPPA OF ASSOCIATION

Measures of association are commonly reported for ordinal ratings as they provide helpful insight about the strength of agreement and near agreement between pairs of ratings (Graham & Jackson, 1993). An association measure may be presented alone or in conjunction with an agreement measure. Bloch and Kraemer (1989) define a well-defined measure of agreement, as a distinct kind of association, as how well a rater's observation of a patient agrees with the same rater's report at another time. Although a measure of agreement focuses only on ratings with identical ordinal scores (e.g., a patient's two scores are both "3"), a measure of association provides weight, or credit, to pairs of classifications scored into adjacent and near-adjacent categories (e.g., if a patient's paired scores are a "2" and a "3").

For a measure of association, the highest credit is usually assigned to pairs of ratings in exact agreement, with pairs of discordant ratings further apart assigned decreasing weights. This is in contrast to a measure of agreement that assigns credit only to pairs of ratings in exact agreement, and ignores all pairs of ratings that disagree. Although any weighting scheme can be applied, two commonly used weighting schemes are linear weights $w_{rs} = 1 - |r - s| / (C - 1)$ and quadratic weights $w_{rs} = 1 - (r - s)^2 / (C - 1)^2$ for a rater's pairs of ratings on a patient in the r th and s th categories, respectively, at the two time points ($r, s = 1, \dots, C$; Graham & Jackson, 1993). Choosing a conventional weighting scheme allows for comparison across different studies. We first define two important components of the proposed kappa measure, observed and chance intrarater association.

3.1 | Measures of observed and chance intrarater association

Observed intrarater association $p_{0, \text{assoc}}$ describes how often a rater classifies a pair of patient's test results into the r th and s th categories, respectively ($r, s = 1, \dots, C$), of the ordinal grading scale, weighted by the chosen weighting scheme. When $r = s$, the rater assigns an identical rating to a patient's test results at both viewings. On the other hand, chance intrarater association $p_{c, \text{assoc}}$ describes how often a rater assigns the r th and s th categories, respectively, to two different patient test results at different time points k and k' (i.e., simply by coincidence because these are different patients), also weighted by the chosen weighting scheme. The expressions in Equation (3) hold in a general population-based setting and are not reliant on any particular statistical model:

$$\begin{aligned}
 p_{0, \text{assoc}} &= \sum_{r=1}^C \sum_{s=1}^C w_{rs} [P(Y_{ijk} = r \cap Y_{ijk'} = s)] \text{ and } p_{c, \text{assoc}} \\
 &= \sum_{r=1}^C \sum_{s=1}^C w_{rs} [P(Y_{ijk} = r \cap Y_{i'jk'} = s)].
 \end{aligned} \tag{3}$$

The measures of observed and chance intrarater association are weighted averages of probabilities and take values between 0 and 1. Under the ordinal GLMM model in Equation (1) observed intrarater association takes the form:

$$\begin{aligned}
 p_{0, \text{assoc}} &= \int_{z_u} \int_{z_{v1}} \sum_{r=1}^C \sum_{s=1}^C w_{rs} \left[\Phi(\alpha_r - \sigma_u z_u - \sigma_{v1} z_{v1}) \right. \\
 &\quad \left. - \Phi(\alpha_{r-1} - \sigma_u z_u - \sigma_{v1} z_{v1}) \right] \times \left[\Phi \left(\frac{\alpha_s - \beta - \sigma_u z_u - \rho_v \sigma_{v2} z_{v1}}{\sqrt{1 + \sigma_{v2}^2 (1 - \rho_v^2)}} \right) \right. \\
 &\quad \left. - \Phi \left(\frac{\alpha_{s-1} - \beta - \sigma_u z_u - \rho_v \sigma_{v2} z_{v1}}{\sqrt{1 + \sigma_{v2}^2 (1 - \rho_v^2)}} \right) \right] f(z_{v1}) f(z_u) dz_{v1} dz_u,
 \end{aligned} \tag{4}$$

while chance intrarater association is defined as

$$\begin{aligned}
 p_{c, \text{assoc}} &= \int_{z_{v1}} \sum_{r=1}^C \sum_{s=1}^C w_{rs} \left[\Phi \left(\frac{\alpha_r - \sigma_{v1} z_{v1}}{\sqrt{1 + \sigma_u^2}} \right) - \Phi \left(\frac{\alpha_{r-1} - \sigma_{v1} z_{v1}}{\sqrt{1 + \sigma_u^2}} \right) \right] \\
 &\quad \times \left[\Phi \left(\frac{\alpha_s - \beta - \rho_v \sigma_{v2} z_{v1}}{\sqrt{1 + \sigma_u^2 + \sigma_{v2}^2 (1 - \rho_v^2)}} \right) - \Phi \left(\frac{\alpha_{s-1} - \beta - \rho_v \sigma_{v2} z_{v1}}{\sqrt{1 + \sigma_u^2 + \sigma_{v2}^2 (1 - \rho_v^2)}} \right) \right] f(z_{v1}) dz_{v1}.
 \end{aligned} \tag{5}$$

3.2 | Proposed paired kappa for intrarater association

We propose here a novel paired kappa statistic κ_{assoc} to assess the strength of intrarater association and the effects of rater training in a large crossed-design study including at least three raters. The proposed intrarater kappa measure κ_{assoc} provides an overall chance-corrected measure of intrarater association over all the ratings from all raters and items included in the study as a population-based measure. It combines the effects of different parameters together into one cohesive summary measure based on the relative effects of the patients' test results and raters and overall shift between the two time points, while correcting for effects of chance intrarater association.

Chance intrarater association is an important consideration when dealing with categorical ratings. It is feasible that an unskilled rater without any expertise could randomly assign the same category to a pair of patient's test results at the two time points by coincidence due to the limited choice of categories. If we do not account for chance association in our measure, we would get an inflated sense of what the strength of association is between a rater's pairs of assessments on the same patient's test result. As we are primarily interested in the amount of association *beyond* that achieved simply by chance, we focus on correcting for chance by minimizing the impact of chance association on κ_{assoc} . We achieve this by selecting

threshold values $\alpha_1, \alpha_2, \dots, \alpha_{C-1}$ that minimize the equation for chance association in Equation (5), and incorporate these values into our κ_{assoc} measure in Equation (6). By incorporating these values $\alpha_{1 \min}, \alpha_{2 \min}, \dots, \alpha_{C-1 \min}$ into κ_{assoc} (which is based on observed association), we obtain the magnitude of association that occurs when the effect of chance association is minimized. Other parameters ($\beta, \sigma_u^2, \Sigma_v$) are not included in this minimization process because these parameters are very informative about the discrepancy and variability between raters' classifications at the two viewings, and consequently association. These minimum values $\alpha_{1 \min}, \alpha_{2 \min}, \dots, \alpha_{C-1 \min}$ are obtained by setting the derivatives of Equation (5) to zero, and can be easily found using an optimization function in R (see the authors' R code in the Supporting Information):

$$\begin{aligned} \kappa_{\text{assoc}} = & \int_{z_u} \int_{z_{v1}} \sum_{r=1}^C \sum_{s=1}^C w_{rs} \left[\Phi(\alpha_{r \min} - \sigma_u z_u - \sigma_{v1} z_{v1}) \right. \\ & \left. - \Phi(\alpha_{r-1 \min} - \sigma_u z_{ui} - \sigma_{v1} z_{v1}) \right] \times \left[\Phi \left(\frac{\alpha_{s \min} - \beta - \sigma_u z_u - \rho_v \sigma_{v2} z_{v1}}{\sqrt{1 + \sigma_v^2(1 - \rho_v^2)}} \right) \right. \\ & \left. - \Phi \left(\frac{\alpha_{s-1 \min} - \beta - \sigma_u z_{ui} - \rho_v \sigma_{v2} z_{v1}}{\sqrt{1 + \sigma_v^2(1 - \rho_v^2)}} \right) \right] f(z_{v1}) f(z_u) dz_{v1} dz_u. \end{aligned} \quad (6)$$

3.3 | Interpretation of the paired kappa

The paired kappa provides a chance-corrected summary measure of association of a typical rater's repeated viewings of a patient's test results, where rater training may be undertaken in between viewings. The kappa measure lies between 0 and 1, and can be interpreted as follows: a low value (near 0) suggests that a rater's classifications of the same patient often disagree substantially between the two viewings. Thus, when the kappa is used to assess the impact of rater training (before vs. after) on the strength of association, a small kappa value indicates the training has modified how a typical rater interprets the patients' paired test items. A high value of kappa (close to 1) suggests strong chance-corrected association and consistency between a typical rater's classifications of the same patient, that is, the same or similar ratings are assigned to a patient's paired test results by a rater. More specifically, a value κ_{assoc} of 0.7–1.0 is considered very strong chance-corrected association between a typical rater's paired scores, a value between 0.5 and 0.7 indicates strong association while a value between 0 and 0.5 to be weak to moderate chance-corrected association.

Asymptotically, κ_{assoc} reaches its maximum value of 1 (perfect chance-corrected association) when all of the following hold: $\beta = 0$, the variance components tend to $+\infty$ and the correlation coefficient ρ_v is 1.0. The measure reaches its minimum value of 0 (zero chance-corrected association) when β heads toward $-\infty$ or $+\infty$ with variance components and correlation coefficients taking any values, though the rate of progression toward 0 will be faster with smaller variance components.

3.4 | A paired kappa for intrarater agreement

In this paper, our focus is on describing an intrarater kappa of association for ordinal paired classifications. However, an intrarater kappa can also be used to calculate the degree of

chance-corrected (exact) *agreement*, κ_{agree} , to assess how often a rater assigns identical ratings to the patient's paired test results. This measure can be calculated using Equation (6) by simply setting $w_{rs} = 1$ when $r = s$, and $w_{rs} = 0$ otherwise. The corresponding standard error is calculated in a similar manner for the intrarater kappa of association described in Section 4.2.

4 | ESTIMATION OF THE ORDINAL GLMM AND PROPOSED PAIRED KAPPA MEASURE

4.1 | Parameter estimation of the ordinal GLMM

The first step requires estimating parameters to get $(\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_{C-1}, \hat{\beta}, \hat{\sigma}_u^2, \hat{\Sigma}_v)$ for the ordinal GLMM model in Equation (1). This is achieved by adopting the multivariate Laplacian method (Shun & McCullagh, 1995), an approximate maximum likelihood approach implemented in the `clmm` function in the `ordinal` package in R (2015). This technique provides an efficient strategy to obtain reasonably unbiased parameter estimates for the GLMM model. Exact maximum likelihood methods cannot be used here due to a lack of a closed form and high dimensionality of the corresponding marginal likelihood function with crossed random effects. Approximate standard errors of the set of parameter estimates can be obtained from the Hessian matrix generated during the model-fitting process. Limited options are currently available in statistical software packages to fit an ordinal GLMM with crossed random effects. Another currently available software package in R for estimating the ordinal GLMM is the `MCMCglmm` package.

4.2 | Estimation of the paired kappa of association

The estimated paired kappa for a specific study is based on the estimated GLMM parameters $\hat{\theta} = (\hat{\beta}, \hat{\sigma}_u^2, \hat{\Sigma}_v)$ as

$$\begin{aligned} \hat{\kappa}_{\text{assoc}} = & \int_{z_u} \int_{z_{v1}} \sum_{r=1}^C \sum_{s=1}^C w_{rs} \left\{ \left[\Phi(\alpha_{r \min} - \hat{\sigma}_u z_u - \hat{\sigma}_{v1} z_{v1}) \right. \right. \\ & - \Phi(\alpha_{r-1 \min} - \hat{\sigma}_u z_u - \hat{\sigma}_{v1} z_{v1}) \left. \right] \times \left[\Phi \left(\frac{\alpha_{s \min} - \hat{\beta} - \hat{\sigma}_u z_u - \hat{\rho}_v \hat{\sigma}_{v2} z_{v1}}{\sqrt{1 + \hat{\sigma}_{v2}^2 (1 - \hat{\rho}_v^2)}} \right) \right. \\ & \left. \left. - \Phi \left(\frac{\alpha_{s-1 \min} - \hat{\beta} - \hat{\sigma}_u z_u - \hat{\rho}_v \hat{\sigma}_{v2} z_{v1}}{\sqrt{1 + \hat{\sigma}_{v2}^2 (1 - \hat{\rho}_v^2)}} \right) \right] \right\} f(z_{v1}) f(z_u) dz_{v1} dz_u. \end{aligned} \quad (7)$$

The approximate standard error of $\hat{\kappa}_{\text{assoc}}$ can be estimated using the multivariate delta method. As a function of vector $\hat{\theta}$, $\text{var}(\hat{\kappa}_{\text{assoc}}) = (\partial \hat{\kappa}_{\text{assoc}} / \partial \hat{\theta}) \Sigma_{\hat{\theta}} (\partial \hat{\kappa}_{\text{assoc}} / \partial \hat{\theta})'$. A sample dataset and R code is provided by the authors in Supporting Information to calculate GLMM parameter estimates and the intrarater kappas for association and agreement.

4.3 | Cohen's averaged intrarater kappa of association

Cohen's (1968) weighted kappa is a familiar measure that can be used to assess intrarater association between a single rater's paired classifications of a sample of test results. For a single rater, this measure takes the form $\kappa_{Cohen} = (p_{0w} - p_{cw}) / (1 - p_{cw})$, where p_{0w} is the weighted proportion of observed agreement and p_{cw} is the weighted proportion of chance agreement, with weights defined similarly to Section 3. Although this measure cannot incorporate more than one rater's paired classifications, we will calculate Cohen's weighted kappa for each individual rater and take the average value over the J raters, which we will call the averaged intrarater kappa of association, $\kappa_{Cohen, assoc}$ for comparison with our proposed kappa measure.

5 | SIMULATION STUDIES

In this section we examine the performance of the proposed paired kappa measure for association via an extensive set of simulation studies under varying conditions. In particular, we investigate the impact of rater and patient sample sizes, misspecification of the random effects distribution and sizes of variance components on the estimation of the paired kappa. We include estimates of the averaged intrarater kappa of association, $\kappa_{Cohen, assoc}$ for comparison with our proposed kappa measure.

5.1 | Setup of simulation studies

One thousand simulated datasets are randomly generated for each set of simulations. We selected small and larger sample sizes of raters and patients ($I = 120, J = 5$), ($I = 200, J = 50$), ($I = 100, J = 50$) and ($I = 250, J = 100$) to demonstrate the performance of the proposed approach. We assumed the use of quadratic weights throughout. For each simulated dataset, paired ratings $Y_{ijk} = c$, ($k = 1, 2$ viewings) from an ordinal scale with $C = 4$ categories are randomly generated using the following technique. The fixed coefficient is set as $\beta = 0.10, 0.15$, or 0.20 and four sets of variance components are examined, $(\sigma_u^2 = 3, \sigma_{v1}^2 = 0.75, \sigma_{v2}^2 = 0.2, \rho_v = 0.4)$, $(\sigma_u^2 = 3, \sigma_{v1}^2 = 0.2, \sigma_{v2}^2 = 0.2, \rho_v = 0.95)$, $(\sigma_u^2 = 1, \sigma_{v1}^2 = 1, \sigma_{v2}^2 = 0.5, \rho_v = 0.20)$, and $(\sigma_u^2 = 10, \sigma_{v1}^2 = 1, \sigma_{v2}^2 = 0.5, \rho_v = 0.90)$. Thresholds are set at $(a_1, a_2, a_3) = (0, 2, 4)$. For each set of simulations, we assume the I patient random effects and J pairs of rater random effects are randomly generated from normal distributions $u_i \sim N(0, \sigma_u^2)$ and $(v_{1j}, v_{2j}) \sim BVN(0, \Sigma_v)$ from Equation (2), respectively, the latter using the `mvrnorm` function in R. A beta distribution, Beta ($a = 0.6, b = 1.4$), is used to generate skewed nonnormally distributed patient random effects $(u_1, \dots, u_I)$ for a second set of simulations to investigate the impact of nonnormal random effects on parameter estimation of the GLMM, which assumes normally distributed random effects. The nonnormal random effects are then scaled to ensure a zero mean and similar variance to the corresponding normal distribution. The `rmultinom` function in R is then used to randomly select samples of ordinal ratings based on the following GLMM:

$$\begin{aligned} \text{First viewing}(k = 1): P(Y_{ij1} \leq c \mid u_i, v_{1j}) &= \Phi(\alpha_c - (u_i + v_{1j})) \\ \text{Second viewing}(k = 2): P(Y_{ij2} \leq c \mid u_i, v_{2j}) &= \Phi(\alpha_c - (\beta + u_i + v_{2j})). \end{aligned}$$

Each simulated dataset is fitted using the `clmm` function in the `ordinal` package in R to estimate the ordinal GLMM parameters $\hat{\theta} = (\hat{\beta}, \hat{\sigma}_u^2, \hat{\Sigma}_v)$ and their standard errors. An estimate of the proposed kappa measure of association in Equation (7) and corresponding standard error are calculated. For each set of simulations, the estimated standard error (est *SE*) of $\hat{\kappa}_{\text{assoc}}$ is calculated as the mean of the 1,000 estimated standard errors, $se(\hat{\kappa}_{\text{assoc}})$ from the 1,000 simulated datasets. The observed standard error (obs. *SE*) for each set of simulations is calculated as the square root of the variance of the 1,000 estimates $\hat{\kappa}_{\text{assoc}}$, from the 1,000 simulated datasets.

5.2 | Simulation study results

Table 1 presents summary tables of the estimation of the proposed kappa measure for each simulation scenario. Full simulation results for all parameters are provided in Tables S2–S5. We observe that the thresholds are estimated throughout all simulation studies with little or no bias. The fixed parameter β is also estimated with no or little bias in each simulation study, with moderate bias in the simulation set with $(I = 120, J = 5)$ and nonnormal patient random effects. The variance of the patient random effects σ_u^2 is estimated with minimal bias throughout the normally distributed random effects simulation studies at each sample size. In the sets of simulations where the patient random effects were generated from a skewed beta distribution, moderate bias is observed in the estimation of patient variance σ_u^2 in the smallest sample size $(I = 120, J = 5)$, which dissipates at the larger sample sizes. For the variance components related to the rater random effects, a moderate level of bias is noted in the estimation of the rater variances at each of the two time points σ_{v1}^2 and σ_{v2}^2 in the smallest sample size $(I = 120, J = 5)$, which again disappears at the larger sample sizes. The rater correlation coefficient ρ_v is well estimated throughout the simulation studies. The presence of some bias in the GLMM parameters at the smaller sample size did not impact the estimation of the proposed kappa measure of association κ_{assoc} , which was estimated with little or no bias in all the sets of simulation studies, and at both small and large sample sizes, demonstrating that its estimation is robust to smaller sample sizes and violation of the normally distributed random effect assumption. The averaged intrarater Cohen's kappa measures for agreement and association, $\kappa_{\text{Cohen, agree}}$ and $\kappa_{\text{Cohen, assoc}}$, respectively, consistently take smaller values than the new proposed measures κ_{agree} and κ_{assoc} in each simulation scenario, though tend to be closer in value when the correlation ρ_v is strong.

Coverage probabilities for κ_{assoc} defined as the percent of 95% confidence intervals from the 1,000 simulated datasets that include the true value of κ_{assoc} are summarized in Table S1. These results indicate very good to excellent coverage for the simulation studies with normally distributed random effects, while the coverage probabilities varied more for the nonnormally distributed random effects simulation studies, ranging from 83.4% to 100.0%.

5.3 | Characteristics of the paired kappa

Figure 1 displays plots to investigate the behavior of the proposed measure. The number of categories is set at $C = 4$. In particular, Figure 1a presents the effects of increasing patient variance σ_u^2 over a broad range of the fixed coefficient β . Other parameters are kept constant

at $\sigma_{v1}^2 = 5$, $\sigma_{v2}^2 = 1$ and $\rho_v = 0.9$. We observe that larger values of σ_u^2 lead to stronger intrarater association across the range of values of β due to patient test results that are very distinguishable from each other. This suggests good visibility of the presence of disease (or not), leading to more consistent ratings of patient test results at the two viewings. Intrarater association is always highest when β is zero, occurring when the distribution of rater responses at both time points across the C categories is the same. As β moves away from 0, the distributions of rater responses at the two time points become more discrepant, indicative of weaker intrarater association and less consistency between measurements made at two time points. We observe that Cohen's averaged intrarater kappa consistently takes lower values for both values of σ_u^2 . Figure 1b demonstrates how $\hat{\kappa}_{\text{assoc}}$ is robust to the underlying prevalence of disease for any specific value of β . Threshold values were selected to create sets of ratings with different distributions of ratings over the C categories. We focused on three scenarios of disease prevalence: a rare disease (80% in category 1 (no disease), 15% in category 2, 4% in category 3, and 1% in category 4 (high probability of disease)); equal prevalence (25% of ratings in each of the four categories); common disease (1% in category 1, 4% in category 2, 15% in category 3, and 80% in category 4). We see that estimation of $\hat{\kappa}_{\text{assoc}}$ is not influenced by the underlying disease prevalence, and that $\hat{\kappa}_{\text{assoc}}$ is strongest at $\beta = 0$. In comparison, Cohen's averaged intrarater kappa is observed to be impacted by disease prevalence.

6 | APPLICATION TO BREAST MRI IMAGING STUDY

Women at high risk for breast cancer may undergo supplemental screening via magnetic resonance imaging (MRI). Background parenchymal enhancement (BPE) presents background noise on an MRI in the search for breast cancers. Women with greater than minimal BPE on an MRI may face increased risk of developing breast cancer (Grimm et al., 2019). BPE is visually categorized on an MRI by a radiologist using an ordinal scale based on the volume and intensity of enhancement. Due to the visual nature of the interpretation, classification of BPE is vulnerable to variability of rater scoring (Dontchos et al., 2015; Grimm et al., 2019; King et al., 2011). Melsaether et al. (2014) recently conducted a study to investigate whether rater training (a 2-hr interactive presentation) improved inter- and intrarater agreement of radiologists' interpretation of BPE. At the first time point, four radiologists each independently rated a sample of 119 breast MRI exams in accordance with the 4-point BI-RADS scale (1 = minimal BPE, <25%; 2 = mild BPE, 25–50%; 3 = moderate BPE, 51–75%; 4 = marked BPE, >75%). The radiologists underwent training 2 weeks later, and 1 week after undergoing the training (time point 2), the radiologists reread the MRIs (in a randomized order). To assess intrarater agreement, the radiologists then viewed the MRIs at a third time point (at least 3 weeks after training to minimize recall bias).

6.1 | Results for breast MRI imaging study—Effectiveness of rater training

Table 2 presents results for BPE assessment by radiologists before versus after training (time point one vs. time point two). In this population-based setting, where we are comparing the interpretation of patients' BPE test results viewed at two time points by each rater, a kappa value of $\hat{\kappa}_{\text{assoc}} = 0.755$ (95% CI [0.715, 0.795]) implies that there is reasonably strong

association between a rater's paired scores of a patient's BPE from the two viewings (before vs. after training) after correcting for chance association. The paired kappa of agreement was lower at $\hat{\kappa}_{\text{agree}} = 0.487$ (95% CI [0.434, 0.540]) indicating only moderate agreement in a radiologist's ratings before versus after training, and a different score was often assigned to the same patient after training. The fixed coefficient $\hat{\beta} = 0.122$ ($SE = 0.414$) is close to 0. The z -statistic for the corresponding Wald hypothesis test $H_0: \beta = 0$, $z = (\hat{\beta} - 0)/SE(\hat{\beta}) = 0.294$ with p -value of 0.769 tells us there is not a statistically significant shift in the distribution of responses across the ordinal scale after training by the group of radiologists, although our results likely do not have high power due to the limited number of raters included in this study and small overall sample size. Melsaether et al. (2014) present intrarater Cohen kappas of agreement for each individual rater, which ranged from 0.45 to 0.84. Our paired kappas provide overall kappa values of agreement and association based on all the raters' paired classifications.

Moderate variability exists between the four radiologists' ratings before training is undertaken where $\hat{\sigma}_{v1}^2 = 0.768$ ($SE = 0.545$), which decreases to $\hat{\sigma}_{v2}^2 = 0.204$ ($SE = 0.144$) after training has occurred. These variance estimates are also reflected in Figure 1, which displays boxplots of the four radiologists' conditional modes (estimates of the random effects) from the *clmm* procedure in the *ordinal* package in R. This suggests that radiologists provided less variable ratings as a group after undergoing training, with less extreme rating behavior exhibited by any radiologist.

6.2 | Results for breast MRI imaging study—Intrarater agreement and association

Comparing radiologists' paired ratings between time points two and three provides a valuable overall assessment of intrarater association and agreement for BPE when no training was conducted in the interim; the radiologists simply reviewed the slides a third time in a randomized order. These results are presented in the right-hand column in Table 2. We note that radiologists' behavior exhibit similar variability at both time points two and three with $\hat{\sigma}_{v2}^2 = 0.189$ and $\hat{\sigma}_{v3}^2 = 0.197$, supported by Figure 2, which displays very similar boxplots of conditional modes across time points two and three. The estimated correlation between paired ratings at time points two and three is also high at $\hat{\rho}_v = 0.96$. Both the estimated paired kappa of association $\hat{\kappa}_{\text{assoc}} = 0.775$ (95% CI [0.745, 0.806]) and paired kappa of agreement $\hat{\kappa}_{\text{agree}} = 0.515$ (95% CI [0.434, 0.540]) suggest improved consistency between a rater's paired scores compared with the first two time points. Melsaether et al.'s (2014) intrarater kappas of agreement (Cohen's kappas) for each individual rater ranged from 0.56 to 0.98.

The paired kappa for intrarater agreement $\hat{\kappa}_{\text{agree}} = 0.515$ ($SE = 0.022$) indicates moderately strong chance-corrected intrarater agreement for the typical rater. The Wald hypothesis test for $H_0: \beta = 0$ yields a p -value of .142 indicating no statistically significant shift in the overall distribution of scores across the time points two and three.

6.3 | Goodness of fit for the breast MRI imaging study

An approach described by Todem, Kim, and Lesaffre (2007) and Hedeker and Gibbons (1994) to assess goodness of fit focuses on comparing the overall observed frequencies of MRIs by the BI-RADS category (observed counts) with frequencies estimated based on the ordinal GLMM model (estimated counts). Marginal probabilities at each of the three time points of being classified into each of the four ordinal categories are calculated using

$$P(Y_{ijk} = c) = \Phi\left(\frac{\alpha_c - \beta x_k}{\sqrt{1 + \sigma_u^2 + \sigma_{vk}^2}}\right) - \Phi\left(\frac{\alpha_{c-1} - \beta x_k}{\sqrt{1 + \sigma_u^2 + \sigma_{vk}^2}}\right) \quad c = 1, \dots, 4. \quad (7)$$

Table 3a presents the observed and estimated counts (based on probabilities estimated using the ordinal GLMM approach) for Melsaether et al.'s (2014) breast-imaging MRI study with $J = 4$ radiologists and $I = 119$ patient MRIs over three time points. The estimated counts align very reasonably with the observed counts, indicating that the GLMM model is a good model fit here.

7 | APPLICATION TO A LARGE-SCALE BREAST DENSITY STUDY

To demonstrate the applicability of our proposed methods to a much bigger study, we consider a second breast-imaging study. Breast density is considered an important risk factor and masking factor when interpreting mammograms for breast cancer. A recent study conducted by the Breast Cancer Surveillance Consortium collected breast density measurements on a stratified random sample of 332 pairs of screening mammograms (Gard et al., 2015; Taplin, Rutter, & Lehman, 2006). Nineteen community radiologists ($J = 19$) each independently interpreted the breast MRIs of $I = 332$ patients at one time point and again after approximately six months. The third edition of the BI-RADS four-category BI-RADS breast density ordinal scale (1 = entirely fat; 2 = scattered fibroglandular density; 3 = heterogeneously dense; and 4 = extremely dense) was used for interpretation by radiologists (American College of Radiology, 1998). This dataset is as part of a larger study investigating the impact of computer-assisted detection in the detection of breast cancer. Of interest here is to investigate the strength of intrarater agreement and association between the radiologists' paired mammogram viewings at the two time points.

7.1 | Results for breast density imaging study—Intrarater agreement and association

Table 4 presents the results for the breast density study. We see that the paired kappa of agreement $\hat{\kappa}_{\text{agree}} = 0.610$ (95% CI [0.577, 0.643]) and association $\hat{\kappa}_{\text{assoc}} = 0.836$ (95% CI [0.816, 0.855]) yield strong intrarater agreement and very strong intrarater association, respectively, between the paired density measurements of radiologists 6 months apart. The estimated correlation coefficient between radiologists' paired ratings is very strong at $\hat{\rho}_v = 0.99$, although it is important to be reminded that a high correlation does not necessarily imply strong agreement or association.

The patient variance $\hat{\sigma}_u^2 = 6.202$ ($SE = 0.512$) suggests good distinguishability of breast density between mammograms, while $\hat{\sigma}_{v1}^2 = 0.513$ ($SE = 0.164$) and $\hat{\sigma}_{v2}^2 = 0.511$ ($SE = 0.163$)

are both very similar at the two time points. Closer inspection of individual radiologists' paired ratings of a patient's test results demonstrate that the majority of radiologists did not deviate much from their first reading of a patient's test result in their second reading six months later, providing results that were identical or just one category apart, leading to the very strong estimation of intrarater association. The Wald hypothesis test for $H_0: \beta = 0$ yields a p -value of .151, which showed no overall shift in the radiologists' overall distribution of ratings across the six months. Our results provide a comprehensive overall summary of agreement and association that extend results presented in Melsaether et al. (2014) at the individual radiologist level.

7.2 | Goodness of fit for the breast density imaging study

Following a similar approach outlined in Section 6.3 and Equation (7), the table of observed and estimated counts presented in Table 3b for the large-scale breast density study (Gard et al., 2015) with $J = 19$ experts and $I = 332$ patient test results at two time points six months apart shows that the ordinal GLMM model fits this dataset reasonably well across the four breast density categories at both time points.

8 | DISCUSSION

Rater variability continues to be an important limitation when interpreting images from screening and diagnostic tests in clinical practice. In this paper, to gain better insight into the sources of the variability, we propose a new measure that provides an overall evaluation of the association among multiple raters' paired scores of patients' ordinal test results from two time points. Any number of raters (more than two) can be accommodated, with no added complexity for a very large number of raters. To the best of the authors' knowledge, no other statistical approaches exist for this purpose, especially in larger studies when the number of raters is more than two or three. Many research studies report intra-agreement statistics for each individual rater, using Cohen's kappa, however this approach becomes complex when a large number of raters contribute ratings. The averaged intrarater Cohen's kappa measures of agreement and association (averaged over the raters included in the study) tend to take smaller values than the proposed kappa measures and are vulnerable to the well-known flaws of Cohen's kappa as outlined in the Introduction. Our proposed summary measure can flexibly be used for either ordinal and binary ratings, can accommodate missing ratings, and is robust to the underlying prevalence of disease. The role of rater and patient-level covariates such as the experience level of a rater, or the age and family history of a patient may also play an important role in understanding the levels of intrarater agreement and impact of rater training. A future research goal is to extend the proposed intrarater kappas to assess the effects of rater- and patient-level covariates.

In the screening and diagnosis of many diseases, no gold standard measurement may be available. In these settings, the study of intrarater agreement (reproducibility) and association plays an important role in evaluating the consistency of raters' classifications and ultimately, the effectiveness of the screening test. When a gold standard exists to determine a patient's true disease status, alternative summary measures such as the area under the receiver operator curve and receiver operator curve methodology are preferred approaches

(Gatsonis, 2009; Pepe, 2003). In this case, our proposed kappa provides additional valuable insight into the overall consistency of a screening procedure and can be used in conjunction with these other statistical procedures to provide a comprehensive evaluation of the screening test.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors are grateful for the support provided by grants R01-CA172463 and R01-CA226805 from the United States National Institutes of Health. We thank Dr. Amy Melsaether for providing the breast-imaging MRI dataset. Data collection and sharing for the large-scale breast density (CAD vs. no CAD) study were supported by the National Cancer Institute-funded Breast Cancer Surveillance Consortium (BCSC; HHSN261201100031C, P01CA154292) and U01CA6373. We also thank the participating women, mammography facilities, radiologists, and BCSC investigators for the data they have provided for this study. You can learn more about the BCSC at <http://www.bsc-research.org/>.

Funding information

United States National Institutes of Health, Grant/Award Numbers: R01-CA172463, R01-CA226805; National Cancer Institute-funded Breast Cancer Surveillance Consortium (BCSC, Grant/Award Numbers: HHSN261201100031C, P01CA154292; Division of Cancer Prevention, National Cancer Institute, Grant/Award Numbers: R01-CA172463, R01-CA226805

REFERENCES

- Agresti A (2010). Analysis of ordinal categorical data (2nd ed.). Hoboken, NJ: Wiley.
- American College of Radiology (1998). Breast imaging reporting and data system (3rd ed.). Reston, VA: American College of Radiology.
- Antonio ALM, & Crespi CM (2010). Predictors of interobserver agreement in breast imaging using the Breast Imaging Reporting and Data System (BI-RADS). *Breast Cancer Research and Treatment*, 120, 539–546. [PubMed: 20300960]
- Baker SG, Freedman LS, & Parmar MKB (1991). Using replicate observations in observer agreement studies with binary assessments. *Biometrics*, 47, 1327–1338. [PubMed: 1786322]
- Banerjee M (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, 27, 3–23.
- Barnhart HX, & Williamson JM (1997). Assessing interrater agreement from dependent data. *Biometrics*, 53, 707–714. [PubMed: 9192459]
- Bloch DA, & Kraemer HC (1989). 2×2 kappa coefficients: Agreement or association. *Biometrics*, 45, 269–287. [PubMed: 2655731]
- Brennan RL, & Prediger DJ (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41, 687–699.
- Breslow NE, & Clayton DG (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9–25.
- Cao H, Sen PK, Peery AF, & Dellon ES (2016). Assessing agreement with multiple raters on correlated kappa statistics. *Biometrical Journal*, 58, 935–943. [PubMed: 26890370]
- Chaves MLF, Camozzato AL, Godinho C, Kochhann R, Schuh A, & de Almeida VL (2007). Validity of the Clinical Dementia Rating Scale for the detection and staging of dementia in Brazilian patients. *Alzheimer Disease & Associated Disorders*, 21, 210–217. [PubMed: 17804953]
- Choa RM, & Giele HP (2015). Inter- and intrarater reliability of osteoarthritis classification at the trapeziometacarpal joint. *Journal of Hand Surgery*, 40, 23–26. [PubMed: 25443166]

- Ciatto S, Houssami N, Apruzzese A, Bassetti E, Brancato B, Carozzi F, ... Scorsolini A (2006). Reader variability in reporting breast imaging according to BI-RADS assessment categories (the Florence experience). *Breast*, 15, 44–51. [PubMed: 16076556]
- Cohen J (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Cohen J (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213–200. [PubMed: 19673146]
- Donner A, Shoukri MM, Klar N, & Bartfay E (2000). Testing the equality of two dependent kappa statistics. *Statistics in Medicine*, 19, 373–387. [PubMed: 10649303]
- Dontchos BN, Rahbar H, Partridge SC, Korde LA, Lam DL, Scheel JR, ... Lehman CD (2015). Are qualitative assessments of background parenchymal enhancement, amount of fibroglandular tissue on MR images, and mammographic density associated with breast cancer risk? *Radiology*, 276, 371–380. [PubMed: 25965809]
- Elmore JG, Longton GL, Carney PA, Geller BM, Onega T, Tosteson A, ... Weaver DL (2015). Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Journal of the American Medical Association*, 313, 1122–1132. [PubMed: 25781441]
- Fleiss JL (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378–382.
- Gard CC, Aiello Bowles EJ, Miglioretti DL, Taplin SH, & Rutter CM (2015). Misclassification of Breast Imaging Reporting and Data System (BIRADS) mammographic density and implications for breast density reporting legislation. *Breast Journal*, 21, 481–489. [PubMed: 26133090]
- Gatsonis CA (2009). Receiver operating characteristic analysis for the evaluation of diagnosis and prediction. *Radiology*, 253, 593–596. [PubMed: 19952022]
- Gonin R, Lipsitz SR, Fitzmaurice GM, & Molenberghs G (2000). Regression modelling of weighted κ by using generalized estimating equations. *Journal of the Royal Statistical Society Series C—Applications*, 49, 1–18.
- Graham P, & Jackson R (1993). The analysis of ordinal agreement data: Beyond weighted kappa. *Journal of Clinical Epidemiology*, 46, 1055–1063. [PubMed: 8263578]
- Grassano L, Pagana G, Daperno M, Bibbona E, & Gasparini M (2018). Asymptotic distributions of kappa statistics and their differences with many raters, many rating categories and two conditions. *Biometrical Journal*, 60, 146–154. [PubMed: 29110316]
- Grimm LJ, Saha A, Ghate SV, Kim C, Scott Soo M, Yoon SC, & Mazurowski MA (2019). Relationship between background parenchymal enhancement on high-risk screening MRI and future breast cancer risk. *Academic Radiology*, 26, 69–75. [PubMed: 29602724]
- Gwet KL (2016). Testing the difference of correlated agreement coefficients for statistical significance. *Educational and Psychological Measurement*, 76, 609–637. [PubMed: 29795880]
- Hedeker D, & Gibbons RD (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, 50, 933–44. [PubMed: 7787006]
- Hsiao CK, Chen P-C, & Kao W-H (2011). Bayesian random effects for interrater and test-retest reliability with nested clinical observations. *Journal of Clinical Epidemiology*, 64, 808–814. [PubMed: 21292442]
- Kerlikowske K, Grady D, Barclay J, Frankel SD, Ominsky SH, Sickles EA, & Ernster V (1998). Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. *Journal of the National Cancer Institute*, 90, 1801–1809. [PubMed: 9839520]
- King V, Brooks JD, Bernstein JL, Reiner AS, Pike MC, & Morris EA (2011). Background parenchymal enhancement at breast MR imaging and breast cancer risk. *Radiology*, 260, 50–60. [PubMed: 21493794]
- Liu I, & Agresti A (2005). The analysis of ordinal categorical data: An overview and a survey of recent developments. *Test*, 14, 1–73.
- Maclure M, & Willett WC (1987). Misinterpretation and misuse of the Kappa statistic. *American Journal of Epidemiology*, 126, 161–169. [PubMed: 3300279]

- Melsaether A, McDermott M, Gupta D, Pysarenko K, Shaylor SD, & Moy L (2014). Inter- and intra-reader agreement for categorization of background parenchymal enhancement at baseline and after training. *American Journal of Roentgenology*, 203, 209–213. [PubMed: 24951217]
- Nelson KP, & Edwards D (2015). Measures of agreement between many raters for ordinal classifications. *Statistics in Medicine*, 34, 3116–3132. [PubMed: 26095449]
- Nelson KP, & Edwards D (2018). A measure of association for ordered categorical data in population-based studies. *Statistical Methods in Medical Research*, 27, 812–831. [PubMed: 27184590]
- Pepe MS (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction* Oxford Statistical Science Series. Oxford: Oxford University Press.
- R Core Team (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing Retrieved from <http://www.R-project.org/>
- Sadler ME, Yamamoto RT, Khurana L, & Dallabrida SM (2017). The impact of rater training on clinical outcomes assessment data: A literature review. *International Journal of Clinical Trials*, 4, 101–110.
- Scott WA (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19, 321–325.
- Shun Z, & McCullagh P (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society Series B Methods*, 57, 749–760.
- Taplin SH, Rutter CM, & Lehman CD (2006). Testing the effect of computer-assisted detection on interpretive performance in screening mammography. *American Journal of Roentgenology*, 187, 1475–1482. [PubMed: 17114540]
- Thompson WD, & Walter SD (1988). A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology*, 41, 949–958. [PubMed: 3057117]
- Todem D, Kim K, & Lesaffre E (2007). Latent-variable models for longitudinal data with bivariate ordinal outcomes. *Statistics in Medicine*, 26, 1034–1054. [PubMed: 16832841]
- Tsai M-Y. (2012). Assessing inter- and intra-rater agreement for dependent binary data: A Bayesian hierarchical correlation approach. *Journal of Applied Statistics*, 39, 173–187.
- Vanbelle S, Mutsvari T, Declerck D, & Lesaffre E (2012). Hierarchical modeling of agreement. *Statistics in Medicine*, 31, 3667–3680. [PubMed: 22736494]
- Williamson JM, Lipsitz SR, & Manatunga AK (2000). Modeling kappa for measuring dependent categorical agreement data. *Biostatistics*, 1, 191–202. [PubMed: 12933519]
- Williamson JM, & Manatunga AK (1997). Assessing interrater agreement from dependent data. *Biometrics*, 53, 704–714.

FIGURE 1.

Properties of the paired kappa κ_{assoc} when varying parameters. Effects on the value of κ_{assoc} for (a) increasing patient variance σ_u^2 over the range of β and other parameters remain fixed; (b) varying prevalence of disease over the C categories. The averaged Cohen's intrarater kappa, $\kappa_{\text{Cohen, assoc}}$, is included for comparison

FIGURE 2.

Boxplots of conditional modes are presented for each time point for (a) Melsaether et al.'s (2014) study with three time points; (b) breast density study with two time points six months apart

TABLE 1

True values and mean estimates (mean standard error) of the proposed paired kappa of association κ_{assoc} for each of eight simulation studies

(a) Normally distributed random effects

Variances ($\sigma_u^2, \sigma_{v1}^2, \sigma_{v2}^2$)	Correlation ρ_v	True κ_{assoc}	<i>I</i> = 120 <i>J</i> = 5	<i>I</i> = 200 <i>J</i> = 50
			Mean $\hat{\kappa}_{\text{assoc}}$ (SE)	Mean $\hat{\kappa}_{\text{assoc}}$ (SE)
(3, 0.75, 0.20)	0.40	0.749	0.744 (0.022)	0.747 (0.011)
(3, 0.20, 0.20)	0.95	0.773	0.768 (0.015)	0.772 (0.009)
			<i>I</i> = 100 <i>J</i> = 50	<i>I</i> = 250 <i>J</i> = 100
(1,1, 0.50)	0.20	0.635	0.636 (0.016)	0.634 (0.011)
(10,1, 0.50)	0.40	0.860	0.856 (0.010)	0.857 (0.006)

(b) Nonnormally distributed random effects

Variances ($\sigma_u^2, \sigma_{v1}^2, \sigma_{v2}^2$)	Correlation ρ_v	True κ_{assoc}	<i>I</i> = 120 <i>J</i> = 5	<i>I</i> = 200 <i>J</i> = 50
			Mean $\hat{\kappa}_{\text{assoc}}$ (SE)	Mean $\hat{\kappa}_{\text{assoc}}$ (SE)
(3, 0.75, 0.20)	0.40		0.763 (0.022)	0.750 (0.010)
(3, 0.20, 0.20)	0.95		0.788 (0.014)	0.775 (0.009)
			<i>I</i> = 100 <i>J</i> = 50	<i>I</i> = 250 <i>J</i> = 100
(1,1, 0.50)	0.20		0.635 (0.011)	0.635 (0.006)
(10,1, 0.50)	0.40		0.870 (0.009)	0.867 (0.006)

Note. Each set of simulations is based on 1,000 simulated datasets with ordinal ($C = 4$) outcomes.

TABLE 2

Assessment of BPE in Melsaether et al. (2014) study

		(a) Before vs. after training (Time points 1 vs. 2)	(b) Intrarater agreement (Time points 2 vs. 3)
Parameter	Symbol	Estimate (SE)	Estimate (SE)
Ordinal GLMM			
Threshold 1	α_1	0.006 (0.475)	−0.075 (0.282)
Threshold 2	α_2	1.922 (0.481)	1.754 (0.290)
Threshold 3	α_3	3.848 (0.504)	3.633 (0.321)
Fixed coefficient	β	0.122 (0.414)	0.157 (0.107)
Patient MRI variance	σ_u^2	3.212 (0.555)	3.069 (0.528)
Rater variances			
First time point	σ_{v1}^2	0.768 (0.545)	—
Second time point	σ_{v2}^2	0.204 (0.144)	0.189 (0.133)
Third time point	σ_{v3}^2	—	0.197 (0.139)
Correlation coefficient	ρ_v	0.401 (0.449)	0.960 (0.088)
Measures of agreement			
Paired kappa	κ_{agree}	0.487 (0.029)	0.515 (0.022)
Cohen's averaged intrarater kappa	$\kappa_{\text{Cohen,agree}}$	0.502 (0.062)	0.716 (0.064)
Measures of association			
Paired kappa	κ_{assoc}	0.755 (0.025)	0.775 (0.015)
Cohen's averaged intrarater weighted kappa	$\kappa_{\text{Cohen,assoc}}$	0.745 (0.086)	0.854 (0.091)

Note. Four radiologists ($J = 4$) interpreted breast MRIs of $I = 119$ patients using the BI-RADS four-point BPE ordinal scale at three time points: 2 weeks before training (time point 1), within 1 week after training (time point 2), and 3+ weeks after training (time point 3).

TABLE 3

Goodness-of-fit assessment for (a) Melsaether et al.'s (2014) breast-imaging MRI study; (b) a large-scale breast density study Gard et al. (2015)

(a) Marginal observed (estimated) counts for breast-imaging MRI study			
BPE category	Time point 1	Time point 2	Time point 3
1	243 (239)	236 (231)	220 (217)
2	146 (145)	148 (151)	156 (155)
3	65 (72)	72 (75)	77 (82)
4	22 (20)	20 (19)	23 (22)
(b) Marginal observed (estimated) counts for breast density study			
Breast density category	Time point 1	Time point 2	
1	383 (362)	352 (375)	
2	2,103 (2,013)	2,178 (2,041)	
3	2,517 (2,650)	2,536 (2,631)	
4	745 (722)	682 (700)	

Note. Observed counts (estimated counts) are displayed for each time point.

TABLE 4

Assessment of intrarater agreement and association in a large breast density study (Gard et al., 2015), using a four-category ordinal scale of increasing breast density

Parameter	Symbol	Estimate (SE)
Ordinal GLMM		
Threshold 1	α_1	-4.250 (0.222)
Threshold 2	α_2	-0.609 (0.216)
Threshold 3	α_3	3.187 (0.219)
Fixed coefficient	β	-0.050 (0.035)
Patient variance	σ_u^2	6.202 (0.512)
Rater variances		
First reading	σ_{v1}^2	0.513 (0.164)
Second reading (6 months later)	σ_{v2}^2	0.511 (0.163)
Correlation coefficient	ρ_v	0.992 (0.004)
Measures of agreement		
Paired kappa	κ_{agree}	0.610 (0.010)
Cohen's averaged intrarater kappa	$\kappa_{\text{Cohen,agree}}$	0.692 (0.042)
Measures of association		
Paired kappa	κ_{assoc}	0.836 (0.010)
Cohen's averaged intrarater weighted kappa	$\kappa_{\text{Cohen,assoc}}$	0.816 (0.058)