Published in final edited form as:

Stat Med. 2019 July 30; 38(17): 3272–3287. doi:10.1002/sim.8200.

A Paired Kappa To Compare Binary Ratings Across Two Medical Tests

Kerrie P. Nelson,

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

Don Edwards

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

Abstract

Agreement between experts' ratings is an important prerequisite for an effective screening procedure. In clinical settings large-scale studies are often conducted to compare the agreement of experts' ratings between new and existing medical tests, for example, digital versus film mammography. Challenges arise in these studies where many experts rate the same sample of patients undergoing two medical tests, leading to a complex correlation structure between experts' ratings. Here we propose a novel paired kappa measure to compare the agreement between the ~~ordinal~~ binary ratings of many experts across two medical tests. Existing approaches can accommodate only a small number of experts, and rely heavily on Cohen's kappa and Scott's pi measures of agreement, thus are prone to their drawbacks. The proposed kappa appropriately accounts for correlations between ratings due to patient characteristics, corrects for agreement due to chance and is robust to disease prevalence and other flaws inherent in the use of Cohen's kappa. It can be easily calculated in the software package R. In contrast to existing approaches, the proposed measure can flexibly incorporate large numbers of experts and patients by utilizing the generalized linear mixed models framework. It is intended to be used in population-based studies, increasing efficiency without increasing modeling complexity. Extensive simulation studies demonstrate low bias and excellent coverage probability of the proposed kappa under a broad range of conditions. Methods are applied to a recent nationwide breast cancer screening study comparing film mammography to digital mammography.

Keywords

Agreement; binary classifications; kappa; breast imaging; screening test

1. Introduction

In many medical settings, clinical findings and images are viewed by experts to classify diseases such as breast cancer, dementia and skin conditions^{1–4}. Screening tests such as mammography play an integral role in early detection of disease, leading to a reduction in deaths⁵, and strong consistency between experts' ratings is an essential prerequisite for an

effective screening procedure. However, widespread discrepancies are often observed between experts' visual interpretations of patient test images, leading to lower levels of accuracy and agreement between experts.

With the goal of improving breast cancer detection rates, large-scale clinical studies are conducted to compare the consistency of experts' ratings between newly developed and existing screening breast imaging procedures⁶⁻⁹. Typically, a large random sample of patients undergoes both screening tests (new and existing) within a short time frame, and test results are independently interpreted by each expert such as a radiologist, in a large group of experts often using a defined binary or ordinal grading scale. Since both screening tests are conducted on the same sample of patients, test results for a patient are strongly linked due to the underlying patient characteristics. As each expert interprets both sets of images, an expert's paired classifications of the two screening test results for an individual patient may also be highly correlated. Consequently, any assessment of agreement between an expert's ratings from two medical tests needs to account for these correlations. Further challenges arise due to the large numbers of experts who participate in these studies.

Much recent work on the topic of agreement has focused on the study of correlated kappa statistics¹⁰⁻¹⁷. However, despite the current trend of large-scale agreement studies, no approach exists for comparing agreement between an expert's ratings from two medical tests when a large number of experts contribute binary ratings. Most existing approaches for binary ratings can only incorporate a small fixed number of experts and are based upon Cohen's kappa¹⁸ and Scott's pi¹⁹, which allow for differing and equal marginal distributions of experts' ratings respectively. Oden²⁰ developed a pooled version of Cohen's kappa to account for correlation between paired binary assessments, where both eyes of a sample of patients were graded by the same two experts using a binary grading scale. Klar et al.¹⁵ describe a modeling approach based upon estimating equations to assess the impact of covariates, such as different screening tests on Cohen's kappa and Scott's pi for binary ratings and two experts for independent samples of patients. Their approach can potentially incorporate more than two experts and correlated ratings. Donner et al.¹⁴ assess the equality of intra-class correlation coefficients (ICCs) between two sets of binary ratings on the same sample of patients classified by two experts, while Dixon et al.¹⁰ focus on inferences for the ICC. Schouten et al.²¹ describes a kappa statistic similar to Fleiss' kappa for assessing pairwise agreement between each pair of experts, but when the same group of experts grade all patients and focuses upon ratings from one screening procedure only. Vanbelle and Albert¹⁷ describe a coefficient extending Cohen's kappa to quantify agreement between two independent groups of experts classifying the same set of patients.

McKenzie et al.¹² proposed a resampling method using bootstrapping techniques to quantify the variance of the difference between two correlated Cohen's kappa statistics for two experts, with pairwise comparisons of kappa statistics when patients were evaluated by three experts. Their approach was later extended by Vanbelle et al.¹⁶ who implemented bootstrapping techniques to test two or more correlated Cohen's kappa statistics for equality for binary assessments by two experts. These approaches tend to be computationally intensive, though require minimal distributional assumptions. Barnhart and Williamson¹³ employ the method of weighted least squares to test the equality of two correlated kappas for

two experts' binary or nominal categorical assessments on a single sample of patients using the SAS statistical software package. Other existing modeling approaches utilize generalized estimating equations (GEEs) to describe correlated Cohen's kappa-like coefficients to assess agreement between nominal ratings¹¹ of a fixed, small number of raters. These approaches can be used for comparing ratings between two test procedures, with potential extension to several testing procedures.

While Cohen's kappa and Scott's pi agreement measures have some desirable properties (are simple, easy to interpret and widely used), they are vulnerable to a number of drawbacks including often overcorrecting for chance agreement leading to conservative assessments of agreement, are affected by the underlying disease prevalence^{22–24} and cannot accommodate more than a small number of experts' ratings or account for correlated ratings.

Consequently, clinical studies often report screening test results for every pair of experts (for example, Ooms et al.²⁵, Comperat et al.²⁶), resulting in complexities in interpretation and a loss of efficiency, rapidly becoming infeasible for a study with a large number of experts.

In this paper we propose a novel paired kappa measure to compare the strength of agreement between an expert's binary ratings from two medical tests when several experts (at least three) provide ratings. It is assumed that a single sample of patients undergoes two medical tests and their test results evaluated by the same group of experts. Strengths of the proposed kappa include incorporating dependencies between an expert's correlated ratings for an individual patient, correcting for chance agreement and being robust to the underlying disease prevalence. Our proposed paired kappa utilizes the class of generalized linear mixed models (GLMMs) to flexibly accommodate the ratings of many experts in large-scale agreement studies without additional complexity. The proposed kappa is intended to be used in population-based studies and to measure agreement in the tests over all possible experts and patients in the underlying populations, not just the experts and patients included in the study. It makes inference on the diagnostic processes themselves by treating patient and expert effects as random effects. Our goal is to draw conclusions about the agreement between a typical expert's binary ratings provided by two medical tests. Test results are obtained from a randomly selected patient in a population-based setting. Furthermore, unique characteristics of individual experts can be assessed for each screening test, where, for example, some experts may be more conservative in assigning positive test scores in one medical test relative to the other test.

The paper proceeds as follows – in Section Two we describe the population-based GLMM framework for accommodating ratings from two medical tests when experts use an binary classification scale. The proposed paired kappa measure to assess agreement between a typical expert's ratings for two medical tests is derived in Section Three. Section Four presents simulation studies conducted to assess performance and properties of the proposed approach while in Section Five, results are presented for a recent nationwide large-scale breast screening study comparing film mammography to digital mammography. Assessment of goodness-of-fit is described in Section Six and we conclude with a brief discussion in Section Seven.

2. A Generalized Linear Mixed Model for Paired Binary Ratings

To compare the agreement between a typical expert's binary ratings between two screening medical tests, a common study design is assumed where a random sample of J experts ($J > 2$) each independently classify the test results of the same randomly selected sample of I patients ($I > 2$) who undergo two screening tests ($K = 2$) using the same ordinal or binary grading scale. A GLMM provides a flexible framework for this setting, with the important advantage of being able to incorporate correlations between the paired measurements provided by each expert on each patient's screening medical test results. It is assumed that the experts view the sets of test results from the two screening tests in a blinded and randomized order with a reasonable time interval between viewings to avoid recall bias. We define $Y_{ijk} = c$ as the binary rating ($C = 2$ categories; for example, $c = 1$ not diseased or negative, $c = 2$ diseased or positive) assigned by expert j ($j = 1, \dots, J$) to the i th patient's test result ($i = 1, \dots, I$) obtained from the k th test ($k = 1, 2$). A GLMM with a probit link function models the probability that a patient's test result is classified as diseased as:

$$P(Y_{ijk} = 2 | u_{ki}, v_{kj}) = 1 - \Phi(\alpha - (\beta x_k + u_{ki} + v_{kj})) \quad k = 1, 2 \quad (1)$$

where α is a fixed intercept term while the covariate term x_k takes the values 0 and 1 for the first ($k = 1$) and second ($k = 2$) screening tests respectively. The corresponding probability that a patient's test result is classified as not diseased is

$P(Y_{ijk} = 1 | u_{ki}, v_{kj}) = \Phi(\alpha - (\beta x_k + u_{ki} + v_{kj}))$. The fixed coefficient β represents any overall adjustment to the prevalence of binary ratings for the second test when compared to the first test. Terms u_{ki} and v_{kj} represent random effects for the i th patient and the j th expert respectively for the k th medical test. A large positive value for a patient random effect u_{ki} indicates a patient's test result where disease is clearly visible on the image, while a value close to 0 suggests a test result where the disease status is not easily distinguishable relative to other test results. A large positive value for random effect v_{kj} reflects an expert more liberal with classifying patients' test results as more severely diseased relative to other experts. A crossed random effects structure accounts for the study design where each expert rates every patient's test result for each screening medical test, and assumes that the patients and experts are independent of each other. Since the same experts grade the same sample of patients for both tests, random effects for an expert across the two tests are assumed to be correlated, and similarly for the patient random effects as

$$\begin{aligned} \begin{pmatrix} v_{1j} \\ v_{2j} \end{pmatrix} &\sim \text{Bivariate Normal} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_v = \begin{bmatrix} \sigma_{v1}^2 & \rho_v \sigma_{v1} \sigma_{v2} \\ \rho_v \sigma_{v1} \sigma_{v2} & \sigma_{v2}^2 \end{bmatrix} \right) \\ \begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} &\sim \text{Bivariate Normal} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_u = \begin{bmatrix} \sigma_{u1}^2 & \rho_u \sigma_{u1} \sigma_{u2} \\ \rho_u \sigma_{u1} \sigma_{u2} & \sigma_{u2}^2 \end{bmatrix} \right) \end{aligned} \quad (2)$$

The variance components provide valuable information on sources of variability contributing to discrepancies observed between experts' ratings. While either a logit or probit link function are common choices for a binary GLMM, we choose a probit link here to reflect the

assumption that the true underlying disease status of a patient is a latent unobserved continuous normally distributed variable²⁷. In a related GLMM framework setting, both link functions lead to nearly identical results^{28,29}.

A key strength of the binary GLMM framework is its capacity to accommodate large numbers of experts without increasing modeling complexity, and unbalanced data when experts do not rate every patient for each medical test. Incorporating experts' ratings for both screening tests in a single unified modeling approach increases efficiency¹⁴ and avoids conducting separate intra-rater analyses for each individual expert, as is often observed in clinical applications. If experts and patients participating in the study are randomly sampled from their respective populations, conclusions can be generalized to typical experts and patients undergoing these medical procedures.

2.1 Estimation of GLMM parameters

Estimation of the parameter vector $\theta = (\alpha, \beta, \Sigma_u, \Sigma_v)$ using exact maximum likelihood methods is challenging due to the high dimensionality of the marginal likelihood function $L(\theta; Y)$ with a crossed random effects structure in (3). Adaptive Gaussian quadrature, often considered an efficient approximate maximum likelihood estimation approach for GLMMs is not feasible here due to the large numbers of random effects³⁰. Instead, a multivariate Laplacian approach³¹ implemented in the *ordinal* package in R³² provides an efficient and stable strategy for providing a reasonably unbiased estimate of θ for binary classifications. Further available packages for estimation of a binary GLMM include the LME4 and MCMCglmm packages in R and the GLIMMIX procedure in SAS³³.

$$\begin{aligned}
 L(\theta; Y) &= \int_{u,v} L(\theta; u, v, y) \, du \, dv = \int_{u,v} f_{Y|u,v}(y; u, v) f_u(u; \Sigma_u) f_v(v; \Sigma_v) \, du \, dv \\
 &= \int_u \int_v \prod_{i=1}^I \prod_{j=1}^J \prod_{k=1}^K [1 - \Phi(\alpha - (\beta x_k + u_{ki} + v_{kj}))]^{d_{ijk}} \{\Phi(\alpha - (\beta x_k + u_{ki} + v_{kj}))\}^{1-d_{ijk}} \\
 &\quad \times \left[\frac{1}{2\pi} (\det \Sigma_u)^{-1/2} \exp\left(-\frac{1}{2} \mathbf{u}_i' \Sigma_u^{-1} \mathbf{u}_i\right) \right] \\
 &\quad \left[\frac{1}{2\pi} (\det \Sigma_v)^{-1/2} \exp\left(-\frac{1}{2} \mathbf{v}_j' \Sigma_v^{-1} \mathbf{v}_j\right) \right] d\mathbf{u} \, d\mathbf{v}
 \end{aligned} \tag{3}$$

where vectors Y , u and v consist of all the experts' ratings, patient and expert random effects respectively. Indicator term $d_{ijk} = 1$ if $Y_{ijk} = 2$ and 0 otherwise. The issue of identifiability or numerical instability arises in the estimation of the variance-covariance parameters in Σ_u

due to the presence of a bivariate random effects vector in the GLMM model $[u_{1i}, u_{2i}]^{34}$. However, in practice and through our extensive simulation studies, we observe excellent numerical stability in our parameter estimation especially when the sample sizes of raters and patients are not too small when using the multivariate Laplacian approach provided in the ordinal package in R. Large-sample approximate standard errors for parameter estimates are estimated as $se(\hat{\theta}) = \sqrt{\text{diag}[-\{H(\hat{\theta})\}^{-1}]}$ where matrix $H = \partial^2 l(\theta; u, v, y) / \partial \theta \partial \theta^t$ is generated during the model-fitting process.

3. A Paired Kappa to Compare Agreement Between Screening Medical Tests

The comparison of agreement between a newly developed test and existing medical test provides useful information regarding how experts interpret the imaging output from each of the two screening tests for the same patient. A paired kappa focuses on the paired ratings provided by individual experts for the two tests on the same patient. Any noticeable discrepancies or disagreement raised by a low paired kappa value can lead to further investigation into potential distinguishing features of each medical test and how the test results are interpreted in practice.

In the population setting over many experts and patients, a well-defined paired measure of agreement describes how well a typical expert's rating of a patient's test result for one medical test agrees with his or her rating of the patient's result for the second medical test³⁵ after accounting for chance agreement. Our proposed paired kappa is a function of two important elements, observed and chance agreement, which are now defined. We assume that experts and patients are randomly sampled from their respective populations.

3.1 Observed Agreement in the Population-Based Setting

Observed agreement describes the proportion of time an expert typically assigns the same binary rating ($c = 1$ or 2) to the results from both tests for the i th patient. Two ratings are said to be in agreement only if they are equal. A natural measure of observed agreement p_0 in this paired population-based setting is not reliant on any particular model:

$$p_0 = \sum_{c=1}^{C=2} P(Y_{ijk} = c \cap Y_{ijk'} = c) \quad 0 \leq p_0 \leq 1$$

Under the GLMM framework in Section 2, observed agreement p_0 takes the form and takes a value between 0 and 1: (proof and derivations in Appendix)

$$p_0 = \iint_{z_{u1} z_{v1}} \left[\left\{ \Phi(\alpha - (\sigma_{u1} z_{u1} + \sigma_{v1} z_{v1})) \times \Phi\left(\frac{\alpha - \beta - \rho_u \sigma_{u2} z_{u1} - \rho_v \sigma_{v2} z_{v1}}{\sqrt{1 + \sigma_{u2}^2(1 - \rho_u^2) + \sigma_{v2}^2(1 - \rho_v^2)}}\right) \right\} + \right. \\ \left. \{1 - \Phi(\alpha - (\sigma_{u1} z_{u1} + \sigma_{v1} z_{v1}))\} \times \left\{ 1 - \Phi\left(\frac{\alpha - \beta - \rho_u \sigma_{u2} z_{u1} - \rho_v \sigma_{v2} z_{v1}}{\sqrt{1 + \sigma_{u2}^2(1 - \rho_u^2) + \sigma_{v2}^2(1 - \rho_v^2)}}\right) \right\} \right] f(z_{u1}) f(z_{v1}) dz_{u1} dz_{v1} \quad (4)$$

3.2 Chance Agreement

Chance agreement p_C occurs in this paired population-based setting when an expert happens to assign the same rating to two unrelated patients' test results ($i \neq i'$) obtained from two different medical tests ($k \neq k'$), that is, by coincidence, taking into consideration the distribution of classifications for each category. When assessing agreement between an experts' categorical ratings, it is important to consider the impact of chance agreement on the overall assessment of agreement. When classifying patients using a binary scale (two categories), it would be relatively easy for an expert, even an untrained expert to assign the same rating to two unrelated patient test results since only a limited number of options (categories) are available. Not accounting for agreement due to chance in an agreement measure can lead to an inflated estimate of the agreement between an expert's ratings. In Section 3.3 below we demonstrate how the proposed kappa statistic is adjusted for chance agreement. In the population-based setting, chance agreement does not depend upon any specific model formulation:

$$p_C = \sum_{c=1}^C \frac{1}{C} P(Y_{ijk} = c \cap Y_{i'jk'} = c) \quad 0 \leq p_C \leq 1$$

When a GLMM modeling framework is used, paired chance agreement takes the following form for binary ratings and lies between 0 and 1 (proof and derivations in Appendix):

$$p_C = \int_{z_{v1}} \left[\left\{ \Phi\left(\frac{\alpha - \sigma_{v1} z_{v1}}{\sqrt{1 + \sigma_{u1}^2}}\right) \times \Phi\left(\frac{\alpha - \beta - \rho_v \sigma_{v2} z_{v1}}{\sqrt{1 + \sigma_{u2}^2 + \sigma_{v2}^2(1 - \rho_v^2)}}\right) \right\} + \right. \\ \left. \left\{ 1 - \Phi\left(\frac{\alpha - \sigma_{v1} z_{v1}}{\sqrt{1 + \sigma_{u1}^2}}\right) \right\} \times \left\{ 1 - \Phi\left(\frac{\alpha - \beta - \rho_v \sigma_{v2} z_{v1}}{\sqrt{1 + \sigma_{u2}^2 + \sigma_{v2}^2(1 - \rho_v^2)}}\right) \right\} \right] f(z_{v1}) dz_{v1} \quad (5)$$

3.3. Proposed Paired Kappa

We now derive the proposed kappa κ_{paired} to assess the consistency between a typical expert's binary ratings across two medical tests in the population-based setting. A "typical" expert here reflects the expected or averaged probabilities of agreement based upon the behavior of the underlying population of experts he or she is randomly sampled from. The

measure κ_{paired} is based upon observed agreement p_0 in equation (4). To adjust for the impact of chance agreement on the agreement measure we first find the value of the intercept α which minimizes the expression for p_c in equation (5) by setting the derivative of p_c with respect to α to zero. We focus on minimizing p_c with respect to the intercept only, as random effects variance components in the model are expert and patient-specific with an expectation of 0, and coefficient β plays an influential role in determining the magnitude of discrepancy in the overall distribution of classifications (diseased versus non-diseased) between the two screening tests and consequently on the assessment of agreement. The minimized value of the intercept is then incorporated into the expression for κ_{paired} in equation (6):

$$\begin{aligned} \kappa_{paired} = & \int_{z_{u1}} \int_{z_{v1}} \left[\Phi(\alpha_{\min} - (\sigma_{u1}z_{u1} + \sigma_{v1}z_{v1})) \right. \\ & \times \Phi\left(\frac{\alpha_{\min} - \beta - \rho_u\sigma_{u2}z_{u1} - \rho_v\sigma_{v2}z_{v1}}{\sqrt{1 + \sigma_{u2}^2(1 - \rho_u^2) + \sigma_{v2}^2(1 - \rho_v^2)}}\right) + \\ & \left. (1 - \Phi(\alpha_{\min} - (\sigma_{u1}z_{u1} + \sigma_{v1}z_{v1}))) \times \left(1 - \Phi\left(\frac{\alpha_{\min} - \beta - \rho_u\sigma_{u2}z_{u1} - \rho_v\sigma_{v2}z_{v1}}{\sqrt{1 + \sigma_{u2}^2(1 - \rho_u^2) + \sigma_{v2}^2(1 - \rho_v^2)}}\right)\right) \right] \\ & f(z_{u1}) f(z_{v1}) dz_{u1} dz_{v1} \end{aligned} \quad (6)$$

The proposed kappa κ_{paired} takes a value between 0 and 1 and is interpreted in a similar manner to Cohen's kappa³⁶ with one exception. In the unusual scenario where observed agreement is less than chance agreement, which may occur on rare occasions in a small sample, the data-driven Cohen's kappa can take a negative value. In contrast, the minimum value of the proposed kappa κ_{paired} is zero. A kappa close to 0 is interpreted as weak chance-corrected agreement between a typical expert's ratings of a patient's two screening test results, while a value close to 1 is considered very strong chance-corrected agreement. The proposed measure κ_{paired} will always be greater or equal to zero in the population-based setting since in the worst-case scenario, there is zero agreement between an expert's ratings across the two screening tests in a population-based setting. Asymptotically, κ_{paired} will reach its maximum value of 1 (perfect chance-corrected agreement) when all of the following hold: $\beta = 0$, the variance components tend to $+\infty$ and both correlation coefficients ρ_u and ρ_v are 1.0. The measure will reach its minimum value of 0 (zero chance-corrected agreement) 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. The incorporation of the minimized thresholds into κ_{paired} to account for the effects of chance agreement ensures that chance agreement has a minimal impact on the kappa, and the resulting summary measure is robust to the underlying distribution of diseased ratings across the binary categories. This attractive feature of the proposed measure, demonstrated via simulations in Section 4 below, allows for easier

comparison of κ_{paired} across different clinical studies which may exhibit differing distributions of diseased classifications.

For any particular dataset, the estimate $\hat{\kappa}_{paired}$ can be calculated from parameter estimates

$\hat{\theta} = (\hat{\alpha}_{min}, \hat{\beta}, \hat{\Sigma}_u, \hat{\Sigma}_v)$ obtained by fitting the binary GLMM described in Section 2. The

approximate variance of $\hat{\kappa}_{paired}$ is estimated using the multivariate delta method based upon the GLMM model parameters in equation (1) as

$\widehat{\text{var}}(\hat{\kappa}_{paired}) = (\partial \hat{\kappa}_{paired} / \partial \hat{\theta})' \text{var}(\hat{\theta}) (\partial \hat{\kappa}_{paired} / \partial \hat{\theta})$. R code for fitting the proposed approach and a sample dataset are also provided as supplementary materials.

3.4 Properties of the Paired Kappa

Figure 1 presents how parameters β , Σ_u and Σ_v play an important role in determining the value of the paired kappa κ_{paired} . Agreement between an expert's paired ratings for two screening tests is strongest when β is 0 which occurs when the overall probabilities of a patient's test result being classified as diseased is the same for both screening tests. Larger absolute values of β lead to increasingly different probabilities for a patient's two test results to be assigned as diseased, leading to weaker consistency between an expert's ratings.

Figure 1(a) displays how increasing the correlation between an experts' paired ratings ρ_v impacts κ_{paired} over a wide range of values for β when other terms remain constant

$(\sigma_{u1}^2 = 10, \sigma_{u2}^2 = 10, \sigma_{v1}^2 = 8, \sigma_{v2}^2 = 8, \rho_u = 0.5)$. A stronger correlation between an expert's

ratings between screening tests, ρ_v , leads to stronger agreement, although this effect is mitigated as β moves further away from 0. A zero value for correlation ρ_v indicates that an expert typically assigns independent ratings to a patient's two test results, and leads to poor consistency. While we do not anticipate a negative expert correlation ρ_v (where an expert

often provides conflicting ratings for a patient's two test results) in the population-based setting, if that occurred, the agreement decreases further. Figure 1(b) presents the impact of increasing variances for all patient and expert random effects on κ_{paired} as β changes (for fixed expert and patient correlations of $\rho_u = 1.00$ and $\rho_v = 1.00$ respectively). We observe that the paired kappa yields the strongest chance-corrected agreement (κ_{paired} achieves close to the maximum value of 1) between a typical expert's ratings for the two screening tests when β is zero and variance components and correlations are very large. In this situation, the disease status of patients' test results are clearly distinguishable from one another. Figures

1(c) displays changes in the proposed paired kappa for increasing patient variances respectively while keeping the other parameters constant. As β moves away from 0 in Figure 1(c) for fixed $\sigma_{v1}^2 = 5, \sigma_{v2}^2 = 5, \rho_u = 0.75, \rho_v = 0.5$ the sizes of the patient variances

σ_{u1}^2 and σ_{u2}^2 make a noticeable impact on the paired kappa value, where larger values of

σ_{u1}^2 and σ_{u2}^2 are linked with screening test results that display an increasingly diverse range of disease clarity, improving the agreement between an expert's paired ratings of the same patient. We demonstrate this approach in a large breast imaging study dataset in Section 6.

3.4 The Impact of Disease Prevalence

The proposed paired kappa is robust to the overall distribution of experts' diseased ratings, which reflect the underlying prevalence of disease. To demonstrate this feature, we selected three different prevalences of diseased ratings: rare disease, equal disease and common disease ($\alpha = -3, 0$ and 3) respectively. Figure 1(d) presents the resulting κ_{paired} values for these three disease prevalences. Each line displays the κ_{paired} values for a fixed set of parameters varying the patient variances σ_{u1}^2 and σ_{u2}^2 , allowing the disease prevalences to change. We observe that the $\hat{\kappa}_{paired}$ values are unaffected by changing disease prevalence.

4. Simulation Studies

4.1 Simulation of Datasets with Normally Distributed Random Effects

Extensive simulation studies were conducted to evaluate the performance of the proposed kappa measure under a diverse range of scenarios as presented in Table 1. Simulated datasets were generated as follows. We assume each patient undergoes two separate screening tests conducted very close in time, usually at the same clinical visit. Sets of correlated patient random effects from the two screening tests (u_{1j}, u_{2j}) and correlated expert random effects (v_{1j}, v_{2j}) ($i = 1, \dots, I, j = 1, \dots, J$) are randomly generated from their respective bivariate normal distributions defined in equation (2) using the *rvnorm* function in R. Binary ratings $Y_{ijk} = c$ ($c = 1$ or 2) are then randomly generated for each expert-patient combination for each screening test ($k = 1, 2$) using the *rmultinom* function in R based upon the GLMM model probabilities in equation (1). For each scenario, one thousand simulation datasets are randomly generated. For each set of simulations, the true value of the proposed kappa measure κ_{paired} is calculated based upon the known values of the parameter vector $\theta = (\alpha, \beta, \Sigma_u, \Sigma_v)$ from which the simulated datasets are generated.

The intercept was set at $\alpha = 1$, and $\beta = -0.10$ or 1.5 to reflect estimates in the breast imaging study in Section 5.

4.2 Simulation of Datasets with Non-normally Distributed Random Effects

A key assumption of the GLMM modeling approach is that random effects are normally distributed. To evaluate the impact on the estimation of the paired kappa when this assumption is violated, we generated correlated patient random effect vectors (u_{1j}, u_{2j}) according to a multivariate beta distribution which is very skewed and non-normal following the approach outlined in Michael and Schucany³⁷. First we randomly generate a set of patient random effects for one of the screening tests u_{1i} ($i = 1, \dots, I$) from a prior beta distribution with shape parameters a and b using the *rbeta* function in R. The u_{1i} values lie between 0 and 1. To generate the set of correlated random effects u_{2i} for the second screening test from a posterior beta distribution, we pre-specify the correlation coefficient ρ_u and set $v = (a + b)\rho_u / (1 - \rho_u)$. Term v represents the amount of information provided by the likelihood and is calculated as the required binomial sample size for the correlation specified. We then create a binomial sample k_i from a binomial distribution with sample size v and probability corresponding to the patient's random effect value u_{1i} using the *rbinom* function in R. Finally, random effects u_{2i} are randomly generated from a posterior beta

distribution with shape parameters $a + k_i$ and $b + v - k_i$ conditional on k_i . For example, if we set the shape parameters $a = 1$ and $b = 4$ and Correlation $\rho_u = 0.5$ then random effects $u_{1i} \sim \text{rbeta}(1, 4)$ and $v = (1 + 4) \times 0.5 / (1 - 0.5) = 5$. Binomial samples are generated as $k_i \sim \text{rbinom}(1, 5, u_{1i})$ and $u_{2i} \sim \text{rbeta}(1 + k_i, 4 + 5 - k_i)$. The generated correlated sets of patient random effects (u_{1i}, u_{2i}) are then scaled to have mean zero and the required variance. Some sets of simulations conducted were based on parameters very similar to the breast imaging dataset described in Section 5 below, and further sets of simulations to provide a more diverse view of the proposed method's behavior.

4.3 Simulation Study Results

For each set of one thousand simulations, we estimated the parameter vector $\theta = (\alpha, \beta, \Sigma_u, \Sigma_v)$ by fitting the GLMM using the *ordinal* package in R, and calculated the paired kappa $\hat{\kappa}_{\text{paired}}$ and variance $\text{var}(\hat{\kappa}_{\text{paired}})$. Table 2 summarizes the mean $\hat{\kappa}_{\text{paired}}$ and its average calculated standard error over the 1000 simulated datasets for each simulation scenario. We observe that the proposed paired kappa κ_{paired} is estimated with negligible bias in each case with normal and non-normally distributed random effects (slightly better for normally distributed random effects). There are marginal improvements in bias for larger sample sizes of experts and patients.

Full simulation results are presented in the Supplementary Materials Tables (S1–S5). Mean parameter estimates were calculated as the average of 1000 simulated estimates for each parameter. Variability was assessed using the mean standard error (average value of the estimated standard error over 1000 simulated datasets) and empirically as the variance of 1000 corresponding parameter estimates. All parameters in $\theta = (\alpha, \beta, \Sigma_u, \Sigma_v)$ were estimated with minimal bias in each simulation scenarios with normally distributed random effects in Tables S1 and S2 for binary ratings. The fixed intercept α and coefficient β were estimated with moderate bias for small sample sizes of experts and patients for non-normally distributed patient random effects (Tables S3 and S4), however this bias improved with the larger sample sizes of experts and patients. Variances were often estimated with a small amount of bias for non-normally distributed patient random effects, which generally improved at the larger sample sizes, while correlations ρ_u and ρ_v were estimated with minimal bias throughout the simulation studies.

Coverage probabilities of $\hat{\kappa}_{\text{paired}}$, defined as the percent of 95% confidence intervals that include the true value of κ_{paired} , are summarized in the Supplementary Materials Table S5. In all simulation scenario the coverage probabilities take reasonable values, ranging from 89.5% to 100%. Some slight skewness in the $\hat{\kappa}_{\text{paired}}$ estimates is observed for the non-normal simulations leading to slightly lower 95% coverage probabilities.

5. Breast Imaging Study Comparing Film versus Digital Mammography

5.1 The ACRIN Breast Imaging Study

The Digital Mammographic Imaging Screening Trial (DMIST) is a nationwide breast screening study conducted by the American College of Radiology from 2001 to 2003 to

compare the diagnostic accuracy of film versus digital mammography³⁸. Participants underwent both digital and film mammography in random order. The two screening tests were performed at the same clinical visit for each patient. As part of this study, 12 radiologists independently classified the film and digital mammograms of a sample of 75 (female) patients at two randomly ordered reading sessions at least six weeks apart between interpretations acquired with the two tests to avoid recall bias. Binary ratings were assigned by radiologists as defined in Hendrick et al⁷ where a negative test interpretation indicates no malignancy and a positive test interpretation indicates the presence of malignancy.

The binary GLMM model in (1) was fitted to the DMIST reader agreement dataset in R using the *clmm* function in the *ordinal* package. The classification $Y_{ijk} = c$ refers to the binary rating ($c = 1, 2$) assigned by radiologist j ($j = 1, \dots, 12$) on the i th patient's test result ($i = 1, \dots, 75$) obtained from the k th screening test ($k = 1$ film mammography; $k = 2$ digital mammography). The authors' R functions to calculate the proposed model-based paired kappa are available in Supplementary materials.

5.2 Results for the Mammogram Study

Table 3 displays the estimated parameters and standard errors for the DMIST Reader Study. We observe that the fixed coefficient $\hat{\beta} = 0.203$ ($s.e. = 0.202$) is near 0 indicating that the overall distribution of probabilities of a test result being classified as positive is slightly higher for digital mammography ($x_k = 1$). However, a Wald hypothesis test conducted to test the hypothesis $H_0: \beta = 0$ suggests no statistically significant difference (i.e. no overall "shift") between the probabilities of classifications for film and digital mammography in this study ($p\text{-value} = 0.687$). Patient random effect variance components $\hat{\sigma}_{u1}^2 = 2.312$ and $\hat{\sigma}_{u2}^2 = 1.857$ for film and digital mammography respectively suggest that disease severity has reasonably similar clarity for both digital and film images. There is a strong estimated correlation $\hat{\rho}_u = 0.991$ between a patient's film and digital mammograms ratings. Expert random effect variances estimates $\hat{\sigma}_{v1}^2$ and $\hat{\sigma}_{v2}^2$ are smaller (0.443 and 0.414 respectively) suggesting little variability between radiologists' ratings for each screening test, that is, no radiologist was excessively liberal or conservative in assigning higher disease scores relative to the other radiologists in the study. The moderately strong estimated correlation $\hat{\rho}_v = 0.638$ between a radiologist's paired ratings for the two screening tests indicates that radiologists often assigned different scores to a patient's film and digital mammograms. The proposed paired kappa is estimated as $\hat{\kappa}_{paired} = 0.729$ ($s.e. = 0.049$) with a 95% confidence interval of (0.632, 0.826) indicating moderately strong chance-corrected agreement between a typical radiologist's binary assessments of a randomly selected patient's two screening test results (film and digital), after accounting for the correlations present between all the ratings. Our overall conclusion for this breast imaging reader study is that radiologists demonstrate good agreement between their binary ratings of the same patient's two mammograms after accounting for chance agreement, although some discrepancies exist in experts' interpretation between the two screening tests.

The GLMM fitting process provides conditional modes of the random effects for patients and experts for each screening test, presented in Figure 2. Experts exhibited similar

variability in their ratings of both the film mammography and digital mammography images. Similar dispersions of ratings for patients' test results is present for both digital and film mammography. Four patients had large positive outlying random effects for both film and digital mammography (for example, $\hat{u}_{1,19} = 3.43$, $\hat{u}_{2,19} = 3.07$) reflecting mammograms with clearly visible disease status where the majority of experts assigned very high disease scores to each of these images.

6. Testing Goodness-of-fit

To ensure the binary GLMM framework in equation (1) provides an appropriate modeling strategy for incorporating the correlated ratings from the two screening tests, goodness-of-fit methods proposed by Hedeker and Gibbons³⁹ and Todem et al.⁴⁰ are applied to the breast imaging dataset in Section 6. This approach involves comparing the number of mammograms classified into each category by experts in the study (observed counts) with their corresponding marginal counts predicted from the binary GLMM (estimated counts) for each screening test separately. Marginal probabilities for the k th screening test ($k = 1$ film, $k = 2$ digital mammography) of being classified into each binary categories ($c = 1, 2$) are calculated as:

$$P(Y_{ijk} = 1) = \Phi\left(\frac{\alpha - \beta x_k}{\sqrt{1 + \sigma_{uk}^2 + \sigma_{vk}^2}}\right) \text{ and } P(Y_{ijk} = 2) = 1 - \Phi\left(\frac{\alpha - \beta x_k}{\sqrt{1 + \sigma_{uk}^2 + \sigma_{vk}^2}}\right) \quad k = 1, 2.$$

Observed and estimated counts of mammograms are presented in Table 4. We see that the estimated marginal frequencies closely align with their corresponding observed frequencies for each of the two screening tests, indicating that the binary GLMM framework provides a good model fit for this breast imaging reader study.

7. Discussion

The proposed paired kappa as a summary measure provides valuable insight into the strength of agreement between a typical expert's ratings across two screening tests. In large-scale agreement studies there is an important need for statistical methods to compare the agreement between test results produced by two screening tests, especially between a new and existing screening test, when many experts contribute ratings. In the setting of breast cancer screening, early detection of breast cancers through screening mammography may lead to a reduced mortality rate and less aggressive treatments⁴¹. However, a downside of widespread cancer screening is unnecessary further testing and anxiety that may be triggered by a false-positive finding on a mammogram. As advances in technology lead to new and updated imaging tests, investigation into the consistency of ratings that an expert provides from their interpretation of patient test results from two different screening tests such as film and digital mammography provides valuable insight into their implementation into current clinical practice, where strong agreement of ratings between experts is an important and necessary prerequisite of an accurate and effective screening procedure.

Current approaches for comparing correlated binary ratings between two screening tests are restricted to a very small number of experts (often just two experts), are often computationally intensive and are prone to flaws often observed in the use of Cohen's kappa and Scott's pi. The novel paired kappa proposed here provides an easily interpretable measure of chance-corrected agreement of a typical expert's ratings for two screening tests where simple measures of agreement such as Cohen's kappa cannot be used. A large number of experts' binary ratings can be incorporated without increasing modeling complexity. Further strengths include appropriately accounting for the correlation structure between the experts' ratings and being robust to the underlying disease prevalence. Incorporating all experts' ratings into one unified modeling approach increases the efficiency and power of the analysis, and avoids the need for stratification. Through extensive simulation studies the paired kappa is demonstrated to be robust to model violations including skewed random effect distributions and has very good coverage probability.

The approaches described in this paper can be easily and quickly implemented in the software package R. R code along with a sample dataset is provided in the supplementary materials for fitting the methods presented in this paper.

The proposed paired kappa measure is used to compare the agreement of ratings provided by two different screening tests in the common situation where neither test provides a "gold standard" measurement of disease status. In some clinical assessments, a gold standard measurement may not exist (for example, breast density, psychology assessments, skin conditions), or is not easily obtainable. Diagnostic accuracy and agreement between experts' ordinal binary ratings are both important characteristics of an effective screening test. Conventional approaches for comparing the diagnostic accuracy between two screening tests when a gold standard measurement exists in a study are based upon sensitivity and specificity⁴². Latent variable models provide a potential approach for assessing diagnostic accuracy of a test when no gold standard exists, although weaknesses and a lack of robustness are often evident in the implementation of these methods^{43,44}. Other authors^{42,45} describe methods for providing insights into the accuracy of a test while taking the variability between raters into consideration. When a gold standard measurement of disease is available, our proposed approach for comparing the agreement between ratings from two screening tests can be used in conjunction with methods to study accuracy to provide valuable insight into understanding potential sources of variability that lead to inconsistent classifications.

A future research goal includes extending the kappa measure to compare results between more than two medical tests.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

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 the American College of Radiology Imaging Network for kindly

providing us with their Digital Mammographie Imaging Screening Trial study dataset. Original data collection for the ACRIN 6652 trial was supported by NCI Cancer Imaging Program grants. We appreciate Aya Mitani's and Thomas Zhou's assistance with the R code.

Appendix A

Derivation of observed paired agreement p_0 (equation 4)

$p_0 = \sum_{c=1}^C \left[P(Y_{ijk} = c \cap Y_{ijk'} = c) \right]$ where screening tests $k = 1$ and $k' = 2$ and $C = 2$ (binary ratings)

$$\begin{aligned}
 &= \sum_{c=1}^2 E \left\{ \left[\Phi(\alpha_c - (\beta x_{k=1} + u_1 + v_1)) - \Phi(\alpha_{c-1} - (\beta x_{k=1} + u_1 + v_1)) \right] \cap \right. \\
 &\quad \left. \left[\Phi(\alpha_c - (\beta x_{k=2} + u_2 + v_2)) - \Phi(\alpha_{c-1} - (\beta x_{k=2} + u_2 + v_2)) \right] \right\} \\
 &= \int_{v_1} \int_{u_1} \sum_{c=1}^2 \left\{ \left[P(Q_1 \leq \alpha_c - (u_1 + v_1)) - P(Q_1 \leq \alpha_{c-1} - (u_1 + v_1)) \right] \times \right. \\
 &\quad \left. \left[P(Q_2 + u_2 + v_2 \leq \alpha_c - \beta | u_1, v_1) - P(Q_2 + u_2 + v_2 \leq \alpha_{c-1} - \beta | u_1, v_1) \right] \right\} f(u_1) f(v_1) du_1 dv_1 \\
 &= \int_{v_1} \int_{u_1} \sum_{c=1}^2 \left\{ \left[\Phi(\alpha_c - (u_1 + v_1)) - \Phi(\alpha_{c-1} - (u_1 + v_1)) \right] \times \right. \\
 &\quad \left. \left[\Phi \left(\frac{\alpha_c - \beta - \frac{\rho_u \sigma_{u2} u_1}{\sigma_{u1}} - \frac{\rho_v \sigma_{v2} v_1}{\sigma_{v1}}}{\sqrt{1 + \sigma_{u2}^2 (1 - \rho_u^2) + \sigma_{v2}^2 (1 - \rho_v^2)}} \right) - \Phi \left(\frac{\alpha_{c-1} - \beta - \frac{\rho_u \sigma_{u2} u_1}{\sigma_{u1}} - \frac{\rho_v \sigma_{v2} v_1}{\sigma_{v2}}}{\sqrt{1 + \sigma_{u2}^2 (1 - \rho_u^2) + \sigma_{v2}^2 (1 - \rho_v^2)}} \right) \right] \right\} f(u_1) f(v_1) du_1 dv_1 \\
 &= \int_{z_{v1}} \int_{z_{u1}} \left\{ \left[\Phi(\alpha - (\sigma_{u1} z_{u1} + \sigma_{v1} z_{v1})) \times \Phi \left(\frac{\alpha - \beta - \rho_u \sigma_{u2} z_{u1} - \rho_v \sigma_{v2} z_{v1}}{\sqrt{1 + \sigma_{u2}^2 (1 - \rho_u^2) + \sigma_{v2}^2 (1 - \rho_v^2)}} \right) \right] + \right. \\
 &\quad \left. \left[1 - \Phi(\alpha - (\sigma_{u1} z_{u1} + \sigma_{v1} z_{v1})) \right] \times \left[1 - \Phi \left(\frac{\alpha - \beta - \rho_u \sigma_{u2} z_{u1} - \rho_v \sigma_{v2} z_{v1}}{\sqrt{1 + \sigma_{u2}^2 (1 - \rho_u^2) + \sigma_{v2}^2 (1 - \rho_v^2)}} \right) \right] \right\} f(z_{u1}) f(z_{v1}) dz_{u1}
 \end{aligned}$$

when rewritten in terms of the standard normal integrands as $z_{u1} = u_1 / \sigma_{u1}$ and $z_{v1} = v_1 / \sigma_{v1}$.

Derivation of chance paired agreement p_c (equation 5):

$$\begin{aligned}
p_C &= \sum_{c=1}^{C=2} \left[P(Y_{ijk} = c \cap Y_{ijk'} = c) \right] \\
&= \sum_{c=1}^2 E \left\{ \left[\Phi(\alpha_c - (\beta x_{k=1} + u_1 + v_1)) - \Phi(\alpha_{c-1} - (\beta x_{k=1} + u_1 + v_1)) \right] \cap \right. \\
&\quad \left. \left[\Phi(\alpha_c - (\beta x_{k=2} + u_2 + v_2)) - \Phi(\alpha_{c-1} - (\beta x_{k=2} + u_2 + v_2)) \right] \right\} \\
&= \int_{v_1} \sum_{c=1}^2 \left\{ \left[P(Q_1 + u_1 \leq \alpha_c - (\beta x_{k=1} + v_1)) - P(Q_1 + u_1 \leq \alpha_{c-1} - (\beta x_{k=1} + v_1)) \right] \times \right. \\
&\quad \left. \left[P(Q_2 + u_2 + v_2 \leq \alpha_c - \beta x_{k=2} | v_{1j}) - P(Q_2 + u_2 + v_2 \leq \alpha_{c-1} - \beta x_{k=2} | v_{1j}) \right] \right\} f(v_1) dv_1 \\
&= \int_{v_1} \sum_{c=1}^2 \left\{ \left[P\left(\frac{Q_1 + u_1}{\sqrt{1 + \sigma_{u1}^2}} \leq \frac{\alpha_c - v_1}{\sqrt{1 + \sigma_{u1}^2}}\right) - P\left(\frac{Q_1 + u_1}{\sqrt{1 + \sigma_{u1}^2}} \leq \frac{\alpha_{c-1} - v_1}{\sqrt{1 + \sigma_{u1}^2}}\right) \right] \times \right. \\
&\quad \left[P\left(\frac{Q_2 + u_2 + v_2}{\sqrt{1 + \sigma_{u2}^2 + \sigma_{v2}^2(1 - \rho_v^2)}} \leq \frac{\alpha_c - \beta}{\sqrt{1 + \sigma_{u2}^2 + \sigma_{v2}^2(1 - \rho_v^2)}} \middle| v_1 \right) - \right. \\
&\quad \left. \left. P\left(\frac{Q_2 + u_2 + v_2}{\sqrt{1 + \sigma_{u2}^2 + \sigma_{v2}^2(1 - \rho_v^2)}} \leq \frac{\alpha_{c-1} - \beta}{\sqrt{1 + \sigma_{u2}^2 + \sigma_{v2}^2(1 - \rho_v^2)}} \middle| v_1 \right) \right] \right\} f(v_1) dv_1 \\
&= \int_{v_1} \sum_{c=1}^2 \left\{ \left[\Phi\left(\frac{\alpha_c - v_1}{\sqrt{1 + \sigma_{u1}^2}}\right) - \Phi\left(\frac{\alpha_{c-1} - v_1}{\sqrt{1 + \sigma_{u1}^2}}\right) \right] \times \right. \\
&\quad \left[\Phi\left(\frac{\alpha_c - \beta - \frac{\rho_v \sigma_{v2} v_1}{\sigma_{v1}}}{\sqrt{1 + \sigma_{u2}^2 + \sigma_{v2}^2(1 - \rho_v^2)}}\right) - \Phi\left(\frac{\alpha_{c-1} - \beta - \frac{\rho_v \sigma_{v2} v_1}{\sigma_{v1}}}{\sqrt{1 + \sigma_{u2}^2 + \sigma_{v2}^2(1 - \rho_v^2)}}\right) \right] \right\} f(v_1) dv_1 \\
&= \int_{z_{v1}} \left\{ \left[\Phi\left(\frac{\alpha - \sigma_{v1} z_{v1}}{\sqrt{1 + \sigma_{u1}^2}}\right) \times \Phi\left(\frac{\alpha - \beta - \rho_v \sigma_{v2} z_{v1}}{\sqrt{1 + \sigma_{u2}^2 + \sigma_{v2}^2(1 - \rho_v^2)}}\right) \right] + \right. \\
&\quad \left. \left[1 - \Phi\left(\frac{\alpha - \sigma_{v1} z_{v1}}{\sqrt{1 + \sigma_{u1}^2}}\right) \right] \times \left[1 - \Phi\left(\frac{\alpha - \beta - \rho_v \sigma_{v2} z_{v1}}{\sqrt{1 + \sigma_{u2}^2 + \sigma_{v2}^2(1 - \rho_v^2)}}\right) \right] \right\} f(z_{v1}) dz_{v1}
\end{aligned}$$

Appendix B

Agreement between two binary tests: range of p_0 is 0 p_0 1

Begin with equation (4):

$$\begin{aligned}
 p_0 &= \sum_{c=1}^2 \Pr[(Y_{ijk} = c) \cap (Y_{ijk'} = c)] \\
 &= \int_{z_{u1}} \int_{z_{v1}} \Phi(\alpha - (\sigma_{u1}z_{u1} + \sigma_{v1}z_{v1})) \times \Phi\left(\frac{\alpha - \beta - \rho_u\sigma_{u2}z_{u1} - \rho_v\sigma_{v2}z_{v1}}{\sqrt{1 + \sigma_{u2}^2(1 - \rho_u^2) + \sigma_{v2}^2(1 - \rho_v^2)}}\right) \\
 &\quad \left\{ + \right. \\
 &\quad \left. \left\{ 1 - \Phi(\alpha - (\sigma_{u1}z_{u1} + \sigma_{v1}z_{v1})) \right\} \times \left\{ 1 - \Phi\left(\frac{\alpha - \beta - \rho_u\sigma_{u2}z_{u1} - \rho_v\sigma_{v2}z_{v1}}{\sqrt{1 + \sigma_{u2}^2(1 - \rho_u^2) + \sigma_{v2}^2(1 - \rho_v^2)}}\right) \right\} \right\} f \\
 &\quad (z_{u1})f(z_{v1})dz_{u1} dz_{v1}
 \end{aligned} \tag{4}$$

The integral can be approximated to arbitrary accuracy limiting z_{u1} and z_{v1} to a finite range, so these values can be considered to be bounded. We note that as

$\alpha \rightarrow +\infty$ and $(\alpha - \beta) \rightarrow \infty$, (e.g. let $\alpha = 2\beta$ and take $\beta \rightarrow \infty$), the first product in the integrand $\rightarrow 1$ and the second product $\rightarrow 0$, so $p_0 \rightarrow 1$. Similarly as

$\alpha \rightarrow -\infty$ and $(\alpha - \beta) \rightarrow \infty$, (e.g. let $\beta = 2\alpha$ and take $\alpha \rightarrow -\infty$) the first product $\rightarrow 0$ and the second $\rightarrow 0$, so $p_0 \rightarrow 0$. Therefore, 0 p_0 1 and both endpoints are achievable in the limit.

Agreement between two tests: range of p_c is 0 p_c 1

Begin with equation (5):

$$\begin{aligned}
 p_c &= \sum_{c=1}^{C=2} \Pr[(Y_{ijk} = c) \cap (Y_{i'jk'} = c)] = \\
 &\int_{z_1} \left\{ \left[\Phi\left(\frac{\alpha - \sigma_{v1}z_{v1}}{\sqrt{1 + \sigma_{u1}^2}}\right) \times \Phi\left(\frac{\alpha - \beta - \rho_v\sigma_{v2}z_{v1}}{\sqrt{1 + \sigma_{u2}^2 + \sigma_{v2}^2(1 - \rho_v^2)}}\right) \right] + \right. \\
 &\quad \left. \left[1 - \Phi\left(\frac{\alpha - \sigma_{v1}z_{v1}}{\sqrt{1 + \sigma_{u1}^2}}\right) \right] \times \left[1 - \Phi\left(\frac{\alpha - \beta - \rho_v\sigma_{v2}z_{v1}}{\sqrt{1 + \sigma_{u2}^2 + \sigma_{v2}^2(1 - \rho_v^2)}}\right) \right] \right\} f(z_{v1}) dz_{v1}
 \end{aligned}$$

Again, the integral can be approximated to arbitrary accuracy limiting z_{v1} to a finite range, so the values of z_{v1} can be considered to be bounded. Note that as

$\alpha \rightarrow \infty$ and $(\alpha - \beta) \rightarrow \infty$, (e.g. let $\alpha = 2\beta$ and take $\beta \rightarrow \infty$), the first product in the integrand $\rightarrow 1$

and the second $\rightarrow 0$, so $p_c \rightarrow 1$. Similarly as

$\alpha \rightarrow -\infty$ and $(\alpha - \beta) \rightarrow \infty$, (e.g. let $\beta = 2\alpha$ and take $\alpha \rightarrow -\infty$) the first product $\rightarrow 0$ and the second $\rightarrow 0$, so $p_c \rightarrow 0$. Therefore $0 \leq p_c \leq 1$ and both endpoints are achievable in the limit.

REFERENCES

1. Elmore JG, Longton GL, Carney PA, Geller BM, Onega T, Tosteson A, Nelson HD, Pepe MS, Allison KH, Schnitt SJ, O'Malley FP, and Weaver DL. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA* 2015; 313: 1122–1132. [PubMed: 25781441]
2. Conant EF, Beaber EF, Sprague BL, Herschorn SD, Weaver DL, Onega T, Tosteson ANA, McCarthy AM, Poplack SP, Haas JS, Armstrong K, Schnall MD, and Barlow WE. Breast cancer screening using tomosynthesis in combination with digital mammography compared to digital mammography alone: a cohort study within the PROSPR consortium. *Breast Cancer Research and Treatment* 2016; 156: 109–116. [PubMed: 26931450]
3. Gonin R, Lipsitz SR, Fitzmaurice GM, Molenberghs G, Gonin R. Regression modelling of weighted κ by using generalized estimating equations. *Journal of Royal the Statistical Society Series C (Applied Statistics)* 2000; 49: 1–18.
4. Nair AK, Gavett BE, Damman M, Dekker W, Green RC, Mandel A, Auerbach S, Steinberg E, Hubbard EJ, Jefferson A and Stern RA. Clock drawing test ratings by dementia specialists: Interrater reliability and diagnostic accuracy. *Journal of Neuropsychiatry Clinical Neurosciences* 2010; 22: 85–92. [PubMed: 20160214]
5. Friedewalde SM, Rafferty EA, Rose SL, Durand MA, Plecha DM, Greenberg JS, Hayes MK, Copit DS, Carlson KL, Cink TM, Barke LD, Greer LN, Miller DP, Conant EF. Breast cancer screening using tomosynthesis in combination with digital mammography. *JAMA* 2014; 311: 2499–2507. [PubMed: 25058084]
6. Beam CA, Conant EF and Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. *Journal of the National Cancer Institute*; 2003; 95: 282–290. [PubMed: 12591984]
7. Hendrick RE, Cole EB, Pisano Etta D, Acharyya S, Marques H, Cohen MA, Jong RA, Mawdsley GE, Kanal KM, D'Orsi CJ, Bebnar M, and Gatsonis C. Accuracy of soft-copy digital mammography versus that of screen-film mammography according to digital manufacturer: ACRIN DMIST retrospective multireader study. *Radiology* 2008; 24: 38–48.
8. Gard CC, Aiello Bowles EJ, Miglioretti DL, Taplin SH and Rutter CM. Misclassification of Breast Imaging Reporting and Data System (BIRADS) mammographic density and implications for breast density reporting legislation. *Breast Journal* 2015; 21: 481–489. [PubMed: 26133090]
9. Raza S, Mackesy MM, Winkler NS, Hurwitz S and Birdwell RL. Effect of training on qualitative mammographic density assessment. *Journal of the American College of Radiology* 2016; 13: 310–315. [PubMed: 26944039]
10. Dixon SN, Donner A and Shoukri MM. Adjusted inference procedures for the interobserver agreement in twin studies. *SMMR* 2016; 25: 1260–1271.
11. Williamson JM, Lipsitz SR, and Manatunga AK. Modeling kappa for measuring dependent categorical agreement data. *Biostatistics* 2000; 1:191–202. [PubMed: 12933519]
12. McKenzie DP, MacKinnon AJ, Peladeau N, Onghena P, Bruce PC, Clarke DM, Harrigan S, and McGorry PD. Comparing correlated kappas by resampling: is one level of agreement significantly different from another? *Journal of Psychiatric Research* 1996; 30: 483–492. [PubMed: 9023792]
13. Barnhart HX and Williamson JM. Weighted least-squares approach for comparing correlated kappa. *Biometrics* 2002; 58: 1012–1019. [PubMed: 12495157]
14. Donner A, Shoukri MM, Klar N, and Bartfay E. Testing the equality of two dependent kappa statistics. *Statistics in Medicine* 2000; 19: 373–387. [PubMed: 10649303]
15. Klar N, Lipsitz SR and Ibrahim JG. An estimating equations approach for modelling kappa. *Biometrical Journal* 2000; 42: 45–58.
16. Vanbelle S and Albert A A bootstrap method for comparing correlated kappa coefficients. *Journal of Statistical Computation and Simulation* 2008; 78: 1009–1015.

17. Vanbelle S and Albert A. Agreement between two independent groups of raters. *Biometrika* 2009; 74: 477–491.
18. Cohen J A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960; 20: 37–46.
19. Scott WA. Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly* 1955; 19: 321–325.
20. Oden NL. Estimating kappa from binocular data. *Statistics in Medicine* 1991; 10: 1303–1311. [PubMed: 1925161]
21. Schouten HJA. Measuring pairwise interobserver agreement when all subjects are judged by the same observers. *Statistica Neerlandica* 1982; 36: 45–61.
22. Hripcsak G and Heitjan DF. Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics* 2002; 35: 99–110. [PubMed: 12474424]
23. Maclure M and Willett WC. Misinterpretation and misuse of the Kappa statistic. *American Journal of Epidemiology* 1987; 126: 161–169. [PubMed: 3300279]
24. Byrt T, Bishop J and Carlin JB. Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 1993; 46: 423–429. [PubMed: 8501467]
25. Ooms EA, Zonderland HM, Eijkemans MJC, Kriege M, Mahdavian Delavary B, Burger CW, and Ansink AC. Mammography: Interobserver variability in breast density assessment. *The Breast* 2007; 16: 568–576. [PubMed: 18035541]
26. Comprat E, Egevad L, Lopez-Beltran A, Camparo P, Algaba F, Amin M, Epstein JI, Hamberg H, Hulsbergen-van de Kaa C, Kristiansen G, Montironi R, Pan C-C, Heloir F, Treurniet K, Sykes J and Van der Kwast TH. An interobserver reproducibility study on invasiveness of bladder cancer using virtual microscopy and heatmaps. *Histopathology* 2013; 63: 756–766. [PubMed: 24102813]
27. Liu I and Agresti A. The analysis of ordered categorical data: an overview and a survey of recent developments. *Test* 2005; 14: 1–73.
28. Nelson KP and Edwards D. On population-based measures of agreement for binary classifications. *The Canadian Journal of Statistics* 2008; 36: 411–426.
29. Nelson KP and Edwards D. Measures of agreement between many raters for ordinal classifications. *Statistics in Medicine* 2015; 34(23): 3116–3132. [PubMed: 26095449]
30. Gueorguieva R A multivariate generalized linear mixed model for joint modeling of clustered outcomes in the exponential family. *Statistical Modeling* 2001; 1: 177–93.
31. Shun Z and McCullagh P. Laplace Approximation of High Dimensional Integrals. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995; 57: 749–760.
32. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria URL, <http://www.R-project.org/>. 2015.
33. Mitani AA and Nelson KP. Modeling agreement between binary classifications of multiple raters in R and SAS. *Journal of Modern Applied Statistical Methods* 2017; 16: 277–309.
34. Lavielle M and Aarons L What do we mean by identifiability in mixed effects models? *Journal of Pharmacokinetics and Pharmacodynamics* 2016; 43: 111–122. [PubMed: 26660913]
35. Bloch DA and Kramer HC. 2×2 kappa coefficients: agreement or association. *Biometrics* 1989; 45: 269–287. [PubMed: 2655731]
36. Landis JR and Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174. [PubMed: 843571]
37. Michael JR and Schucany WR. The mixture approach for simulating bivariate distributions with specified correlations. *The American Statistician* 2002; 56: 48–54.
38. Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, Conant EF, Fajardo LL, Bassett L, D’Orsi C, Jong R, Rebner M and the Digital Mammographic Imaging Screening Trial (DMIST) Investigators. Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening. *New England Journal of Medicine* 2005; 353: 1773–1783. [PubMed: 16169887]
39. Hedeker D and Gibbons RD (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* 50, 933–44. [PubMed: 7787006]

40. Todem D, Kim K, and Lesaffre E. Latent-variable models for longitudinal data with bivariate ordinal outcomes. *Statistics in Medicine* 2007; 26: 1034–1054. [PubMed: 16832841]
41. Siu AL Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. *Annals of Internal Medicine* 2016; 164: 279–296. [PubMed: 26757170]
42. Zhou X-H, Obuchowski NA and McClish DK (2011). *Statistical methods in diagnostic medicine*. New Jersey: Wiley & Co., 2011.
43. Hui SL and Walter SD. Estimating the error rates of diagnostic tests. *Biometrics* 1980; 36: 167–171. [PubMed: 7370371]
44. Collins J and Albert PS. Estimating diagnostic accuracy without a gold standard: a continued controversy. *Journal of Biopharmaceutical Statistics* 2016; 26: 1078–1082. [PubMed: 27548004]
45. Ishwaran H and Gatsonis C. A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *Canadian Journal of Statistics* 2000; 28: 731–750.
46. Siu AL Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. *Annals of Internal Medicine* 2016; 164: 279–296. [PubMed: 26757170]

Figure 1.

Characteristics of the paired kappa over the range of β for (a) increasing correlation ρ_v between the expert random effects with parameters set at $\sigma_{v1}^2 = \sigma_{v2}^2 = \sigma_{u1}^2 = \sigma_{u2}^2 = 10$, $\rho_u = 0.5$, (b) increasing sets of variances where $\rho_u = \rho_v = 1$, (c) increasing variability $\sigma_{u1}^2 = \sigma_{u2}^2$ for each screening test with $\sigma_{v1}^2 = \sigma_{v2}^2 = 10$ and $\rho_u = \rho_v = 0.5$. Figure (d) presents the impact of increasing disease prevalence for three different sets of correlation coefficients $\rho_u = \rho_v = 0, 0.5$ and 0.9 . Variances are $\sigma_{u1}^2 = \sigma_{u2}^2 = 10$ and $\sigma_{v1}^2 = \sigma_{v2}^2 = 100$.

Expert and Patient Random Effects by Screening Test

Figure 2. Estimation of radiologist ($J=12$) and patient ($I=75$) random effects as conditional modes for the ACRIN Breast Imaging Study comparing film to digital mammography.

Table 1.

Parameter values for the simulation scenarios with 1000 datasets generated for each combination for binary ratings ($C = 2$).

Variances $(\sigma_{u1}^2, \sigma_{u2}^2, \sigma_{v1}^2, \sigma_{v2}^2)$	Correlations (ρ_u, ρ_v)	Random Effects Distributions	Number of items I , Number of raters J
(2.5, 2.5, 0.1, 0.1)	(0.95, 0.5)	$(u_{01i}, u_{02i}) \sim \text{MVN}(0, \Sigma_u)$ and $(v_{01j}, v_{02j}) \sim \text{MVN}(0, \Sigma_v)$	$(I = 75, J = 15)$
(10, 5, 5, 10)	(0.5, 0.75)	$(u_{01i}, u_{02i}) \sim \text{MVBeta}(a, b)$ and $(v_{01j}, v_{02j}) \sim \text{MVN}(0, \Sigma_v)$	$(I = 250, J = 100)$

Table 2.

True values and mean estimates (mean standard error) of the proposed paired kappa agreement measure κ_{paired} for each simulation study. Each set of simulations is based upon 1000 simulated datasets with binary ($C = 2$) outcomes.

(a) Normally distributed Random effects				
Variances ($\sigma_{u1}^2, \sigma_{u2}^2, \sigma_{v1}^2, \sigma_{v2}^2$)	Correlations (ρ_u, ρ_v)	True κ_{paired}	Mean $\hat{\kappa}_{mpaired}$ (S.E.)	Mean $\hat{\kappa}_{mpaired}$ (S.E.)
			$I = 75 \quad J = 15$	$I = 250 \quad J = 100$
(2.5, 2.5, 0.5, 0.5)	(0.95, 0.5)	0.726	0.722 (0.051)	0.724 (0.021)
	(0.5, 0.75)	0.632	0.630 (0.029)	0.631 (0.011)
(10, 5, 5, 10)	(0.95, 0.5)	0.721	0.721 (0.051)	0.721 (0.020)
	(0.5, 0.75)	0.686	0.681 (0.036)	0.687 (0.013)
(b) Non-normally distributed Random effects				
			$I = 75 \quad J = 15$	$I = 250 \quad J = 100$
(2.5, 2.5, 0.5, 0.5)	(0.95, 0.5)	0.726	0.737 (0.056)	0.732 (0.022)
	(0.5, 0.75)	0.632	0.638 (0.032)	0.633 (0.011)
(10, 5, 5, 10)	(0.95, 0.5)	0.721	0.721 (0.052)	0.720 (0.020)
	(0.5, 0.75)	0.686	0.672 (0.036)	0.684 (0.013)

Table 3.

Results for the DMIST Reader Study (Hendrick et al., 2008) comparing the binary classifications (1 = a negative test interpretation indicating no malignancy; 2 = a positive test interpretation indicating the presence of malignancy) of $J = 12$ radiologists for $I = 75$ patients undergoing both screen-film and digital mammography.

Parameter	Symbol	Estimate	S.E.
Binary GLMM:			
Intercept	α	1.258	0.280
Fixed coefficient	β	0.203	0.202
Patient Random Effect Variances:			
Screening Test #1 Film	σ_{u1}^2	2.312	0.376
Screening Test #2 Digital	σ_{u2}^2	1.857	0.302
Correlation coefficient	ρ_u	0.991	0.002
Rater Random Effect Variances:			
Screening Test #1 Film	σ_{v1}^2	0.443	0.181
Screening Test #2 Digital	σ_{v2}^2	0.414	0.170
Correlation coefficient	ρ_v	0.638	0.186
Agreement Measure:			
Paired Kappa	κ_{paired}	0.729	0.049

Table 4.

Goodness-of-fit assessment for the DMIST Breast Cancer Study with $J = 12$ experts and $I = 75$ patient test results for film and digital mammography. Observed counts from the study dataset and estimated counts based upon model-based probabilities estimated from the binary GLMM are displayed in brackets for (a) each separate screening test; (b) pairs of patients' two screening test results that fall into each paired category.

(a) Marginal Observed (Estimated) Counts			
Disease Category		Film Mammography	Digital Mammography
Non-diseased		672 (657.32)	653 (638.06)
Diseased		214 (228.68)	233 (247.94)
(b) Paired Observed (Estimated) Counts			
		Digital Mammography	
		Not diseased	Diseased
Film	Not diseased	588 (549.99)	65 (107.25)
Mammography	Diseased	84 (88.01)	149 (136.87)