Published in final edited form as:

Phys Med Biol. 2012 May 21; 57(10): 2873–2904. doi:10.1088/0031-9155/57/10/2873.

Measuring agreement between ratings interpretations and binary clinical interpretations of images: a simulation study of methods for quantifying the clinical relevance of an observer performance paradigm

Dev P. Chakraborty, Ph. D.

University of Pittsburgh Department of Radiology, 4771 Presbyterian South Tower 200 Lothrop St
Pittsburgh, PA 15213, USA 412-605-1553 (office) 412-605-1554 (fax) dpc10@pitt.edu

Abstract

Laboratory receiver operating characteristic (ROC) studies, that are often used to evaluate medical imaging systems, differ from “live” clinical interpretations in several respects which could compromise their clinical relevance. The aim was to develop methodology for quantifying the clinical relevance of a laboratory ROC study. A simulator was developed to generate ROC ratings data and binary clinical interpretations classified as correct or incorrect for a common set of images interpreted under clinical and laboratory conditions. The area under the trapezoidal ROC curve was used as the laboratory figure-of-merit and the fraction of correct clinical decisions as the clinical figure-of-merit. Conventional agreement measures (Pearson, Spearman, Kendall and kappa) between the bootstrap-induced fluctuations of the two figures-of-merit were estimated. A jackknife pseudo-value transformation applied to the figures-of-merit was also investigated as a way to capture agreement existing at the individual image level that could be lost at the figure-of-merit level. It is shown that the pseudo-values define a relevance ROC curve the area under which (rAUC) measures the ability of the laboratory figure-of-merit based pseudo-values to correctly classify incorrect vs. correct clinical interpretations, and is a measure of the clinical relevance of an ROC study. The conventional measures and rAUC were compared under varying simulator conditions. It was found that design details of the ROC study, namely the number of bins, the difficulty level of the images, the ratio of disease-present to disease-absent images, and the unavoidable difference between laboratory and clinical performance levels, can seriously underestimate the agreement as indicated by conventional agreement measures, even for perfectly correlated data, while rAUC showed high agreement and was relatively immune to these details. At the same time rAUC was sensitive to factors such as intrinsic correlation between the laboratory and clinical decision variables and differences in reporting thresholds that are expected to influence agreement both at the individual image level and at the figure-of-merit level. Suggestions are made for how to conduct relevance ROC studies aimed at assessing agreement between laboratory and clinical interpretations.

Keywords

observer performance; ROC; bootstrap; jackknife; agreement; correlation; agreement; statistical modeling; clinical relevance

1. Introduction

Receiver operating characteristic (ROC) studies are widely used in medical imaging to compare the performance of imaging systems (Metz 1978; Metz 1989; Dodd, Wagner et al. 2004; Wagner, Metz et al. 2007; Kundel, Berbaum et al. 2008). It is not often appreciated

that (i) ROC studies represent laboratory simulations of the clinical task, (ii) direct measurement of clinical performance is difficult, and (iii) it is not proven that laboratory measurements track clinical performance measurements in the sense of predicting the same ordering of modalities. ROC studies aim to simulate the basic aspects of the clinical task, namely having radiologists find disease on actual images under conditions of uncertainty, but the laboratory simulations cannot reproduce all factors affecting clinical interpretations. They are conducted using a ratings scale, usually with 5 or more ratings, while in the clinic the decisions are often binary. The ratings scale allows finer definition of the ROC curve and more accurate estimation of threshold-independent performance measures such as the area under the ROC curve. With binary decisions such estimation is not possible and one must rely on metrics such as sensitivity and specificity, which are reporting threshold dependent. The ROC image sets are usually enriched to contain a larger fraction of diseased patients than is normally encountered in the clinic. For example in the United States about 3 – 5 per thousand screened women actually have cancer, and conducting an ROC study with such low prevalence is out of the question. The low prevalence could induce a “vigilance” effect (Wolfe, Horowitz et al. 2005), where very infrequently occurring cancers can be missed, which is not simulated in laboratory studies. The ROC studies tend to be more subtle (ambiguous normals and subtle lesions) than encountered in the clinic. ROC studies are designed so that the same readers interpret the same set of cases in the different modalities, thereby allowing analysis that attenuates the effect of reader and case variability, leading to greater statistical power (Dorfman, Berbaum et al. 1992). This type of matching is of course impossible in the clinic. In the laboratory there is no “clinical pressure” since there are no consequences to correct or incorrect decisions. In the clinic the reader generally has access to more patient related information than in a laboratory study (e.g., the reason for the examination, images from other modalities, etc.). While these differences are there for a purpose, namely to yield a more controlled and precise measurement of the difference in areas under the laboratory ROC curves between two modalities, i.e., greater statistical power, the differences from “real life” could degrade the clinical relevance of the laboratory ROC study and predict incorrect modality ordering from that obtaining in the clinic.

The differences between the ratings ROC paradigm and binary clinical interpretations have stirred much interest and debate in the observer performance methodology development community. A “laboratory effect” has been identified wherein performance has been observed to be higher in the clinic than in a laboratory ROC study conducted with the same images and readers (Gur, Bandos et al. 2007; Gur, Bandos et al. 2008). It has been suggested that ROC studies conducted with binary ratings may have advantages over multiple ratings in certain scenarios (Gur, Bandos et al. 2010) but this view has been disputed (Samuelson, Gallas et al. 2011). The argument for binary ratings ROC studies is that the radiologist's operating point should be a primary consideration in designing an observer performance study, and while the ROC rating paradigm provides more detailed definition of the ROC curve, it does not reflect the actual operating point or reporting threshold adopted by the radiologist in the clinic. Alternatives to ROC measurements are being explored which involve collecting more information than a single ROC rating per image (Bunch, Hamilton et al. 1978; Swenson 1996; Obuchowski, Lieber et al. 2000; Chakraborty 2011) but there is controversy surrounding the clinical relevance of one of the paradigms (Gur and Rockette 2008; Chakraborty 2009).

We believe that much of the controversy and resulting confusion can be traced to the fact that there is currently no accepted definition of clinical relevance and no way of quantifying it. A hierarchical classification system of efficacies of different types of measurements (Fryback and Thornbury 1991; Thornbury 1994; NCRP 1995) is the closest we could find to a description of the true meaning of clinical relevance. Level 1 is technical efficacy, measured by spatial resolution, modulation transfer function, etc., the traditional physical

measures of image quality performed routinely in radiology departments. Level 2 is diagnostic accuracy efficacy, and includes measures such as area under the ROC curve, sensitivity and specificity, i.e., the traditional ROC metrics. Level 3 is diagnostic-thinking efficacy, which refers to the extent to which knowledge about the result of the diagnostic test changes the physician's estimate of the probability of disease. This is measured by positive and negative predictive values, as is done in clinical audits of mammography screening programs (Spring and Kimbrell-Wilmot 1987; Sickles, Ominsky et al. 1990). Level 4 is therapeutic efficacy, which refers to the impact of the results of the diagnostic test on the treatment of patients. Level 5 is patient-outcome efficacy, refers to whether or not the health of the patient is improved by the diagnostic test, and could be measured by long-term survival. Level 6 is societal efficacy, which refers to the value to society, addressing the issue of whether the cost borne by society is justifiable, measured by cost-benefit analysis. The ordered levels of the hierarchical classification system can be regarded as corresponding to increasing clinical relevance, i.e., clinical relevance (Level 3) > clinical relevance (Level 2) > clinical relevance (Level 1), etc., with the measurements become increasingly more difficult at the higher levels.

In this paper we consider a simulation model of “live” binary clinical interpretations of a set of cases by a reader and subsequent laboratory ROC ratings of the same cases by the same reader. The trapezoidal area under the ratings ROC curve is used as the laboratory performance measure (figure-of-merit). The clinical interpretations can be retrospectively classified as correct or incorrect, and the fraction of correct interpretations is used as a measure of clinical figure-of-merit. Because they share the same cases, the laboratory and clinical figures-of-merit are expected to be correlated. The simulator will incorporate all of the differences between laboratory and clinical interpretations described above with the exception of “clinical pressure” and “vigilance” effects. The simulated data yields two figures-of-merit, one laboratory and one clinical. Of course, one cannot calculate the agreement between just two numbers. Needed is a way to induce statistically valid fluctuations in the two numbers. This is done using the bootstrap resampling method (Efron 1982). The basic idea behind the bootstrap is quite simple: one assumes that the observed data defines the population. Therefore, one re-samples the data with replacement (by definition an infinite population cannot be exhausted). The resulting fluctuations in the two figures-of-merit allow estimation of their agreement.

What is the implication of high agreement between laboratory and clinical figures-of-merit? High agreement implies quite literally, that positive fluctuations of the laboratory figure-of-merit tend to be associated with positive fluctuations of the clinical figure-of-merit, and negative with negative. A higher-than-usual laboratory figure-of-merit implies an easier dataset. This work, and others like it, involves an often unstated assumption that the high agreement holds regardless of what caused the dataset to be easier to interpret. It follows that if modality A makes the cases easier to interpret in the laboratory than modality B, then modality A is also going to yield greater clinical performance than modality B. High agreement gives greater confidence that the modality ordering found in the laboratory study will be confirmed in the clinic. [This assumption is implicit in related work that has studied this type of issue (Gifford, King et al. 2000; Narayanan, Gifford et al. 2002), namely, one measures the agreement when certain imaging parameters are varied, and if the agreement is strong one assumes that the agreement also applies when modalities are changed.]

In the following sections we describe the simulation model. A few conventional methods of measuring agreement between the laboratory and clinical figures-of-merit are described. Because of intrinsic limitations of these measures in the present context, a new method, termed the relevance-ROC, is proposed. Unlike the conventional agreement measures, the relevance-ROC measures agreement at the image-level. It will be argued that the area under

the relevance ROC curve qualifies as a measure of clinical relevance of the laboratory ROC study. Results are presented under varying conditions of simulator parameters. Some statistical issues with the new analysis are discussed and suggestions are made on how future studies that aim to measure laboratory vs. clinical agreement should be conducted. [In the following “agreement” refers to the generic measure while “correlation” is used when there is an associated correlation coefficient.]

2. Methods

The simulator needs to generate decision variables under laboratory and clinical conditions. It needs to allow specification of the difficulty levels of the two sets of interpretations, the correlation between the ratings, the number of bins used in the laboratory interpretations, the offset between the central laboratory threshold and the clinical reporting threshold.

2.1. Modeling the ratings

Let $\mathcal{N}(\mu, \sigma^2)$ denote a random sample from a normal distribution with mean μ and variance σ^2 . The simulator, illustrated schematically in Fig. 1, assumes that the laboratory decision variable is obtained by sampling from one of two unit-variance normal distributions: an $\mathcal{N}(0,1)$ distribution for disease-free images and an $\mathcal{N}(\mu_{lab}, 1)$ distribution for the disease-present images, upper half of the figure. Likewise, the clinical decision variables are obtained by sampling from two unit variance normal distributions: an $\mathcal{N}(0,1)$ distribution for the disease-free images and an $\mathcal{N}(\mu_{clin}, 1)$ distribution for the disease-present images, see lower half of figure. Positive correlation between the laboratory and clinical decision variables is modeled by assigning a common random component to them. If the common component is large relative to other random term(s), then the two will tend to be highly correlated. This is the basic idea behind the variance component modeling of correlated ROC data (Roe and Metz 1997). We consider a simplified version of this model that applies to a perfectly reproducible reader interpreting a set of images in the clinical condition and subsequently interpreting the same set of images in the laboratory condition. The continuous decision variables, henceforth referred to as *z-samples*, are modeled by

$$z_{ikt} = \mu_t + \Delta\mu_{it} + C_{kt} + (\tau C)_{ikt}, \quad \text{Eqn. 1}$$

In Eqn. 1 $i = 1$ corresponds to the laboratory interpretations and $i = 2$ to the clinical interpretations. We adopt the usual convention of indexing the images with two concatenated integers kt , where N_N disease-free images ($t = 1$) are indexed by $k1$ with $k = 1, 2, \dots, N_N$ and N_A disease-present images ($t = 2$) are indexed by $k2$ with $k = 1, 2, \dots, N_A$. If we impose the constraints $\mu_1 \equiv 0$, $\Delta\mu_{11} = 0$, $\Delta\mu_{21} = 0$ and $\Delta\mu_{12} = 0$ then by definition $\mu_1 \equiv 0$ corresponds to the common means of the disease-free distributions for the laboratory and clinical interpretations, and μ_2 specifies the mean of the disease-present distribution for the *laboratory* interpretations, to emphasize which we define $\mu_{lab} \equiv \mu_2$. The constraints are equivalent to a (positive) right-shift on (only) the diseased distribution for the clinical reader, as depicted in the figure. It follows that $\mu_{clin} \equiv \mu_2 + \Delta\mu_{22} = \mu_{lab} + \Delta\mu_{22}$. The term $\Delta\mu_{22}$ determines the difference in performance between the laboratory and the clinical reader. Since the clinical reader generally has more information (clinical history, additional views, images from other modalities, etc.) and greater motivation, one expects greater performance for the clinical reader than for the laboratory reader.

The last two quantities in Eqn. 1 represent independent random samples defined by

$$\begin{aligned} C_{kt} &\sim N(0, \sigma_c^2) \\ (\tau C)_{ikt} &\sim N(0, \sigma_{\tau c}^2). \end{aligned} \quad \text{Eqn. 2}$$

In Eqn. 2, the tilde symbols denote random samples from the specified distributions. Since the total image variance is unity,

$$\sigma_c^2 + \sigma_{\tau c}^2 = 1. \quad \text{Eqn. 3}$$

Eqn. 3 implies that the correlation $\rho_{lab,clin}^z$ between the z-samples corresponding to the laboratory ratings and the clinical ratings is

$$\rho_{lab,clin}^z = 1 - \frac{Var(z_{1kt} - z_{2kt})}{2Var(z_{ikt})} = 1 - \frac{2\sigma_{\tau c}^2}{2(\sigma_c^2 + \sigma_{\tau c}^2)} = \sigma_c^2 / (\sigma_c^2 + \sigma_{\tau c}^2) = \sigma_c^2. \quad \text{Eqn. 4}$$

The correlation arises from the fact that the same cases are being interpreted under the two conditions. The superscript z is used to emphasize that this is a ratings-level correlation. For example, if $\sigma_c^2 = 1$, which implies $\sigma_{\tau c}^2 = 0$, the treatment-dependent contribution to the z-sample is zero, and the z-samples for the laboratory and clinical interpretations are perfectly correlated and $\rho_{lab,clin}^z = 1$, and conversely if $\sigma_{\tau c}^2 = 1$, the variables become uncorrelated and $\rho_{lab,clin}^z = 0$.

2.2. Converting the z-samples to integer ratings

The simulator produces continuous z-samples. In practice, observer studies are conducted using a finite number of ratings. The laboratory z-samples z_{1kt} were binned using an *odd* number R ($R = 1, 3, 5, \dots$) of equally spaced cutoffs ζ_r in the closed interval $[-2.2, \mu_{clin} + 2.2]$ such that central cutoff was always at $\mu_{clin}/2$. If the z-sample was below ζ_1 the image was rated 1, if the z-sample was between ζ_1 and ζ_2 the image was rated 2, etc., and if the z-sample exceeded ζ_R the image was rated $R+1$. The integer rating is denoted by r_{kt} which takes on values 1, 2, ..., $R+1$ (the i -subscript is unnecessary since r_{kt} is implicitly a laboratory rating). If $R = 1$, then there is a single cutoff at $\mu_{clin}/2$ and the binning is binary. In the schematic shown in Fig. 1, $R = 5$ and $\zeta_1 = -2.2$, $\zeta_3 = \mu_{clin}/2$ and $\zeta_5 = \mu_{clin}/2 + 2.2$. [Excluding the case $R = 1$, by design few counts are expected in bins 1 and $R+1$, so the effective number of bins is actually $R-1$ where $R = 3, 5, 7$, etc. An odd number of cutoffs, corresponding to an even number of bins, allowed unambiguous selection of a central cutoff. For example, in a 6-rating ROC study, ratings 1,2 and 3 can be cumulated into one bin, and ratings 4,5 and 6 into another bin, thus defining a central cutoff ζ_3 . This allowed us to separate the effects of the number of bins and cutoff mismatch between laboratory and clinical conditions. The latter is described by an offset parameter to be described shortly.]

2.3. Simulating the correct / incorrect clinical data

To generate binary clinical data the clinical cutoff was set at $\zeta_{clin} = \mu_{clin}/2 + \zeta_{offset}$. A disease-free image ($t = 1$) was classified as correct ($c_{k1} = 1$) if $z_{2k1} < \zeta_{clin}$ and otherwise the image was classified as incorrect ($c_{k1} = 0$). A disease-present image ($t = 2$) was classified as correct ($c_{k2} = 1$) if $z_{2k2} > \zeta_{clin}$ and otherwise the image was classified as incorrect ($c_{k2} = 0$). The quantity ζ_{offset} represents an offset between the central cutoff $\mu_{clin}/2$ used to generate the binned laboratory data and that used to generate the binary clinical data, the issue raised by (Gur, Bandos et al. 2010). If $\zeta_{offset} > 0$ then the clinical reader is more conservative at calling an image as disease-present, resulting in fewer true positives ($c_{k2} = 1$) and false positives ($c_{k1} = 0$). In Fig. 1 the offset parameter ζ_{offset} is negative.

2.4. Laboratory figure-of-merit

The area A under the trapezoidal ROC curve, the laboratory figure-of-merit, was calculated using

$$A = \frac{1}{N_N N_A} \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}). \quad \text{Eqn. 5}$$

Here $r_{k_1 1}$ is the binned laboratory rating of disease-free image $k_1 1$, $r_{k_2 2}$ is the binned laboratory rating of disease-present image $k_2 2$, and the function ψ is defined as unity if the second argument exceeds the first, 0.5 if the two are equal, and 0 otherwise. For the simulator model described by Eqn. 1, it can be shown that the asymptotic value of A (observable with a large number of images and no binning) is given by (Burgess 1995)

$$A = \Phi\left(\frac{\mu_{lab}}{\sqrt{2}}\right).$$

2.5. Clinical figure-of-merit

The average of the classification vector c_{ktb} see §2.3, was defined as the clinical figure-of-merit C :

$$C = \frac{1}{N_N + N_A} \left[\sum_{k=1}^{N_N} \psi(z_{2k1}, \zeta_{clin}) + \sum_{k=1}^{N_A} \psi(\zeta_{clin}, z_{2k2}) \right]. \quad \text{Eqn. 6}$$

The first summation is over the true negative decisions and the second summation is over the true positive decisions. One divides by the total number of cases to get the fraction of correct decisions. [The probability is zero that the (continuous variable) z-sample will exactly equal the clinical threshold, so the ψ functions in Eqn. 6 always yield 0 or 1.]

2.6. Measuring the agreement between fluctuations in laboratory and clinical figures-of-merit

Each simulated dataset yielded a single value for the laboratory figure-of-merit (A) and a single value for the clinical figure-of-merit (C). To measure agreement between the laboratory and clinical figures-of-merit one needs to induce statistically valid fluctuations in these quantities. This can be accomplished using the bootstrap technique (Efron 1982), which consists of regarding the dataset as defining the population, and to sample with replacement from it (by definition a population cannot be exhausted, therefore one must sample with replacement). The fluctuations associated with these samples allow estimation of agreement measures. For example, if positive fluctuations of the laboratory figure-of-merit tend to be associated with positive fluctuations of the clinical figure-of-merit, then agreement is high.

For each simulation condition, 50 independent simulated datasets, indexed by the simulation index, $s = 1, 2, \dots, S = 50$, were generated. Each s yields a laboratory figure-of-merit A_s and a clinical figure-of-merit C_s . For each s two hundred bootstrap samples were generated, indexed by b ($b = 1, 2, \dots, B = 200$). For the b^{th} bootstrapped sample from the s^{th} simulated dataset, the laboratory figure-of-merit, A_{sb} and the clinical figure-of-merit C_{sb} , were calculated (the notation reflects the fact that these depend on both simulation and bootstrap indices).

At this point one can calculate various standard agreement measures (e.g., Pearson, Spearman and Kendall correlation coefficients) between the B-dimensional arrays and A_{sb} and C_{sb} ($b = 1, 2, \dots, B$). The kappa measure of agreement (Cohen 1960) can also be calculated provided the fluctuations are categorized into two (or more) levels. In this study the differences from the corresponding non-bootstrapped values were replaced with 0s and 1s according to the signs of the differences from the corresponding simulation values. The measures were averaged over the S simulated datasets.

2.7. Measuring the agreement between fluctuations in laboratory and clinical figures-of-merit at the image level

The method just described is the standard statistical approach to measuring agreement between fluctuations of the laboratory and clinical figures-of-merit. The figures-of-merit represent averages over all images. Such averaging could obscure more detailed agreement observable at the image level. For example, consider an easy disease-present image where the disease is highly visible. In a 6-rating laboratory ROC study this would elicit a rating of “6” (high confidence disease-present) and moreover the clinical interpretation is likely to be correct ($c_{kt} = 1$), but this agreement is not readily apparent from examination of the two figures-of-merit. If one could devise an image-level measure of the *correctness* of the laboratory decisions, denoted by χ_{kt} , then a more sensitive measure of agreement might be possible. A correctness measure can be defined using the concept of jackknife pseudovalues. Corresponding to the laboratory figure-of-merit A the jackknife pseudovalue A_{kt}^* for image kt (since A is the laboratory figure-of-merit the i -subscript is redundant and therefore it is suppressed; implicitly $i=1$) is defined by (Efron 1982)

$$A_{kt}^* = KA - (K - 1)A_{(kt)}. \quad \text{Eqn. 7}$$

Here $K = N_N + N_A$ is the total number of images and $A_{(kt)}$ is the laboratory figure-of-merit with image kt removed from the analysis. Examination of this equation yields the following insight: removal of an easy image, which could be disease-present or disease-free, leaves behind a slightly more difficult image set, which means that $A_{(kt)}$ will decrease, perhaps ever so slightly, relative to the baseline value A obtained when all images are included.

Since it appears with a negative sign, the jackknife pseudovalue A_{kt}^* will increase, perhaps ever so slightly. Likewise, removal of a difficult image leaves behind an easier image set, and the pseudovalue A_{kt}^* will decrease. The pseudovalue fluctuations mirror the correctness of the laboratory decisions on the individual images. The laboratory correctness measure χ_{k1} and χ_{k2} for disease-free and disease-present images, respectively, are tentatively defined by

$$\begin{aligned} \chi_{k1} &= A_{k1}^* - \frac{1}{N_N} \sum_{k=1}^{N_N} A_{k1}^* \\ \chi_{k2} &= A_{k2}^* - \frac{1}{N_A} \sum_{k=1}^{N_A} A_{k2}^*. \end{aligned} \quad \text{Eqn. 8}$$

The quantity χ_{kt} is expected to be positive for images on which high confidence (ones and sixes) correct laboratory decisions were made, i.e., high confidence true positives and true negatives, and negative for images on which high confidence incorrect laboratory decisions were made, i.e., high confidence false positives and false negatives, and near zero for images on which the decisions were neutral (three rating). [It helps to keep in mind that the extremes of the ROC rating scale, 1 and 6 in the example, represent high confidence decisions, for absence and presence of disease, respectively.]

In Appendix A it is shown that the separate averages of the pseudovalues, over disease-free and disease-present cases, each individually equals the trapezoidal area A defined in Eqn. 5. In other words,

$$A = \frac{1}{N_N} \sum_{k=1}^{N_N} A_{k1}^* = \frac{1}{N_A} \sum_{k=1}^{N_A} A_{k2}^*. \quad \text{Eqn. 9}$$

Eqn. 9 is an extension of a similar result (Hillis 2007) that showed that

$$\frac{1}{N_N + N_A} \left[\sum_{k=1}^{N_N} A_{k1}^* + \sum_{k=1}^{N_A} A_{k2}^* \right] = A. \quad \text{Eqn. 10}$$

In other words, the earlier result showed that the average of the pseudovalues over *all cases* equals the figure-of-merit A . Our result implies the earlier result (as can be verified by substitution), but the converse is not true.

Eqn. 9 is important as it gives a physical interpretation to the pseudovalues, namely the pseudovalues can be regarded as the individual image contributions to the area under the trapezoidal ROC. Some disease-free images (the easy ones) contribute more to the figure-of-merit than the hard ones, but the average over all disease-free images of A_{k1}^* is the net figure-of-merit over *all* images, and likewise the average over all disease-present images of A_{k2}^* is the net figure-of-merit over all images. Using Eqn. 9, Eqn. 8 reduces to

$$\begin{aligned} \chi_{k1} &= A_{k1}^* - A \\ \chi_{k2} &= A_{k2}^* - A. \end{aligned} \quad \text{Eqn. 11}$$

It is also shown in Appendix A that χ_{k1} is proportional to the probability that the average disease-present image rating exceeds the rating of *specific* disease-free image k_1 minus the probability that the average disease-present image rating exceeds the rating of an *average* disease-free image. In other words, χ_{k1} is a measure of how far from average a specific disease-free image is. Likewise χ_{k2} is a measure of how from average a specific disease-present image is. It is also shown in Appendix A, that by a suitable rescaling of the correctness measure, the two constants of proportionality become unity. The redefinition of the correctness-measures, that overrides Eqn. 8 and Eqn. 11, is

$$\begin{aligned} \chi_{k1} &= \frac{(N_N - 1)}{(K - 1)} (A_{k1}^* - A) \\ \chi_{k2} &= \frac{(N_A - 1)}{(K - 1)} (A_{k2}^* - A). \end{aligned} \quad \text{Eqn. 12}$$

The pseudovalue transformation was also applied to the clinical figure-of-merit C . It is shown in Appendix B that the k^{th} pseudovalue is identical to c_{kt} . In other words,

$$C_{kt}^* = KC - (K - 1) C_{(kt)} = c_{kt}. \quad \text{Eqn. 13}$$

Now χ_{kt} is a K -dimensional array of real numbers and one seeks to determine how well this array correlates with the K -dimensional array of binary (0,1) integers that make up the classification vector. The Mann-Whitney-Wilcoxon U-statistic (Wilcoxon 1945; Mann and Whitney 1947) is a non-parametric measure of the association between the continuous χ_{kt} values and the binary c_{kt} values. It is close to unity if large values of χ_{kt} are predominantly associated with $c_{kt} = 1$ and small values with $c_{kt} = 0$. It is defined by

$$U = \frac{1}{N_0 N_1} \sum_{r=1}^{N_0} \sum_{s=1}^{N_1} \psi(\chi_r, \chi_s). \quad \text{Eqn. 14}$$

Here N_0 is the number of zeroes in the c_{kt} array and N_1 is the number of ones in the c_{kt} array, r is the image index (i.e., N_0 values in the range 1 to K , but not necessarily consecutive values) of the r^{th} zero-element in the c_{kt} array with correctness measure χ_r , s is the image index (i.e., N_1 values also in the range 1 to K , not necessarily consecutive) of the s^{th} one-element in the c_{kt} array with correctness measure χ_s . The U measure was calculated once for each independent simulated data set and the values were average over $S = 50$ simulations. [Note that χ_s does not have a truth subscript: both $t = 1$ and $t = 2$ images can contribute to incorrect decision images; in fact any difficult image, regardless of its truth status, can be in the set of images indexed by r . Likewise χ_r does not have a truth subscript because any easy image, regardless of its truth status, can be in the set of images indexed by s .]

2.8. Physical interpretation of the U-statistic

In Eqn. 14, U can be interpreted as the area under the trapezoidal “ROC” curve defined by the correctness measures χ_r, χ_s interpreted as the “ratings” of N_0 “disease-free” and N_1 “disease-present” images that measures the classification accuracy of the correctness measures in separating the incorrect and correct clinical decision images. We use quotation marks as this is not a conventional ROC; rather we term it a relevance-ROC (rROC). Unlike a conventional ROC, the area under which is the probability that the rating of a disease-present image is greater than that of a disease-absent image, the area under the relevance ROC is the probability that the correctness measure for a correct clinical decision ($c_{kt} = 1$) image is greater than the correctness measure for an incorrect clinical decision ($c_{kt} = 0$) image. Henceforth the U-statistic defined in Eqn. 14 will be referred to as the (trapezoidal) area under the relevance ROC curve (rAUC). It is a measure of how good the laboratory correctness measure is at correctly classifying correct and incorrect clinical interpretation images. This suggests that rAUC is a measure of laboratory-clinical agreement or the clinical relevance of the laboratory paradigm. The x-axis of the relevance ROC is relevance-FPF, defined as an incorrect clinical interpretation with laboratory correctness measure exceeding a threshold. The y-axis is relevance-TPF, defined as a correct clinical interpretation with laboratory correctness measure exceeding a threshold. Table 1 compares the conventional ROC to the relevance ROC.

2.9. Comparing different methods of measuring agreement

We have described different methods for measuring agreement between the laboratory and clinical interpretations. How does one choose between them? This is done by examining their behavior when the simulator parameters are changed. The general rule is that that method is preferred that more faithfully reflects agreement at the z-sample level. Consider the case where two z-samples are perfectly correlated. If the laboratory and clinical readers also have identical average performances, i.e., $\mu_{lab} = \mu_{clin}$, a measure that reports less than perfect agreement, depending furthermore on the number of bins chosen for the laboratory study and the common difficulty level of the selected cases, is less desirable than a method that reports perfect agreement, independent of these design-details of the ROC study. Two investigators working with the same sets of images might reasonably arrive at different choices for the number of bins (some investigators prefer few ratings, some prefer one hundred bins) and common difficulty level of the cases (some investigators prefer more difficult cases than others). Since the laboratory reader has less information than the clinical reader one expects $\mu_{lab} < \mu_{clin}$. An agreement measure that depends strongly on the performance difference would be undesirable, especially if it did not reflect perfect

agreement existing at the z -sample level. Likewise, an agreement measures that depends strongly on disease-prevalence (fraction of cases that are disease-present) is undesirable, as this is also a design detail of the study and an unavoidable complication, given that most clinical interpretations occur under very low disease-prevalence conditions compared to laboratory ROC studies. On the other hand the agreement measures should depend on the intrinsic correlation at the z -sample level, decreasing to chance level agreement as the correlation decreases to zero. Likewise, agreement should depend on the offset parameter and there is no guarantee that the laboratory observer will adopt the same reporting central threshold as the clinical observer. In fact the agreement measures are expected to become smaller as the magnitude of ζ_{offset} increases. This is because although the laboratory figure-of-merit is unaffected by the non-zero offset the clinical figure-of-merit is affected, see Fig. 1: making the offset more negative than shown will result in more false positives.

3. RESULTS

3.1. Example plots of fluctuations in A vs. fluctuations in C and the corresponding relevance ROC curves

Before showing detailed results a few representative scatter plots of $\Delta A \equiv A_{sb} - A_s$ vs. $\Delta C \equiv C_{sb} - C_s$ and corresponding trapezoidal relevance-ROC curves, are shown in Fig. 2 panels (a-d). See §2.6 for definitions of these quantities and recall that b is the bootstrap index and s is the simulation index. The plots are for $s = 1$ and b varying from 1 to 200, so ΔA and ΔC are each 200-dimensional arrays. For each simulation, arrays such as these were used to calculate the conventional correlation coefficients and kappa.

Panel (a) is a scatter plot of ΔA vs. ΔC for 50 disease-free and 50 disease-present images. Other simulation conditions were $\rho_{lab,clin}^z = 1$ (perfectly correlated pre-binned z -sample data), $\mu_{common} \equiv \mu_{lab} = \mu_{clin}$ (equal laboratory and clinical performances), $\zeta_{offset} = 0$ (central laboratory threshold and clinical reporting threshold were identical) and the laboratory data was binned into 4 bins ($R = 3$). There are 200 data points on this plot and the dependence is very close to linear ($r^2 = 0.995$). The high correlation is explained by the fact that the 4 bin case is effectively close to 2 bins, since the extreme bins have very few counts, so the laboratory and clinical ratings are almost identical. When we used 2 bins ($R = 1$) the scatter-plot yielded perfect correlation (not shown). Note that panel (a) is for a particular simulation. Repeating the simulation (different s) would have yielded slightly different values. Standard deviations for all estimates are reported later.

Panel (b) is the corresponding relevance-ROC which yielded $rAUC = 1.000$. With 4 bins and only 100 images the number of distinct values of A , and hence pseudovalues and laboratory correctness measures that are possible with the bootstrapped datasets, is quite limited, which explains the few points on the relevance-ROC. Panel (c) shows a plot of ΔA vs. ΔC when the data was not binned and all other simulation conditions remained the same. The correlation decreased to $r^2 = 0.755$ but the relevance-ROC shown in panel (d) remained close to perfect with $rAUC = 0.999$. Because the data was not binned there are many more points on the relevance-ROC.

3.2. Variation with number of bins and common μ

Unless noted otherwise all simulations in this work were conducted with 500 disease-free images and 500 disease-present images. This was done to reduce sampling variability to allow trends to be seen more clearly. With $\rho_{lab,clin}^z = 1$, $\Delta\mu_{22} = 0$ and $\zeta_{offset} = 0$ the pre-binned laboratory and the and clinical z -samples are identical, Eqn. 1. Fig. 3 panel (a) is a plot of Pearson correlation coefficient vs. $\mu_{common} \equiv \mu_{lab} = \mu_{clin}$ for different number of bins as indicated by the numeric symbols (“inf” = infinite number of bins, i.e., no binning was

performed). The seven values of μ_{common} shown correspond to, in increasing order, areas under the ROC curves of 0.60, 0.65, 0.70, 0.80, 0.85, 0.90 and 0.95, respectively. It is observed that (1) even though the pre-binned laboratory and the clinical z-samples are identical, the Pearson correlation is generally less than unity, (2) fine binning decreases correlation relative to coarse binning and (3) the decrease in correlation between a finite number of bins and an infinite number of bins is greater at large μ_{common} than at small μ_{common} . Fig. 3 panel (b) is the corresponding plot for rAUC. The numeric symbols line up one top of each other and the rAUC measure is almost constant at unity, independent of μ_{common} and number of bins.

Fig. 3 is important insofar as it shows the strong effect of ROC study design details, namely the difficulty level of the images and the number of bins, on the Pearson correlation (this effect was also observed for Spearman, Kendall and kappa, see below) but that the rAUC measure was almost completely unaffected, and remained close to the expected value of unity based on the perfectly correlated pre-binned data. This figure can be understood, intuitively, by noting that binning, especially into coarse bins, makes the laboratory data “look more like” the binary clinical data, and therefore the correlation increases. This explains the trend in Fig. 3 (a) for the coarser bins (smaller numbers) to be associated with larger Pearson correlation values. To understand the greater detrimental effect on correlation of fine binning for large μ_{common} vs. small μ_{common} , and bearing in mind that for the binary clinical data most of the fluctuations are within-bin, note that within-bin fluctuations of the z-samples contribute towards correlated variations of the two figures-of-merit, but between-bin fluctuations contribute towards uncorrelated variations. The variability of μ_{common} , and hence of the z-samples, depends on the number of images and the value of μ_{common} . Using (i) an expression derived in (Burgess 1995) for the variance of μ_{common} in terms of the variance of A and (ii) an expression quoted in (Hanley and McNeil 1982) for the variance of A in terms of A , the numbers of normal and abnormal cases, it can be shown that the standard deviation of μ_{common} increases as μ_{common} increases, and the increase is particularly rapid for large μ_{common} . The large standard deviation implies an increased amount of between-bin fluctuations that reduces correlation with clinical figure-of-merit, and this occurs especially at large μ_{common} . [Not shown in Fig. 3 are data for two bins where the correlation was unity regardless of μ because the binned z-samples for the laboratory and clinical data become identical. Also not shown are data for 40 bins which are essentially identical, but slightly above, those shown for infinite number of bins.]

3.3. Simulation results for other agreement measures for perfectly correlated data and zero offset

The other agreement measures (Spearman, Kendall and kappa) tracked the Pearson correlation but tended to be ordered as Pearson > Spearman > Kendall > kappa. This is expected because of the increasingly non-parametric nature of the other measures. For the simulations shown in Fig. 3, Table 2 lists the maximum, minimum, mean and coefficient of variation (CV = standard deviation divided by the mean) of the common separations parameter μ_{common} , the clinical figures-of-merit C , the Pearson, Spearman, Kendall and kappa, and the rAUC agreement measure. Unlike the conventional agreement measures, the rAUC agreement measure was relatively insensitive to the number of bins and μ_{common} and was very close to unity (maximum = 1, minimum = 0.999, CV = 0.000) consistent with the fact that the laboratory and clinical pre-binning data in this example were identical.

3.4. Variation with offset parameter ζ_{offset}

The results shown in the preceding sections were for $\zeta_{offset} = 0$ (central laboratory threshold and clinical reporting threshold were identical). In reality one has little control over the two thresholds. Fig. 4 shows the Pearson correlation coefficient (the filled circles) and the rAUC

agreement measure (the asterisks) plotted vs. ζ_{offset} for different combinations of μ_{common} and R. Panel (a) is for relatively large $\mu_{common} = 2.33$ (equivalent to $A = 0.95$) and two bins ($R = 1$) while panel (b) is for infinite number of bins ($R = \text{infinity}$). Panels (c) and (d) are corresponding results for $\mu_{common} = 0.358$ (equivalent to $A = 0.60$). In Fig. 4(a) Pearson correlation has a maximum of unity at zero offset, because this is the two bin case when the clinical and laboratory ratings become identical. On the other hand in Fig. 4 (b) Pearson correlation reaches a maximum of about 0.8 at zero offset; the reduced value of the maximum is due to effect of binning discussed in connection with Fig. 3. Similar comments apply to Figs. 4 (c) and (d). In each panel the agreement measures are seen to decrease, as expected, as the magnitude of ζ_{offset} increases, see §2.9. These panels demonstrate examples of where there is a genuine effect and a flat dependence is undesirable. The parameter ζ_{offset} reflects the difference in average cutoffs used by the two readers. Although the z-samples are identical, any difference in the offset parameter implies different correctness relationships between laboratory and clinical interpretations, which should degrade agreement and *both* rAUC and Pearson correlation show this dependence. This is the effect postulated by Gur et al (Gur, Bandos et al. 2010) for preferring a binary laboratory study. Table 3 lists summary statistics for the data shown in Fig. 4. The Pearson correlation measure varied between 1 (perfect agreement) and 0.341 (chance level agreement would be 0) while the rAUC agreement measure also varied between perfect agreement and 0.592 (chance level agreement would be 0.5).

3.5. Variation with intrinsic correlation $\rho_{lab,clin}^z$

So far results were presented for perfect z-sample correlation, i.e., $\sigma_{\tau_c}^2 = 0$ in Eqn. 4. In reality the laboratory reader may not be the same as the clinical reader, in which case one has to include the effect of the different responses of the two readers to each case (inter-reader variability). This can be accommodated by increasing the variance term $\sigma_{\tau_c}^2$ in Eqn. 1, which results in $\rho_{lab,clin}^z < 1$. Likewise, the $\sigma_{\tau_c}^2$ term can also model intra-reader variability where the same reader interprets in both conditions, but the reader is not perfectly consistent. Fig. 5 shows the variation of the Pearson correlation coefficient (filled circles) and the rAUC agreement measure (asterisks) vs. $\rho_{lab,clin}^z$ for $R=5$, $\zeta_{offset} = 0$ and $\mu_{common} = 1.81$ (equivalent to $A = 0.9$). As $\rho_{lab,clin}^z$ increased, all agreement measures increased, as expected, because the laboratory and clinical z-samples became increasingly correlated (of all the examples considered, this one is perhaps the easiest to understand). Table 4 shows summary statistics for the various measures. The maximum value of kappa was 0.783 while that for the rAUC agreement measure was unity. The corresponding minimum values were -0.021 and 0.500, respectively. As with the variation with ζ_{offset} the agreement measures are expected to increase with increasing correlation, and all measures had the desired behavior. This is another example of when a flat dependence is undesirable.

3.6. Variation with μ_{lab} for fixed μ_{clin}

So far results were presented for identical performance of the laboratory and the clinical readers. Because the laboratory reader generally has less information and motivation than the clinical reader, performance is expected to be smaller for the laboratory reader, so some difference between μ_{lab} and μ_{clin} is unavoidable and it is important, if possible, to design the study so that the agreement measures is less sensitive to it. Fig. 6 shows the variation of Pearson correlation coefficient (filled circles) and the rAUC agreement measure (asterisks) vs. μ_{lab} for fixed clinical performance $\mu_{clin} = 1.81$ (equivalent to $A = 0.90$) indicated by the vertical line; the other simulation parameters were $\rho_{lab,clin}^z = 1$ and $\zeta_{offset} = 0$. Panel (a) is for $R = 1$, panel (b) for $R = 3$, panel (c) for $R = 5$, panel (d) for $R = 7$ and panel (e) is for $R = \text{infinity}$. For finite number of cutoffs both measures peak near μ_{clin} and fall off away from

this value, but the fall off is more pronounced for the Pearson measure. For $R = 1$ and $R = 3$ the peaks are at unity, as both correspond (essentially) to binary binning, but for higher values of R the peak of the Pearson measure falls increasingly below unity, while that for the rAUC measure stays at unity – this is the behavior discussed in §3.2. For $R = 5$ both measures display secondary peaks at $\mu_{lab} = 0.2$ and 3.4 . We believe this occurs because for these values the cutoffs “bisect” the laboratory z-sample distributions exactly as the clinical cutoff literally bisects the clinical z-sample distributions ($\zeta_{clin} = \mu_{clin}/2$). For this to occur, the cutoff spacing must equal $\mu_{lab}/2$. For $R = 5$ the spacing is 1.6 , so subtracting this value from 1.8 yields 0.2 , the location of the secondary peak to the left of the main peak, and adding the spacing to 1.8 yields 3.4 , the location of the other peak. For $R = 7$ the spacing is 1.0 , so the secondary peaks occur at 0.8 and 2.8 , as is observed in panel (d). Pearson correlation is more sensitive to the difference between μ_{lab} and μ_{clin} , but for infinite number of bins rAUC is almost independent of it. For $R = 7$ the rAUC value is near unity, so this suggests that 8 bins (6 bins in practice because the two extreme bins have few counts) may be a good compromise. Table 5 list the summary statistics for the various agreement measures for the simulations shown in Fig. 6. The rAUC measure is nearly constant, with minimum value 0.877 and coefficient of variation $= 0.036$. Corresponding values for Pearson correlation were 0.500 and 0.190 , respectively. [The reason why the peak of the Pearson measure occurs below μ_{clin} for $R = \text{infinity}$ is not presently clear.]

3.7. Variation with disease prevalence

Disease prevalence D is defined by the ratio of the diseased to the total number of cases. So far results were presented for equal numbers of disease-free and disease-present images, $N_N = N_A = 500$, i.e., $D = 50\%$. In practice clinical interpretations are performed under low prevalence conditions, and one can ask what will happen to the agreement measures if laboratory studies are also performed under such conditions. In other words these simulations investigated the effect of actually performing a ROC study under the identical prevalence conditions existing in the clinic (which is totally impractical in real life). Fig. 7 (a) shows for $R = 1$ the variation of Pearson correlation coefficient (filled circles) and the rAUC agreement measure (asterisks) for varying prevalence in the range 0.01 to 0.8 , in logarithmic steps, for fixed total number of images $N_N + N_A = 2000$. Other parameters were $\rho_{lab,clin}^z = 1$, $\zeta_{offset} = 0$ and $\mu_{common} = 1.81$ (equivalent to $A = 0.9$). Fig. 7 (b) is the corresponding plot for infinite number of bins. The Pearson correlation peaks at $D = 0.5$ but rAUC is flat. In Fig. 7(a) the peak occurs at unity, because with 2 bins the data becomes perfectly correlated, while in Fig. 7(b) the peak is at a smaller value, which is the binning effect, investigated earlier. Table 6 lists summary statistics for the two figures-of-merit and all agreement measures for the simulations shown in Fig. 7. The conventional agreement measures were sensitive to this design detail of the ROC study (coefficient of variation $= 0.435$ for the Pearson correlation), but the rAUC measure was almost completely insensitive to it (coefficient of variation $= 0.002$).

The strong dependence of the Pearson correlation measure on disease prevalence can be understood by considering the simpler case of binary ratings. It is shown in Appendix C that for $D = 0.5$, perfectly correlated laboratory and clinical z-samples, binary laboratory ratings, zero offset and $\mu_{common} \equiv \mu_{lab} = \mu_{clin}$, that the trapezoidal area A defined in Eqn. 5 equals the clinical figure-of-merit C defined in Eqn. 6. This explains the peak at $r = 1$ in Fig. 7(a) of the Pearson correlation coefficient when the prevalence is 0.5 (note that the fitted line does not pass through this point). For D not equal to 0.5 , this equality no longer holds and the correlation falls. Disease prevalence has a strong effect on clinical performance but almost no effect (due to the normalizations) on laboratory performance.

4. DISCUSSION

This work was aimed at examining various statistical methods for measuring the agreement between clinical and laboratory interpretations. The conventional agreement measures considered in this work are generally considered acceptable by statisticians that we have consulted. However, as we have shown, these could grossly underestimate the agreement, even when the pre-binned decision variables are known to be perfectly correlated. The reason is that they measure agreement at the ROC-area level, not at the individual image level. The ROC-area represents an average over all images, and such averaging obscures agreement existing at the individual image level. Moreover, the agreement decreases with increasing number of bins and it depends on the value of the common performance level (Fig. 3), it decreases with increasing difference in performance levels (Fig. 6), and it decrease with increasing departure from 50% disease-prevalence in the case sample (Fig. 7). We have shown that the difference of the pseudovalue for an image from the average pseudovalue, appropriately scaled as in Eqn. 12, can be regarded as the correctness measure for the laboratory interpretation of the image. The correctness measure for an image determines, in probabilistic terms, how far from “average” that image is. The concept of the relevance ROC curve was introduced and rAUC was defined as the trapezoidal area under the relevance ROC curve. The rAUC agreement measure is proposed as a measure of laboratory – clinical agreement that is expected to be much more sensitive to detecting agreement at the individual image level than conventional agreement measures. It is shown that the rAUC is a better measure of agreement in so far it more faithfully reports any agreement existing at the z -sample level.

The method described in this paper has been applied in a clinical study comparing the clinical relevance of ROC and the free-response ROC (FROC) paradigms (Chakraborty and Berbaum 2004; Chakraborty, Hagood et al. 2012). The original prospective interpretations of 80 digital chest radiographs were classified by the truth panel as correct or incorrect depending on correlation with additional imaging. FROC mark-rating data was acquired for 21 radiologists and ROC data was inferred using the highest ratings. The areas under the ROC and alternative FROC curves were used as laboratory figures-of-merit. With an FROC study one uses the corresponding figure-of-merit instead of Eqn. 5, but otherwise the agreement analysis remains the same. Also, with clinical data one cannot simulate fresh datasets and must resort to bootstrapping to obtain confidence intervals for the agreement measures. Low conventional correlation measures and near chance level rAUC values attributable to large differences between the clinical and laboratory interpretation methods (for example, the clinical readers had access to CT images), were observed. The rAUC agreement measure was consistent with the traditional agreement measures but was more sensitive to the differences in clinical relevance (smaller confidence intervals). The clinical paper of course involved only one dataset, and to validate the method one needs many simulations under known conditions, such as was performed in the current work.

Possibly because of the warning in a well-known textbook (Efron 1982) that “although pseudovalues are intriguing, it is not clear whether they are a useful way of thinking about the jackknife. We won't pursue them further here.” the pseudovalue-based rAUC method proposed in this work could be regarded by some as controversial by statisticians (S. L. Hillis, private communication, 2011). The very designation of these quantities as “pseudovalues” carries, in the author's experience, a negative connotation with some statisticians. Efron has also stated (Efron 1982) “Attempts to extract additional information from the [pseudovalues], beyond the [variance estimates], have been disappointing” (in brackets are words in place of Efron's symbols). Estimating agreement is equivalent to determining the correlated and uncorrelated variances of two random variables, so our proposal is consistent with the second comment from Efron. Moreover, the *physical*

interpretation of the pseudovalues derived in this work, namely their relationship to the correctness of individual laboratory interpretations, gives in the author's opinion, a powerful reason for not giving up on pseudovalues, at least in this particular application. The proposed method has been tested with a simulator which captures, in our opinion, the essence of the laboratory ratings ROC task and the binary clinical task. An open source version of the simulator in the widely used (in statistics) and free R-programming language (R 2011) is being made available on the web (www.devchakraborty.com).

The choice of fraction correct as a measure of clinical performance may be criticized on the grounds that it depends on disease prevalence. One of the advantages for using appropriately normalized measures, such as AUC, is that then performance becomes independent of disease prevalence. However, what is an advantage in the context of a laboratory performance measurement need not be an advantage in the context of clinical performance measurement. As an example consider two radiologists with identical areas under the ROC curves, but who operate at different points on the curve: radiologist #1 has high sensitivity and low specificity and radiologist #2 has low sensitivity and high specificity. In a sufficiently low prevalence population radiologist #2 will get more credit because the skill at ruling out negatives will be literally more relevant to the patient population being served than radiologist #1's skill at ruling in positives. This would be reflected in the fraction correct (higher for radiologist #2) but not in AUC. In fact over time radiologist #1 will learn the characteristics of the population and move the operating point down the curve – an example of sensitivity and specificity being influenced by prevalence. Any measure that reflects clinical performance must depend, at the very least, on prevalence and the choice of reporting threshold. This is because these determine the numbers of incorrect (false positives and false negatives) and correct (true positives and true negatives) decisions, and any cost-benefit analysis must explicitly account for them (Metz 1978).

Sensitivity and specificity are particularly unsuited as measures of clinical performance (Begg 1987), who note that “contrary to popular opinion, they are generally not invariant to the population under investigation. ... The (false) claim that the sensitivity and specificity are ‘fixed’ characteristics of a test has led to their adoption as measures preferred to the predictive values, the probabilities of disease or non-disease given the test result. A second drawback is the fact that the individual measures are arbitrary, although a pair of sensitivity/specificity values is not. ... The third drawback is the lack of direct relevance of sensitivity and specificity to decision making. Measures which are more appropriate in this context are the predictive values, which focus attention on the probability of disease in a given patient rather than the probabilities of different test results.” This is also the conclusion of (Fryback and Thornbury 1991) who ranked the change in clinical thinking, which is measured by the predictive values, higher (Level 3) than diagnostic accuracy (Level 2), see Introduction. Both positive and negative predictive values depend on prevalence and the reporting threshold. However, like sensitivity and specificity they are coupled, and therefore cannot be used as a scalar performance measure.

The purpose of this paper is not to suggest how clinical performance ought to be measured - the literature on this subjective is fairly extensive, particularly in mammography which is a tightly regulated area of imaging - rather, *given* a scalar measure of clinical performance, how one should measure its agreement with laboratory ROC performance. The clinical relevance measure could be chosen to whatever is considered to be most justifiable, and as long as it is a scalar, the method described here is applicable. Fraction correct is a particularly simple measure with intuitive appeal, and it seemed reasonable to the clinicians who have worked with us (Chakraborty, Hagood et al. 2012). Methods need to be developed for non-scalar measures, e.g., positive and negative predictive values, which are more widely accepted than fraction correct.

Based on the simulation results one can make some preliminary suggestions on how to conduct a study aimed at measuring the agreement between laboratory and clinical interpretations using the rAUC agreement measure. A rating scale with sufficient number of bins should be used, because then the effect of performance mismatch is minimized. About 8 bins (which is actually equivalent to 6 bins because the way the cutoffs are defined there are very few counts outside of the extreme cutoffs in Fig. 1) should be sufficient, see Fig. 6 (d). The second precaution, to minimize the drop in agreement due to mismatched thresholds, see Fig. 4, is to arrange it so that the central laboratory cutoff is comparable to the clinical cutoff. This can be determined by estimating the sensitivity and specificity of the clinical reader on the common image set and training the laboratory reader to attain similar values for their Yes/No decision – this may not be easy. The cases can be enriched to facilitate estimation of the ROC area. As shown in Fig. 7 asymmetry in the case-mix has almost no effect on the rAUC agreement measure. If these precautions are taken, then the measured agreement should be limited primarily by the intrinsic correlation between the laboratory and clinical raw ratings. [Fig. 3 panel (a) should not be construed as implying that fewer bins are desirable for an ROC study. Pearson correlation is greater for 4 bins than a larger number of bins only when the offset parameter is zero and the performances are identical. If either of these conditions is not satisfied, then the Pearson correlation drops. Rather, the point of Fig. 3, viewing both panels, is that the relevance AUC is superior to Pearson correlation at detecting the perfect correlation existing at the z-sample level independent of the degree of binning and the value of the common performance level.]

Pseudovalues were introduced into ROC analysis in the context of multiple-reader multiple-case (MRMC) study significance testing. The widely-used Dorfman-Berbaum-Metz (DBM) significance testing method (Dorfman, Berbaum et al. 1992) for MRMC analysis calculates the pseudovalues for each modality, reader and image. The pseudovalues are used as-if they were contributions from individual cases. In earlier versions of MRMC software it was assumed that the pseudovalues should behave like real figures-of-merit, and “unphysical” values greater than unity were “clipped” to unity. To the best of our knowledge the current version of the software does not implement this artificial clipping. As we have shown, the individual contributions can assume apparently “unphysical” values, but their average is always exactly equal the actual trapezoidal area in the absence of jackknifing, so artificial clipping is not called for. Note that the result is true only for the trapezoidal area. If other figures-of-merit are used, e.g., the fitted area under the ROC curve (Dorfman and Alf 1968), then the result is no longer strictly valid, and work needs to be done to investigate whether the departure from exactness has practical consequences. Finally, a first-principles analysis of the expected behavior of the Pearson correlation and rAUC for the simulator model is needed, which might explain why, for example, the peak in Fig. 6(e) occurs below the point where the two performances are identical.

5. CONCLUSIONS

Standard AUC-based agreement measures could severely underestimate agreement that exists at the individual image level. The pseudovalue-based relevance ROC method is more sensitive to agreement at the individual image level, and the area under the relevance ROC is proposed as a measure of the clinical relevance of a laboratory observer performance paradigm.

Acknowledgments

DPC was supported in part by grants from the Department of Health and Human Services, National Institutes of Health, R01-EB005243 and R01-EB008688.

APPENDIX A

The pseudo-value A_{kt}^* for an arbitrary image kt is defined by Eqn. 7,

$$A_{kt}^* = KA - (K - 1)A_{(kt)}. \quad \text{Eqn. A1}$$

The removal of a specific image can be accommodated by subtracting the appropriate contribution to the double summation in Eqn. 5. Therefore, the pseudo-value for disease-free image $k_1 1$ is given by

$$A_{k_1 1}^* = \frac{K}{N_N N_A} \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) - \frac{K-1}{(N_N - 1) N_A} \left[\sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) - \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) \right] \quad \text{Eqn. A2}$$

In Eqn. A2, the first term is K times the trapezoidal area calculated over all cases, and the second term is $(K-1)$ times the trapezoidal area calculated over all cases except disease-free image $k_1 1$. Summing over all disease-free cases, one gets

$$\sum_{k_1=1}^{N_N} A_{k_1 1}^* = \frac{K N_N}{N_N N_A} \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) - \frac{K-1}{(N_N - 1) N_A} \left[N_N \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) - \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) \right] \quad \text{A3}$$

This can be simplified as follows:

$$\begin{aligned} \sum_{k_1=1}^{N_N} A_{k_1 1}^* &= \left[\frac{K}{N_A} - \frac{(K-1)N_N}{(N_N-1)N_A} + \frac{K-1}{(N_N-1)N_A} \right] \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) \\ \sum_{k_1=1}^{N_N} A_{k_1 1}^* &= \frac{1}{N_A} \left[K - \frac{(K-1)N_N}{N_N-1} + \frac{K-1}{N_N-1} \right] \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) \\ \sum_{k_1=1}^{N_N} A_{k_1 1}^* &= \frac{1}{N_A} \left[K + \frac{K-1}{N_N-1} (1 - N_N) \right] \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) \\ \sum_{k_1=1}^{N_N} A_{k_1 1}^* &= \frac{1}{N_A} [K - (K-1)] \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) \end{aligned}$$

Therefore,

$$\sum_{k_1=1}^{N_N} A_{k_1 1}^* = \frac{1}{N_A} \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}). \quad \text{Eqn. A4}$$

Likewise, the pseudo-value for disease-present image $k_2 2$ is

$$A_{k_2 2}^* = \frac{K}{N_N N_A} \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) - \frac{K-1}{(N_A - 1) N_N} \left[\sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) - \sum_{k_1=1}^{N_N} \psi(r_{k_1 1}, r_{k_2 2}) \right] \quad \text{Eqn. A5}$$

The summation of A5 over disease-present cases is

$$\sum_{k_2=1}^{N_A} A_{k_2 2}^* = \frac{N_A K}{N_N N_A} \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) - \frac{K-1}{(N_A - 1) N_N} \left[N_A \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) - \sum_{k_2=1}^{N_A} \sum_{k_1=1}^{N_N} \psi(r_{k_1 1}, r_{k_2 2}) \right] \quad \text{A6}$$

Upon simplification this reduces to

$$\sum_{k_2=1}^{N_A} A_{k_2 2}^* = \frac{1}{N_N} \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}). \quad \text{Eqn. A7}$$

Eqn. A4 and A7 lead to the result Eqn. 9, namely, the *averages* of the pseudovalues, calculated *separately* over disease-free and disease-present cases, are *individually* equal to the (common) trapezoidal area A :

$$\begin{aligned} \frac{1}{N_N} \sum_{k_1=1}^{N_N} A_{k_1}^* &= \frac{1}{N_N N_A} \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) = A \\ \frac{1}{N_A} \sum_{k_2=1}^{N_A} A_{k_2}^* &= \frac{1}{N_N N_A} \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) = A. \end{aligned} \quad \text{Eqn. A8}$$

Subtracting the average value, i.e., A , from the pseudovalue for disease-free image $k_1 1$ one has (see tentative definition of $\chi_{k_1 1}$ in Eqn. 8):

$$\chi_{k_1 1} = A_{k_1 1}^* - A = \frac{K-1}{N_A N_N} \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) - \frac{K-1}{(N_N-1) N_A} \left[\sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) - \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) \right] \quad \text{Eqn. A9}$$

Upon simplification this reduces to

$$\chi_{k_1 1} = \frac{-(K-1)}{(N_N-1) N_A} \left[\frac{1}{N_N} \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) - \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) \right]. \quad \text{Eqn. A10}$$

Similarly, for disease-present image $k_2 2$ and using Eqn. 8 one has

$$\chi_{k_2 2} = \frac{K-1}{N_A N_N} \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) - \frac{K-1}{(N_A-1) N_N} \left[\sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) - \sum_{k_1=1}^{N_N} \psi(r_{k_1 1}, r_{k_2 2}) \right] \quad \text{Eqn. A11}$$

Upon simplification this reduces to

$$\chi_{k_2 2} = \frac{-(K-1)}{N_N (N_A-1)} \left[\frac{1}{N_A} \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) - \sum_{k_1=1}^{N_N} \psi(r_{k_1 1}, r_{k_2 2}) \right]. \quad \text{Eqn. A12}$$

Define

$$\begin{aligned} P_{k_1 1} &= \frac{1}{N_A} \sum_{k_2=1}^{N_A} \psi(r_{k_1 1}, r_{k_2 2}) \\ P_{k_2 2} &= \frac{1}{N_N} \sum_{k_1=1}^{N_N} \psi(r_{k_1 1}, r_{k_2 2}). \end{aligned} \quad \text{Eqn. A13}$$

Then

$$\begin{aligned} \chi_{k_1 1} &= \frac{K-1}{N_A (N_N-1)} [N_A P_{k_1 1} - N_A A] = \frac{K-1}{(N_N-1)} [P_{k_1 1} - A] \\ \chi_{k_2 2} &= \frac{K-1}{N_N (N_A-1)} [N_N P_{k_2 2} - N_N A] = \frac{K-1}{(N_A-1)} [P_{k_2 2} - A]. \end{aligned} \quad \text{Eqn. A14}$$

Probabilistic interpretation

P_{k_11} is the probability that a disease-present image rating exceeds the rating of a *specific* disease-free image k_11 . The first equation in A14 shows that χ_{k_11} is a constant times the difference of probability P_{k_11} from its average value A , i.e., it is proportional to the probability that the average disease-present image rating exceeds the rating of *specific* disease-free image k_11 minus the probability that the average disease-present image rating exceeds the rating of an *average* disease-free image. In other words, χ_{k_11} is a measure of the difference of a *specific* disease-free image k_11 from the *average* disease-free image, or how far from average a specific disease-free image is. If image k_11 is a particularly easy image, then the difference χ_{k_11} will be positive, and conversely if it is particularly difficult, the difference will be negative.

Likewise, P_{k_22} is the probability that a disease-free image rating is smaller than the rating of a *specific* disease-present image k_22 , and the second equation in A14 shows that χ_{k_22} is a measure of the difference of a *specific* disease-present image k_22 from the *average* disease-present image, or how from average a specific disease-present image is. If image k_22 is a particularly easy image, then the difference χ_{k_22} will be positive, and conversely if it is particularly difficult, the difference will be negative.

It follows from the preceding equations that using redefined correctness measures, as defined in Eqn. 12, one has (note that the proportionality constant is now unity)

$$\begin{aligned}\chi_{k_11} &= P_{k_11} - A \\ \chi_{k_22} &= P_{k_22} - A.\end{aligned}\quad \text{Eqn. A15}$$

APPENDIX B

The clinical figure-of-merit is defined in Eqn. 6, namely,

$$C = \frac{1}{N_N + N_A} \left[\sum_{k_1=1}^{N_N} \psi(z_{2k_11}, \zeta_{clin}) + \sum_{k_2=1}^{N_A} \psi(\zeta_{clin}, z_{2k_22}) \right]. \quad \text{Eqn. B1}$$

Equivalently,

$$C = \frac{1}{K} \left[\sum_{k_1=1}^{N_N} \psi(z_{2k_11}, \zeta_{clin}) + \sum_{k_2=1}^{N_A} \psi(\zeta_{clin}, z_{2k_22}) \right]. \quad \text{Eqn. B2}$$

If one deletes disease-free image k_1 the clinical figure-of-merit is

$$C_{k_11} = \frac{1}{K-1} \left[\sum_{k_1=1}^{N_N} \psi(z_{2k_11}, \zeta_{clin}) - \psi(z_{2k_11}, \zeta_{clin}) + \sum_{k_2=1}^{N_A} \psi(\zeta_{clin}, z_{2k_22}) \right]. \quad \text{Eqn. B3}$$

Applying Eqn. 7 (the pseudovalue formula) one gets the following expression for the clinical figure-of-merit pseudovalue for disease-free image k_1

$$C^*_{k_11} = KC - \frac{K-1}{K-1} \left[\sum_{k_1=1}^{N_N} \psi(z_{2k_11}, \zeta_{clin}) - \psi(z_{2k_11}, \zeta_{clin}) + \sum_{k_2=1}^{N_A} \psi(\zeta_{clin}, z_{2k_22}) \right] = \psi(z_{2k_11}, \zeta_{clin}) = \text{Eqn. B4}$$

This yields the desired result, namely, the pseudo-value for disease-free image k_1 is identical to the clinical figure-of-merit for that image.

Similarly, for disease-present image k_2 the derivation is

$$C_{k_2} = \frac{1}{K-1} \left[\sum_{k_1=1}^{N_N} \psi(z_{2k_1}, \zeta_{clin}) + \sum_{k_2=1}^{N_A} \psi(\zeta_{clin}, z_{2k_2}) - \psi(\zeta_{clin}, z_{2k_2}) \right]. \quad \text{Eqn. B5}$$

$$C_{k_2}^* = KC - \frac{K-1}{K-1} \left[\sum_{k_1=1}^{N_N} \psi(z_{2k_1}, \zeta_{clin}) + \sum_{k_2=1}^{N_A} \psi(\zeta_{clin}, z_{2k_2}) - \psi(\zeta_{clin}, z_{2k_2}) \right] = \psi(\zeta_{clin}, z_{2k_2}) = \zeta_{clin}. \quad \text{Eqn. B6}$$

The pseudo-value for disease-present image k_2 is identical to the clinical figure-of-merit for that image, which is zero or one depending on whether or not the clinical z-sample *does not exceed* the clinical threshold ζ_{clin} .

APPENDIX C

Consider the special case of perfectly correlated z-sample data, i.e., $\rho_{lab,clin}^z = 1$, binary laboratory ratings ($R = 1$), zero offset $\zeta_{offset} = 0$ and identical performances, $\mu_{common} \equiv \mu_{lab} = \mu_{clin}$. Let F_1 denote the number of disease-free images rated 1 and F_2 denote the number of disease-free images rated 2. Likewise, let T_1 denote the number of diseased images rated 1 and T_2 denote the number of diseased images rated 2. By Eqn. 5,

$$\begin{aligned} N_N N_A A &= F_1 T_2 + 0.5 (F_1 T_1 + F_2 T_2) \\ N_N N_A A &= F_1 T_2 + 0.5 (F_1 (N_A - T_2) + (N_N - F_1) T_2) \\ N_N N_A A &= F_1 T_2 + 0.5 (N_A F_1 + N_N T_2 - 2F_1 T_2) \\ N_N N_A A &= 0.5 (N_A F_1 + N_N T_2) \\ A &= 0.5 \left(\frac{F_1}{N_N} + \frac{T_2}{N_A} \right) \end{aligned}$$

The clinical figure-of-merit is defined by

$$C = \frac{F_1 + T_2}{N_N + N_A}$$

In the special case of balanced data, i.e., $N_N = N_A = N$, the last expression can be seen to be identical to A , i.e.,

$$C = \frac{F_1 + T_2}{2N} = 0.5 \left(\frac{F_1}{N} + \frac{T_2}{N} \right) = A$$

Therefore, for balanced data the two figures-of-merit will be perfectly correlated.

REFERENCES

- Begg CB. Biases in the assessment of diagnostic tests. *Statistics in Medicine*. 1987; 6:411–423. [PubMed: 3114858]

- Bunch PC, Hamilton JF, et al. A Free-Response Approach to the Measurement and Characterization of Radiographic-Observer Performance. *J of Appl Photogr. Eng.* 1978; 4(4):166–171.
- Burgess AE. Comparison of receiver operating characteristic and forced choice observer performance measurement methods. *Med. Phys.* 1995; 22(5):643–655. [PubMed: 7643805]
- Chakraborty DP. Counterpoint to “Performance Assessment of Diagnostic Systems Under the FROC Paradigm” by Gur and Rockette. *Acad Radiol.* 2009; 16:507–510. [PubMed: 19268864]
- Chakraborty DP. New Developments in Observer Performance Methodology in Medical Imaging. *Semin Nucl Med.* 2011; 41(6):401–418. [PubMed: 21978444]
- Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: Modeling, analysis and validation. *Med Phys.* 2004; 31(8):2313–2330. [PubMed: 15377098]
- Chakraborty DP, Hagood TM, et al. Quantifying the clinical relevance of a laboratory observer performance paradigm. *Br J Radiol.* 2012 in press.
- Cohen J. A coefficient of agreement for nominal scales. *Edu. and Psych. Meas.* 1960; 20:37–46.
- Dodd LE, Wagner RF, et al. Assessment methodologies and statistical issues for computer-aided diagnosis of lung nodules in computed tomography: contemporary research topics relevant to the lung image database consortium. *Acad Radiol.* 2004; 11(4):462–475. [PubMed: 15109018]
- Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory- a direct solution. *Psychometrika.* 1968; 33(1):117–124. [PubMed: 5239566]
- Dorfman DD, Berbaum KS, et al. ROC characteristic rating analysis: Generalization to the Population of Readers and Patients with the Jackknife method. *Invest. Radiol.* 1992; 27(9):723–731. [PubMed: 1399456]
- Efron, B. The Jackknife, the Bootstrap and Other Resampling Plans. Capital City Press; Montpelier: 1982.
- Efron, B. CBMS-NSF Regional Conference Series in Applied Mathematics. Capital City Press; Montpelier, Vermont, USA: 1982. The Jackknife, the Bootstrap, and Other Resampling Plans..
- Fryback DG, Thornbury JR. The Efficacy of Diagnostic Imaging. *Med Decis Making.* 1991; 11(2):88–94. [PubMed: 1907710]
- Gifford HC, King MA, et al. Channelized hotelling and human observer correlation for lesion detection in hepatic SPECT imaging. *Journal of nuclear medicine.* 2000; 41(3):514–521. [PubMed: 10716327]
- Gur D, Bandos AI, et al. The “Laboratory” Effect: Comparing Radiologists’ Performance and Variability during Prospective Clinical and Laboratory Mammography Interpretations. *Radiology.* 2008; 249(1):47–53. [PubMed: 18682584]
- Gur D, Bandos AI, et al. The Prevalence Effect in a Laboratory Environment: Changing the Confidence Ratings. *Acad Radiol.* 2007; 14:49–53. [PubMed: 17178365]
- Gur D, Bandos AI, et al. Is an ROC-type Response Truly Always Better Than a Binary Response in Observer Performance Studies? *Academic Radiology.* 2010; 17(5):639–645. [PubMed: 20236840]
- Gur D, Rockette HE. Performance Assessment of Diagnostic Systems under the FROC paradigm: Experimental, Analytical, and Results Interpretation Issues. *Acad Radiol.* 2008; 15:1312–1315. [PubMed: 18790403]
- Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982; 143(1):29–36. [PubMed: 7063747]
- Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC studies. *Statistics in Medicine.* 2007; 26:596–619. [PubMed: 16538699]
- Kundel HL, Berbaum KS, et al. Receiver Operating Characteristic Analysis in Medical Imaging, ICRU Report 79. 2008; 8(1)
- Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics.* 1947; 18:50–60.
- Metz CE. Basic principles of ROC analysis. *Seminars in Nuclear Medicine.* 1978; 8(4):283–298. [PubMed: 112681]
- Metz CE. Some Practical Issues of Experimental Design and Data Analysis in Radiological ROC studies. *Investigative Radiology.* 1989; 24:234–245. [PubMed: 2753640]

- Narayanan MV, Gifford HC, et al. Optimization of iterative reconstructions of ^{99m}Tc cardiac SPECT studies using numerical observers. *IEEE Transactions on Nuclear Science*. 2002; 49(5): 2355–2360.
- NCRP. An Introduction to Efficacy in Diagnostic Radiology and Nuclear Medicine (Justification of Medical Radiation Exposure), Commentary No. 13.. NCRP Reports. 1995.
<http://www.ncrppublications.org/>
- Obuchowski NA, Lieber ML, et al. Data Analysis for Detection and Localization of Multiple Abnormalities with Application to Mammography. *Acad. Radiol.* 2000; 7(7):516–525. [PubMed: 10902960]
- R. A language and environment for statistical computing: R Development Core Team. R Foundation for Statistical Computing; Vienna, Austria: 2011.
- Roe CA, Metz CE. Dorfman-Berbaum-Metz Method for Statistical Analysis of Multireader, Multimodality Receiver Operating Characteristic Data: Validation with Computer Simulation. *Acad Radiol.* 1997; 4:298–303. [PubMed: 9110028]
- Samuelson F, Gallas BD, et al. The Importance of ROC Data. *Academic Radiology*. 2011; 18(2):257–258. [PubMed: 21232688]
- Sickles EA, Ominsky SH, et al. Medical audit of a rapid-throughput mammography screening practice: methodology and results of 27,114 examinations. *Radiology*. 1990; 175(2):323–327. [PubMed: 2326455]
- Spring DB, Kimbrell-Wilmot K. Evaluating the success of mammography at the local level: how to conduct an audit of your practice. *The Radiologic clinics of North America*. 1987; 25(5):983–992.
- Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. *Med. Phys.* 1996; 23(10):1709–1725. [PubMed: 8946368]
- Thornbury JR. Eugene W. Caldwell Lecture. Clinical efficacy of diagnostic imaging: love it or leave it. *Am. J. Roentgenol.* 1994; 162(1):1–8. [PubMed: 8273645]
- Wagner RF, Metz CE, et al. Assessment of Medical Imaging Systems and Computer Aids: A Tutorial Review. *Acad Radiol.* 2007; 14(6):723–748. [PubMed: 17502262]
- Wilcoxon F. Individual Comparison by Ranking Methods. *Biometrics*. 1945; 1:80–83.
- Wolfe JM, Horowitz TS, et al. Rare items often missed in visual searches. *Nature*. 2005; 435(26):439. [PubMed: 15917795]

Fig. 1.

This shows a schematic of the laboratory – clinical simulator. The upper panel illustrates the sampling for the laboratory reader with performance μ_{lab} and the lower panel for the clinical reader with performance μ_{clin} . The laboratory reader uses a nominal 6-point rating scale corresponding to 5-cutoffs ζ_i depicted by the heavy dotted lines. The central cutoff $\zeta_3 = \mu_{clin}/2$ is depicted by the light dotted line. The clinical cutoff is depicted in the lower panel of the figure with heavy dotted line, $\zeta_{clin} = \mu_{clin}/2 + \zeta_{offset}$ where ζ_{offset} is shown negative in this example. Two z-samples are shown for the laboratory reader denoted by z_{151} for the 5th normal case and by z_{132} 3rd abnormal case. The corresponding z-samples for the clinical reader are denoted by z_{251} and z_{232} , respectively. The z-samples are modeled by Eqn. 1. [$N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2 .]

Fig. 2.

The purpose of this figure is to illustrate the differences between the conventional method of measuring agreement and the proposed rAUC method. Panel (a) shows a plot of ΔA vs. ΔC for 50 normal and 50 abnormal images, perfectly correlated z-sample data, equal laboratory and clinical performances, and $\zeta_{offset} = 0$, when the laboratory data was binned into 4 bins ($R = 3$). The dependence is very close to linear with $r^2 = 0.995$ (with 2 bins the dependence would be exactly linear). Panel (b) is the corresponding relevance ROC with area $rAUC = 1.000$. Panel (c) shows a plot of ΔA vs. ΔC when the data was not binned and all other conditions remained the same. The correlation decreased to $r^2 = 0.755$ but the relevance ROC shown in Panel (d) remained close to perfect with $rAUC = 0.999$. This shows a strong effect of binning vs. not binning on the Pearson correlation, and that the rAUC agreement measure is relatively unaffected.

Fig. 3.

Panel (a) shows a plot, for perfectly correlated z-sample data $\rho_{\text{lab,clin}}^z=1$, $\Delta\mu_{22}=0$ and ζ_{offset} of the Pearson correlation coefficient vs. μ_{common} for different number of bins as indicated by the numeric symbols (“inf” = infinite number of bins, i.e., no binning was performed). It is observed that (1) even though the pre-binned laboratory and the clinical z-samples are identical, the Pearson correlation is generally less than unity, (2) fine binning decreases correlation relative to coarse binning, and (3) the decrease in correlation between a finite number of bins and an infinite number of bins is greater at large μ than at small μ . Panel (b) is the corresponding plot for rAUC. The numeric symbols line up one top of each other and the rAUC measure is almost constant at unity, independent of μ_{common} and number of bins. This figure shows that the relevance AUC is superior to Pearson correlation

at detecting the perfect correlation existing at the z-sample level independent of the degree of binning and the value of the common performance level.

Fig. 4.

This figure shows the Pearson correlation coefficient (the filled circles) and the rAUC agreement measure (the asterisks) plotted vs. the offset parameter for different combinations of μ and R . Panel (a) is for relatively large $\mu_{\text{common}} = 2.33$ (equivalent to $A = 0.95$) and two bins ($R = 1$) while panel (b) is for infinite number of bins ($R = \text{infinity}$). Panels (c) and (d) are corresponding results for $\mu_{\text{common}} = 0.358$ (equivalent to $A = 0.60$). In each panel the agreement measures are seen to decrease, as $|\zeta_{\text{offset}}|$ increases. This is expected because although the laboratory figure-of-merit and its fluctuations are unaffected by non-zero

offset, the clinical figure-of-merit and its fluctuations are affected. Minimizing the offset effect is important in future studied aimed at detecting laboratory vs. clinical correlation.

Fig. 5.

This figure shows the variation of the Pearson correlation coefficient (filled circles) and the rAUC agreement measure (asterisks) vs. $\rho_{lab,clin}^z$ for $R=5$, $\zeta_{offset}=0$ and $\mu_{common}=1.81$ (equivalent to $A=0.9$). As $\rho_{lab,clin}^z$ increased, all agreement measures increased, as expected, because the laboratory and clinical z-samples became increasingly correlated. At $\rho_{lab,clin}^z=0$ all agreement measures were at chance level (0 for r and 0.5 for rAUC).

Fig. 6.

This figure shows the variation of Pearson correlation coefficient (filled circles) and the rAUC agreement measure (asterisks) vs. μ_{lab} for fixed clinical performance $\mu_{clin} = 1.812$ (equivalent to $A = 0.90$) indicated by the vertical lines. The other simulation parameters were $\rho_{lab,clin}^z = 1$ and $\zeta_{offset} = 0$. Panel (a) is for $R = 1$, panel (b) for $R = 3$, panel (c) for $R = 5$, panel (d) for $R = 7$ and panel (e) is for $R = \text{infinity}$. For finite number of cutoffs both measures peak near μ_{clin} and fall off away from this value, but the fall off is more pronounced for the Pearson measure. For $R = 7$ and infinity, the rAUC agreement measure had less degradation than the correlation measures. The reasons for the secondary peaks (both measures exhibit it) are explained in the text. The purpose of this figure is to show that $R = 7$, which corresponds in reality to a 6-bin laboratory study, is a good choice for a future relevance ROC study.

Fig. 7.

This figure shows the variation of Pearson correlation coefficient (filled circles) and the rAUC agreement measure (asterisks) for varying disease prevalence in the range 0.01 to 0.8. Panel (a) is for $R = 1$ (two bins) and panel (b) is for infinite number of bins. Note the logarithmic scale of the x-axis. The purpose of this figure is to show the almost total insensitivity of the relevance AUC measure to the disease prevalence, because it is sensitive to the perfect agreement at the z-sample level, while the Pearson measure degrades rapidly as not moves away from equal numbers of disease-absent and disease-present cases ($D = 0.5$).

Table 1

This table compares the conventional ROC and the relevance ROC.

Conventional ROC		Relevance ROC	
Quantity	Meaning	Quantity	Meaning
Abnormal images	Disease present	Correct clinical interpretation images	Radiologist's clinical interpretation was judged correct by correlation with additional imaging.
Normal images	Disease absent	Incorrect clinical interpretation images	Radiologist's clinical interpretation was judged incorrect by correlation with additional imaging
Ratings	Subjective impression of disease presence	laboratory correctness measure	Derived from pseudovalues, i.e., differences of laboratory figures-of-merit
FP	Normal image rated above a threshold	Relevance FP	Incorrect clinical interpretation image with laboratory correctness measure exceeding a threshold
TP	Abnormal image rated above a threshold	Relevance TP	Correct clinical interpretation image with laboratory correctness measure exceeding a threshold
FPF	# FPs divided by number of normal images	Relevance FPF	# relevance FPs divided by number of incorrect clinical interpretation images
TPF	# TPs divided by number of normal images	Relevance TPF	# relevance TPs divided by number of correct clinical interpretation images
ROC	Plot of TPF vs. FPF	Relevance ROC	Plot of relevance TPF vs. relevance FPF
Area under ROC (AUC)	Ability of ratings to discriminate between normal and abnormal images	Area under relevance ROC (rAUC)	Ability of laboratory correctness measure to discriminate between correct and incorrect clinical interpretation images

Table 2

Summary statistics for various agreement measures for the simulations shown in Fig. 3 where the common performance level and the binning was varied. Listed are the maximum (MAX), minimum (MIN), the average (MEAN) and the coefficient of variation (CV = standard deviation divided by the mean). Note the wide variability of the correlation measures even though the laboratory and clinical data are perfectly correlated. In contrast the rAUC agreement measure remains very close to unity.

Vary number of bins and $\mu_{common} \equiv \mu_{lab} = \mu_{clin}$						
	A	C	Correlation measures			rAUC
			Pearson	Spearman	Kendall	Kappa
MAX	0.951	0.879	0.999	0.999	0.979	0.958
MIN	0.572	0.568	0.814	0.801	0.617	0.609
MEAN	0.753	0.716	0.919	0.911	0.777	0.768
CV	0.158	0.150	0.062	0.067	0.150	0.149
						1.000
						0.999
						1.000
						0.000

Summary statistics for various agreement measures for the simulations shown in Fig. 4 where the offset parameter ζ_{offset} was varied. In this case one expects the correlation measures to vary between perfect and near random agreement, and all statistics including the rAUC agreement measure follow the expected trend.

Table 3

Vary ζ_{offset} , $\mu_{common} \equiv \mu_{lab} = \mu_{clin}$						
	A	C	Correlation measures			rAUC
			Pearson	Spearman	Kendall	Kappa
MAX	0.952	0.879	1.000	1.000	1.000	1.000
MIN	0.566	0.544	0.341	0.327	0.228	0.225
MEAN	0.749	0.699	0.679	0.662	0.502	0.493
CV	0.224	0.204	0.237	0.247	0.313	0.317
						0.136

Summary statistics for the various agreement measures for the simulations shown in Fig.5 where the intrinsic correlation $\rho_{lab,clin}^z$ between the laboratory and the clinical decision variables was varied. Here one expects the agreement measures to vary between perfect and random agreement, and all statistics including the rAUC agreement measure followed the expected trends.

Table 4

Vary $\rho_{lab,clin}^z, \mu_{common} = \mu_{lab} = \mu_{clin}$						
	A	C	Correlation measures			rAUC
			Pearson	Spearman	Kendall	Kappa
MAX	0.863	0.821	0.943	0.936	0.794	0.783
MIN	0.857	0.815	-0.010	-0.008	-0.005	-0.021
MEAN	0.859	0.818	0.242	0.234	0.170	0.166
CV	0.001	0.002	1.078	1.094	1.186	1.216
						0.222

Table 5

Summary statistics for the various agreement measures for the simulations shown in The rAUC agreement measure had less undesirable variation than the correlation measures.

Vary μ_{lab} fixed $\mu_{clin} = 1.81$						
	A	C	Correlation measures			rAUC
			Pearson	Spearman	Kendall	
MAX	0.995	0.822	0.995	0.994	0.960	1.000
MIN	0.500	0.814	0.479	0.459	0.323	0.877
MEAN	0.803	0.818	0.758	0.742	0.567	0.967
CV	0.190	0.002	0.131	0.138	0.191	0.036

Summary statistics for the various agreement measures for the simulations shown in Fig. 7, when the disease-prevalence was varied from 0.01 to 0.8. The conventional correlation measures were sensitive to this design detail of the ROC study, but the rAUC measure was almost completely insensitive to it.

Table 6

Vary prevalence, $\mu_{common} \equiv \mu_{lab} = \mu_{clin}$						
	A	C	Correlation measures			rAUC
			Pearson	Spearman	Kendall	Kappa
MAX	0.903	0.820	1.000	1.000	1.000	1.000
MIN	0.812	0.815	0.167	0.160	0.109	0.101
MEAN	0.858	0.818	0.614	0.604	0.476	0.476
CV	0.049	0.002	0.435	0.435	0.521	0.507
						0.999
						0.002