Quantifying the clinical relevance of a laboratory observer performance paradigm

¹D P CHAKRABORTY, PhD, ²T M HAYGOOD, PhD, MD, ³J RYAN, PhD, MIOPeG, ²E M MAROM, MD, PhD
⁴M EVANOFF, PhD, ABR, ³M F McENTEE, MIOPeG and ³P C BRENNAN, MIOPeG

¹Department of Radiology, University of Pittsburgh, Pittsburgh, PA, USA, ²University of Texas MD Anderson Cancer Center, Houston, TX, USA, ³The University of Sydney, Sydney, NSW, Australia, and ⁴American Board of Radiology, Tucson, AZ, USA

Objective: Laboratory observer performance measurements, receiver operating characteristic (ROC) and free-response ROC (FROC) differ from actual clinical interpretations in several respects, which could compromise their clinical relevance. The objective of this study was to develop a method for quantifying the clinical relevance of a laboratory paradigm and apply it to compare the ROC and FROC paradigms in a nodule detection task.

Methods: The original prospective interpretations of 80 digital chest radiographs were classified by the truth panel as correct ($C=1$) or incorrect ($C=0$), depending on correlation with additional imaging, and the average of C was interpreted as the clinical figure of merit. FROC data were acquired for 21 radiologists and ROC data were inferred using the highest ratings. The areas under the ROC and alternative FROC curves were used as laboratory figures of merit. Bootstrap analysis was conducted to estimate conventional agreement measures between laboratory and clinical figures of merit. Also computed was a pseudo-value-based image-level correctness measure of the laboratory interpretations, whose association with C as measured by the area (rAUC) under an appropriately defined relevance ROC curve, is as a measure of the clinical relevance of a laboratory paradigm.

Results: Low correlations (e.g. $\kappa=0.244$) and near chance level rAUC values (e.g. 0.598), attributable to differences between the clinical and laboratory paradigms, were observed. The absolute width of the confidence interval was 0.38 for the interparadigm differences of the conventional measures and 0.14 for the difference of the rAUCs.

Conclusion: The rAUC measure was consistent with the traditional measures but was more sensitive to the differences in clinical relevance. A new relevance ROC method for quantifying the clinical relevance of a laboratory paradigm is proposed.

Received 3 September
2011
Revised 17 November 2011
Accepted 20 December
2011

DOI: 10.1259/bjr/45866310

© 2012 The British Institute of
Radiology

Observer performance studies are widely used in radiology [1–5] for imaging systems assessment. Typically, radiologists are asked to view images and find abnormalities with two modalities, and then their diagnostic accuracy in one modality is compared with their accuracy in the other modality. When the diagnostic task involves finding focal lesions, there are two fundamentally different approaches for assessing accuracy: approaches that ignore information regarding the locations of suspected lesions, an example of which is the receiver operating characteristic (ROC) approach, and approaches that account for this information, an example of which is the free-response ROC (FROC) approach [6–8].

Assume the task is detecting malignant nodules in digital chest radiographs. In the ROC approach the

radiologist rates each radiograph for confidence in the presence of one or more nodules in the image. In the FROC approach the radiologist marks regions suspicious for presence of nodules and rates each suspicious region for confidence that there is actually a nodule at the indicated location. A mark is classified as a lesion localisation (the lesion was correctly localised) or a non-lesion localisation (a lesion-free region was marked) according to its distance from the nearest nodule in the image. The choice of nearness criterion [9] is usually guided by clinical considerations, e.g. nodule diameter or 3 mm, whichever is greater. The ROC ratings and the FROC mark ratings are used to estimate summary measures of performance (figures of merit) that reward correct decisions and penalise incorrect ones.

Observer performance studies such as ROC and FROC represent laboratory simulations of the clinical task. They are designed to simulate the basic aspects of the clinical task, namely having radiologists find disease on actual images under conditions of uncertainty, but the simulations cannot reproduce all factors affecting clinical interpretations. The image collections are usually enriched to contain a larger fraction of diseased patients

Address correspondence to: Dr Dev Chakraborty, Department of Radiology, Presby South Tower, Room 4771, 200 Lothrop Street, University of Pittsburgh, Pittsburgh, PA 15213, USA. E-mail: dpc10@pitt.edu

This work was supported by grants from the Department of Health and Human Services, National Institutes of Health, R01-EB005243 and R01-EB008688. TMH was supported in part by a grant from the John S. Dunn, Sr. Distinguished Chair Funds.

than is normally encountered in the clinic, the images tend to be more subtle (difficult normals and subtle lesions) and there is no "clinical pressure" since there are no consequences to correct or incorrect decisions in the laboratory. The differences from "real life" could degrade the clinical relevance of the laboratory paradigms [10, 11]. The aim of this work was to develop a method for assessing the clinical relevance of a laboratory paradigm and to use it to compare the ROC and FROC paradigms.

The method relies (1) on an additional item of information beyond that necessary for a conventional ROC study and (2) on a physical interpretation of jackknife-derived pseudovalues. These will be explained more fully in the Methods and materials section, but an introduction may be helpful. The additional information is the prospective (*i.e.* the original clinical) interpretation of the images, classified as correct or incorrect. For example, a nodule-free image that, in the clinic, was interpreted as nodule free constitutes a correct prospective interpretation, as does a nodule-containing image in which the nodule was reported. Likewise, a nodule-free image that was interpreted as containing a nodule or a nodule-containing image in which the nodule was missed constitute incorrect prospective interpretations.

Now, with regard to the pseudovalues, suppose that one performs an ROC study under laboratory conditions with the same images and a representative group of readers other than the original interpreting radiologists. The ROC interpretations are analysed to obtain a figure of merit, *e.g.* the area under the ROC curve. The pseudovalue is defined in terms of the difference of figures of merit induced by removal of an image. It will be shown later that the pseudovalues can be interpreted as the correctness of the laboratory interpretations of individual images. In ROC analysis, a lesion-free image correctly interpreted with high confidence (*e.g.* "definitely lesion free") will result in a larger pseudovalue than a lesion-free image that receives a lower confidence in the absence of a lesion. This is because a "definitely lesion free" rating is the most correct laboratory interpretation possible on a lesion-free image. A lesion-free image that receives a "probably lesion free" rating will result in a smaller pseudovalue, and at the other extreme a lesion-free image that receives a "definitely lesion present" rating will result in the smallest pseudovalue, because such a rating is the worst possible laboratory interpretation on a lesion-free image. Likewise, a lesion-present image that receives a "definitely lesion free" rating will result in a smaller pseudovalue and one that receives a "definitely lesion present" rating will result in a larger pseudovalue. So, in general, correct laboratory interpretations with result in larger pseudovalues than incorrect laboratory interpretations.

The laboratory-determined pseudovalues can be interpreted as quasi-ratings in the binary task of classifying the images according to their correct/incorrect prospective interpretations. The quasi-ratings yield a ROC-like curve termed the relevance ROC (rROC). To the extent that larger pseudovalues are associated with correct prospective interpretations, and smaller ones with incorrect prospective interpretations, the area under the rROC curve (rAUC) will be >0.5 . The rAUC measures the correct/incorrect

classification accuracy of the pseudovalues, and is therefore a measure of clinical relevance. The procedure can be repeated with a different method, *e.g.* FROC, which yields different pseudovalues and a different rROC curve will be obtained. The method (ROC or FROC) yielding the greater area under the corresponding rROC curve has greater clinical relevance, because the changes in the laboratory-determined pseudovalues are better at classifying the actual correct/incorrect prospective interpretations of the images. A methodology that is better at mirroring already known clinical decisions may then also be better at predicting the accuracy of clinical decisions that may be made in the future, *i.e.* it will be clinically more relevant.

In the following sections the image selection procedure, the correct/incorrect classifications, the image acquisition and display, the observer study and the statistical analysis are described. Results are presented illustrating the potential and some of the problems with this type of measurement, and suggestions are made for conducting future rROC studies.

Methods and materials

Image selection

The images and radiology reports used in this study were obtained under an institutional review board-approved protocol and are Health Insurance Portability and Accountability Act compliant. The need for informed consent from patients was waived, but the radiologist readers gave informed consent. Two board-certified diagnostic radiologists (the truth panel) with extensive experience in the interpretation of chest radiographs and chest CT images, chose the 80 postero-anterior chest radiographs included in this study. 40 of the included radiographs had 1–4 (average 1.225; total 49) visible pulmonary nodules agreed upon by both radiologists and confirmed by CT. Likewise, the interpretation of the CT was also agreed upon by both radiologists. The nodule sizes ranged from 4.7 to 44.0 mm (average 17.0 mm). Images with calcified granulomas were avoided as it would be ambiguous to the readers whether these should count as a nodule or not. The truth panel included several images that had additional nodules at CT if the nodules were <1 cm in diameter and both radiologists agreed that the additional nodules were not visible on the radiographs.

There were 40 images without nodules. The method of proving these images varied depending on the original interpretation of the image. If the image had been prospectively interpreted as having no nodules, it qualified for inclusion if both radiologists who chose the images agreed that there were no nodules both on the image and on a follow-up chest radiograph obtained a minimum of 6 months later and which was also prospectively interpreted as having no nodule. If the image had been prospectively interpreted as having a nodule or possible nodule, it qualified for inclusion if a CT demonstrated no nodule, both by the prospective interpretation of the CT and by unanimous opinion of both radiologists who chose that image.

When choosing these images, particularly the nodule-free images, precautions were taken to avoid giving subtle clues as to which images did and which did not have nodules. For example, included among both the radiographs with nodules and the radiographs without nodules were various incidental abnormalities such as surgical clips or scoliosis. Thus, the images without nodules were not as a group pristinely lesion free. Age range and sex distribution were reasonably similar between the groups as well. The patients whose radiographs were included as having no nodules ranged in age from 17 to 88 years (average 53 years); 23 were male and 17 female. The patients whose radiographs were included as having nodules ranged in age from 20 to 91 years (average 59 years); 19 were male and 21 female.

Correct/incorrect classification of the prospective interpretations

In addition to determining disease status ground truth, the truth panel retrieved the prospective interpretations of the images and classified them as correct or incorrect. Various radiologists practising chest radiography at our institution, none of whom participated in the observer study to be described later, made these interpretations. A prospective interpretation of an image as having (or not having) nodules was classified as correct if the presence (or absence) of nodules was confirmed by the truth panel. Likewise, a prospective interpretation of an image as having (or not having) nodules was classified as incorrect if the absence (or presence) of nodules was confirmed by the truth panel. There were no instances in our data set of images in which the prospective reader identified an incorrect location on a nodule-containing image. Such images are relatively rare in our clinic, where the images are viewed with the lateral view, prior studies, including CTs, and radiology reports, and we avoided them as it would not be clear how to classify them.

Images with nodules consisted of 20 correct prospective interpretation images and 20 incorrect prospective interpretation images. 16 out of the 20 incorrect prospective interpretation images with nodules (*i.e.* those that had nodules but had been interpreted as having none) had contemporaneous CT. While the reason for these CTs was not entirely clear, they may have been protocol driven and would have been obtained regardless of the chest radiograph interpretation. In four cases the CT was obtained later, after a follow-up radiograph was interpreted as having nodules. 40 nodule-free radiographs were split between 34 correct prospective interpretations and 6 incorrect prospective interpretations.

Image acquisition and display

71 of the images were acquired on a General Electric Advantx Revolution XQ/i machine (GE Healthcare, Waukesha, WI) and 9 on a General Electric Definium 8000 (GE Healthcare). Images were displayed using a ViewSonic® VP201m 5.1-cm liquid crystal display (ViewSonic, Walnut, CA), with a maximum luminance of 250 cd m^{-2} , contrast ratio of 600:1 and a non-interlaced matrix of

1600×1200 . The monitor's maximum luminance did not fall below 250 cd m^{-2} , minimum luminance did not exceed 1 cd m^{-2} ; and geometry, luminance uniformity, temporal stability, resolution and veiling glare were within American Association of Physicists in Medicine Task Group 18 recommended levels [12]. The display was calibrated each morning of the study to the greyscale display function standard [12] with Verilum® software (Image Smiths Inc., Bethesda, MD) and luminance pod. Ambient lighting was maintained at a constant level throughout image interpretation and measured using a calibrated photometer (model 07-621, Nuclear Associates, Carle Place, NY). Images were presented using RocViz software (Vizova Technologies, Dublin, Ireland), which allowed images to be displayed in random order and allowed collection of the FROC data, *i.e.* the locations and ratings of suspected nodule-containing regions.

Observer study

21 radiologists, all board certified by the American Board of Radiology, participated as readers in the study. The radiologists came from varying specialties, and their years of experience post certification ranged from 7 to 59. The number of cases they had read during the preceding year ranged from <10 a week (500 per year) to 400 cases per week (20 800 cases per year). Readers were instructed that the task was to find and rate pulmonary nodules and they were instructed in how to operate the computer program that displayed the images and how to indicate the nodules and their locations. The readers were not told the number of nodules on each image or how many of the images contained nodules.

For each image the radiologist used the computer mouse to click on the centre of each suspected nodule and, using a menu at the bottom of the screen, assigned to it a confidence level rating of 1, 2, 3 or 4. A rating of 1 corresponded to the lowest confidence level that there was a true nodule at the indicated site, and a rating of 4 corresponded to the highest confidence. If the distance from the radiologist's indicated nodule centre to the centre of the closest nodule as defined by the truth panel was <70 pixels (12 mm at mid-patient plane), then the mark was scored as a lesion localisation; otherwise, it was scored as non-lesion localisation. ROC data were inferred from the FROC data by identifying the rating of the highest rated mark on an image as its inferred ROC rating. If the image had no marks a zero rating was inferred. For example, if a radiologist marked two regions and assigned ratings of 2 and 4 to them, then the image as a whole would be considered to have an inferred ROC rating of 4 for the presence of a nodule somewhere on the image. Therefore, the 4-point FROC rating scale 1, 2, 3 or 4 translated to the 5-point ROC rating scale 0, 1, 2, 3 or 4.

Statistical analysis

Laboratory figures of merit

The ROC curve is defined as the plot of true-positive fraction (TPF) *vs* false-positive fraction (FPF). Lesion

localisation fraction (LLF) is defined as the fraction of lesions with ratings exceeding a threshold and which lesions are localised to within the adopted proximity criteria. The highest rating assumption is used to infer the ROC rating of a nodule-free image, and the fraction of nodule-free images with highest ratings exceeding a threshold is the FPF. The alternative FROC (AFROC) curve is defined as the plot of the LLF *vs* the FPF. ROC performance was quantified by the area A under the trapezoidal (empirical) ROC curve, and FROC performance by the area θ under the trapezoidal AFROC curve [13–16]. These were calculated for each reader (j is the reader index, $j=1, 2, \dots, J$, for J readers) using

$$A_j = \frac{1}{N_N N_A} \sum_{k=1}^{N_N} \sum_{k'=1}^{N_A} \psi(X_{jk}, Z_{jk'}) \quad (1a)$$

$$\theta_j = \frac{1}{N_N N_L} \sum_{k=1}^{N_N} \sum_{k'=1}^{N_A} \sum_{l_k'=1}^{N_{k'}} \psi(X_{jk}, Y_{jk'l_k'}) \quad (1b)$$

Here N_N is the number of lesion-free images; N_A is the number of lesion-present images; N_L is the total number of lesions; $N_{k'}$ is the number of lesions on lesion-present image k' ; X_{jk} is the rating, for reader j , of the highest rated mark on lesion-free image k , or zero if the image has no marks (if the rating scale includes negative numbers, then the “zero” is replaced by any value smaller than the smallest allowed rating); $Z_{jk'}$ is the rating, for reader j , of the highest rated mark on lesion-present image k' , or zero if the image has no mark; $Y_{jk'l_k'}$ is the rating, for reader j , of lesion l_k' on lesion-present image k' , or zero if the lesion is not marked; 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. The reader-averaged ROC area $A_\bullet$ (the dot indicates the average over the corresponding symbol) was interpreted as the laboratory figure of merit when using the ROC method and $\theta_\bullet$ was interpreted as the laboratory figure of merit when using the AFROC method.

In the case of the ROC area A_j [Equation (1a)], the trapezoidal estimate of the area under the ROC curve is equivalent to the Mann–Whitney Wilcoxon U -statistic [17, 18]. Using this definition it has been shown [19] that the variance of the AUC is the sum of two terms, one with an inverse dependence on the number of lesion-free images and the other with an inverse dependence on the number of lesion-present images. Clearly, if either of these numbers is small, then the statistical precision of the measurement will be compromised. For this reason, the numbers of lesion-free and lesion-present images in the study were chosen to be equal.

Clinical figure of merit

The correct/incorrect classifications of the images were coded by the classification vector C_k , $k=1, 2, \dots, N_N+N_A$, whose first N_N elements denoted the classifications of the lesion-free images, 0 for an incorrect classification and 1 for a correct classification. The remaining N_A elements of C denoted the classifications of the lesion-present images. The average (over images) of the classification vector $C_\bullet$ was interpreted as the clinical figure of merit.

Traditional agreement measure estimation and confidence interval

The ROC study yields a single value for the laboratory figure of merit ($A_\bullet$) and the clinical interpretations yield a single value for the clinical figure of merit ($C_\bullet$). To measure agreement one needs to induce statistically valid fluctuations in these quantities. This can be accomplished using the bootstrap technique [20], which consists of regarding the data set as defining the population, and to sample with replacement from it (since a population cannot be exhausted, one must sample with replacement). The fluctuations associated with these samples allow estimation of an agreement measure. For example, if positive excursions of the laboratory figure of merit tend to be associated with positive excursions of the clinical figure of merit (and negative with negative) then agreement is good. The agreement can be quantified by one of several available traditional measures, such as the Pearson, the Spearman or the Kendall correlation coefficients. Note that the procedure yields a single value for the chosen agreement measure, *i.e.* it yields a point estimate. To measure its variability, or equivalently a 95% confidence interval, one needs to repeat the entire process using bootstrapped versions of the original data set as the starting points, termed the bootstrap-after-bootstrap technique. Then the standard bootstrap percentile method [20] can be applied to the array of agreement measures so determined in order to estimate a confidence interval.

One can perhaps distinguish two levels of bootstrapping—an “inner loop”, which is necessary to obtain a point estimate of the agreement measure, and an “outer loop”, which generates as many point estimates of the agreement measure as there are outer-loop values—whose spread allows estimation of a confidence interval for the agreement measure. A simple example with 10 images may be helpful (Figure 1). The images are labelled by the consecutive numbers 1, 2, 3, ..., 10. A particular outer-loop bootstrap sample might be the numbers 10, 3, 4, 2, 10, 10, 1, 6, 4, 8. It is seen that, for this particular bootstrap sample, some images occur more than once and some do not occur at all. An inner-loop bootstrap sample is obtained by sampling with replacement from the outer-loop image set. One example could be the set of 10 images denoted by the numbers 2, 3, 4, 8, 10, 4, 6, 4, 3, 4. One calculates laboratory and clinical figures of merit for this set of images. The procedure is repeated with more inner-loop bootstraps. For example, the second inner-loop bootstrap might be the images 10, 10, 1, 2, 6, 6, 1, 6, 4, 8, which will yield another pair of estimates of the laboratory and clinical figures of merit. When one has accumulated a sufficient number of pairs, an agreement measure can be calculated between the laboratory and clinical figures of merit. The procedure described so far yields a single point estimate of the agreement measure. The entire procedure is repeated with different outer-loop bootstrap samples. For example, the second outer-loop bootstrap sample might be 1, 2, 4, 5, 9, 9, 1, 7, 1, 5. After drawing a sufficient number of inner-loop bootstrap data sets from this sample one calculates a second value for the agreement measure. The procedure is repeated for a sufficient number of outer-loop bootstrap samples. The spread of the agreement measure between the different outer-loop samples

Figure 1. The bootstrap-after-bootstrap method used to estimate confidence intervals for the conventional correlation measures and their differences is shown in Table 2. The left tall dotted rectangle illustrates the outer loops, and the other dotted rectangles illustrate the inner loops. The data set is assumed to consist of 10 images labelled by the integers 1–10. Assuming the alternative free-response receiver operating characteristic (AFROC) figure-of-merit-based measure, for the Pearson, Spearman and Kendall correlation measures the quantities x and y are defined in Equation (2a,b) and for kappa they are defined in Equation (3a,b). For the receiver operating characteristic (ROC) figure of merit-based measures the corresponding equations are Equation (2a,c) and Equation (3a,c). For each chosen correlation measure at each inner-loop step the procedure yields two correlations and a difference (AFROC–ROC). ρ_1 is the correlation measure corresponding to the first outer-level bootstrap etc, and CI_ρ is the 95% confidence interval for the correlation measure.

determines the confidence interval. While the procedure described may appear computationally overwhelming, it is well within the capability of current desktop computers. Readers were not bootstrapped (*i.e.* readers were treated as a fixed effect).

Formally, $B'=500$ outer-loop bootstrap data sets, indexed by $b'=1, 2, \dots, B'$, were generated from the original data set. For each b' the laboratory figure of merit values were averaged over readers yielding $A_{\bullet b'}$, $\theta_{\bullet b'}$ and the average of the C-array over cases $C_{\bullet b'}$ was calculated (notation: the dot on A and θ represents the average over readers and that on C represents the average over images; in either case, the second index is the outer-loop bootstrap index). For each b' , $B=200$ inner-loop bootstrap samples were generated, indexed by b ($b=1, 2, \dots, B$). For

inner-loop bootstrap sample b , the reader-averaged laboratory figures of merit, $A_{\bullet b' b}$ and $\theta_{\bullet b' b}$, and the clinical figure of merit, $C_{\bullet b' b}$, were calculated (the varying number of subscripts in the notation reflects the fact that these quantities depend on both b' and b).

At this point one can calculate conventional agreement measures, for fixed b' , between $x_{b' b}$ and $y_{b' b}^\theta$ and between $x_{b' b}$ and $y_{b' b}^A$, where

$$x_{b' b} = C_{\bullet b' b} - C_{\bullet b'} \quad (2a)$$

$$y_{b' b}^\theta = \theta_{\bullet b' b} - \theta_{\bullet b'} \quad (2b)$$

$$y_{b' b}^A = A_{\bullet b' b} - A_{\bullet b'} \quad (2c)$$

For example, one can calculate the Pearson correlation coefficient, the Spearman correlation coefficient and the Kendall correlation coefficient [22]. The kappa measure of agreement [23] can also be calculated provided the fluctuations are categorised into two or more levels. In this work, we categorised into two levels, *i.e.* the differences from the corresponding outer-loop values were replaced with 0s and 1s according to the signs of the differences:

$$\Delta C_{b'b} = H(C_{b'b} - C_{\bullet b'}) \quad (3a)$$

$$\Delta \theta_{b'b} = H(\theta_{b'b} - \theta_{\bullet b'}) \quad (3b)$$

$$\Delta A_{b'b} = H(A_{b'b} - A_{\bullet b'}) \quad (3c)$$

where H is the Heaviside unit step function (positive for non-negative argument, and zero otherwise.). For fixed b' and b running from 1 to B , there will be one kappa measure between $\Delta C_{b'b}$ and $\Delta \theta_{b'b'}$ and another between $\Delta C_{b'b}$ and $\Delta A_{b'b}$. Thus, for each b' and analysis method, three correlation measures and the kappa agreement statistic were calculated, and the procedure was repeated with the other analysis method. Using the percentile method, 95% confidence intervals for these statistics were calculated from the $B'=500$ outer-loop bootstraps.

Proposed new analysis

The previous method measures correlations or agreement between fluctuations of the laboratory and clinical figures of merit from their baseline values. The figures of merit represent averages over all images [this is explicit in the notation $C_{\bullet}$ for the clinical figure of merit, but not in the definition of the laboratory figure of merit A_j via Equation (1a): the ROC area for reader j is equivalent to the average probability that a rating sample from a diseased image is greater than that from a non-diseased image] and such averaging could mask more transparent agreement observable at the image level. For example, consider an easy disease-containing image in which the lesion is highly visible. In the laboratory study this might elicit a "4" rating (high confidence nodule present) and the clinical interpretation is likely to be correct ($C=1$). Conversely, an easy nodule-free image in which the normal structures are visualised clearly would also be likely to elicit a "1" rating in the laboratory study (high confidence disease absent) and the clinical interpretation of the image is also likely to be correct. This type of individual image-level agreement may not be easily evident from the figure-of-merit fluctuations, which represent averages over all images. If one could devise an image-level measure of the correctness of the laboratory decisions, denoted by χ_k for image k , then a more sensitive measure of agreement might be possible.

A correctness measure can be devised by using the concept of jackknife pseudovalues [20] and an interpretation of the pseudovalues [21]. Suppressing the reader index for now, then, corresponding to the laboratory figures of merit A and θ , defined by Equation (1a,b), the jackknife pseudovalues A_k^* and θ_k^* for image k are defined by

$$A_k^* = KA - (K-1)A_{(k)} \quad (4a)$$

$$\theta_k^* = K\theta - (K-1)\theta_{(k)} \quad (4b)$$

Here, K is the total number of images, $A_{(k)}$ is the ROC figure of merit with image k removed from the analysis and $\theta_{(k)}$ is the AFROC figure of merit with image k removed from the analysis. Examination of this equation yields the following insight. Removal of an easy image, which could be nodule present or nodule free, leaves behind a slightly more difficult image set, which means that $A_{(k)}$ and $\theta_{(k)}$ will decrease, perhaps ever so slightly. Since it appears with a negative sign, the jackknife pseudovalues A_k^* and θ_k^* will increase, perhaps ever so slightly. Likewise, removal of a difficult image leaves behind an easier data set and the pseudovalues A_k^* and θ_k^* will decrease. Removal of an image that is of average difficulty level will leave behind a data set with unchanged difficulty level, and the pseudovalues will be relatively unchanged. The pseudovalues mirror the correctness of the laboratory decisions on the individual images [21]. Subtracting the average pseudovalues over all images, the correctness measures χ_k^A and χ_k^θ are defined by

$$\chi_k^A = A_k^* - A_{\bullet}^* \quad (5a)$$

$$\chi_k^\theta = \theta_k^* - \theta_{\bullet}^* \quad (5b)$$

The correctness measures are expected to be positive for images on which correct laboratory decisions were made, and negative for images on which incorrect laboratory decisions were made, and near zero for images on which the decisions were neutral (*e.g.* ROC rating equal to 3).

The pseudovalues transformation was also applied to the clinical figure of merit. It is easily shown, using Equation (4), that the k th pseudovalues is identical to C_k . In other words,

$$C_k^* = KC_{\bullet} - (K-1)C_{\bullet(k)} = C_k \quad (6)$$

Therefore, with good ROC laboratory *vs* clinical agreement one expects $\chi_k^A > 0$ and $C_k = 1$ on images with correct laboratory and correct clinical decisions, and $\chi_k^A < 0$ and $C_k = 0$ on images with incorrect laboratory and incorrect clinical decisions. Considering the ROC case first, χ_k^A 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 C_k . Analogous to ROC analysis, one can define a relevance ROC (rROC) curve, the area under which (rAUC) is a measure of how effective χ_k^A is at discriminating between correct and incorrect clinical interpretation images. This suggests that rAUC is a measure of laboratory-clinical association or the clinical relevance of the laboratory paradigm. The x -axis of the rROC is the relevance FPF, defined as an incorrect clinical interpretation with χ_k^A exceeding an arbitrary threshold. The y -axis is the relevance TPF, defined as a correct clinical interpretation with χ_k^A exceeding a threshold. The threshold defines a specific operating point on the rROC curve, and changing the threshold sweeps out the rROC curve. Table 1 compares

Table 1. This table depicts the analogies and differences between a conventional receiver operating characteristic (ROC) and the relevance ROC

Conventional ROC		rROC	
Quantity	Meaning	Quantity	Meaning
Lesion-present images	Disease present	Correct prospective interpretation images	Radiologist's prospective interpretation was judged correct by correlation with additional imaging
Lesion-free images	Disease absent	Incorrect prospective interpretation images	Radiologist's prospective interpretation was judged incorrect by correlation with additional imaging
Ratings	Subjective impression of disease presence	Correctness measure χ	Derived from pseudovalues, <i>i.e.</i> differences of figures of merit, which could be AFROC-based or ROC-based
FP	Lesion-free image rated above a threshold	Relevance FP	Incorrect prospective interpretation image with correctness measure χ exceeding a threshold value
TP	Lesion-present image rated above a threshold	Relevance TP	Correct prospective interpretation image with correctness measure χ exceeding a threshold value
FPF	Number of FPs divided by number of lesion-free images	Relevance FPF	Number of relevance FPs divided by number of incorrect prospective interpretation images
TPF	Number of TPs divided by number of lesion-present images	Relevance TPF	Number of relevance TPs divided by number of correct prospective interpretation images
ROC	Plot of TPF vs FPF	rROC	Plot of relevance TPF vs relevance FPF
Area under ROC	Ability of ratings to discriminate between lesion-free and lesion-present images	Area under rROC	Ability of correctness measure χ to discriminate between correct and incorrect prospective interpretation images

AFROC, alternative free-response receiver operating characteristic; FP, false positive; FPF, false-positive fraction; rROC, relevance ROC; TP, true positive; TPF, true-positive fraction.

the conventional ROC with the rROC. The correct and incorrect clinical interpretations are analogues of normal and abnormal images, respectively, χ_k^A is the analogue of the ROC rating and the area under the rROC curve is the analogue of the area under the ROC curve. Because it links the correctness of interpretations in the laboratory study with the correctness of actual clinical interpretations, a large rAUC implies high clinical relevance. The rAUC measure is close to unity if positive values of χ_k are predominantly associated with $C_k = 1$ and negative values are associated with $C_k = 0$. If there is little (or random) association, the rAUC is expected to be close to 0.5.

The area under the trapezoidal rROC curve is defined by [compare with Equation (1a)]

$$\text{rAUC} = \frac{1}{N_0 N_1} \sum_{r=1}^{N_0} \sum_{s=1}^{N_1} \psi(\chi_r, \chi_s) \quad (7)$$

Here, N_0 is the number of zeros in the C_k array and N_1 is the number of ones, 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_k array, s is the image index (N_1 values also in the range 1 to K , not necessarily consecutive) of the s th one-element in the C_k array, rAUC is the area under the trapezoidal ROC curve defined by the correctness measures (χ_r, χ_s) interpreted as N_0 "lesion-free" and N_1 "lesion-present" image "ratings", which measures their classification ability with respect to the 0s and 1s that make up the C_k array.

The procedure described above for the ROC figure of merit was repeated with the FROC figure of merit, *i.e.* using χ_k^0 instead of χ_k^A . Clearly, rAUC depends on the choice of figure of merit, so corresponding to the ROC figure of merit one has rAUC^A and corresponding to the

FROC figure of merit one has rAUC^0 . Both were calculated for each reader; for each cycle of the outer loop, a reader average was obtained, and 95% confidence intervals were constructed for rAUC^A , rAUC^0 and their difference $\text{rAUC}^0 - \text{rAUC}^A$.

Equation (7) is similar to Equation (1a), except that the roles of lesion-free and lesion-present images are replaced by the numbers of incorrect ($C=0$) and correct ($C=1$) prospective interpretations, respectively. Therefore, in order to achieve maximum precision, one needs equal numbers of incorrect and correct prospective interpretations. For reasons to be discussed later we were not quite successful at achieving this (there were 54 correct interpretations and 26 incorrect interpretations).

Results

Figure 2a shows the AFROC figure-of-merit fluctuations *vs* the clinical figure-of-merit fluctuations [the fluctuations are defined in Equation (2)]. The outer-loop index b' is fixed whereas the inner-loop index b takes on 200 values, yielding 200 points in the plot. The linear fit follows $y=0.209x$, with intercept near zero, and Pearson correlation coefficient equal to 0.301. Figure 2b shows the ROC figure-of-merit fluctuations *vs* the clinical figure-of-merit fluctuations. The linear fit follows $y=0.275x$, with Pearson correlation coefficient equal to 0.423. For this single outer-loop bootstrap sample the correlation was stronger for ROC. 500 outer-loop samples were used to estimate confidence intervals.

Table 2 summarises the results of the bootstrap analysis. It lists the various agreement measures (Pearson, Spearman and Kendall correlation coefficients, kappa

Figure 2. (a) The alternative free-response receiver operating characteristic (AFROC) figure-of-merit fluctuations vs the clinical figure-of-merit fluctuations [as defined in Equation (2)]. The outer-loop index b' is fixed whereas the inner-loop index b takes on 200 values, yielding 200 points in the plot. The linear fit follows $y=0.209x$, with intercept near zero, and Pearson correlation coefficient equal to 0.301. (b) The receiver operating characteristic (ROC) figure-of-merit fluctuations vs the clinical figure-of-merit fluctuations. The linear fit follows $y=0.275x$, with Pearson correlation coefficient equal to 0.423. For this single outer-loop bootstrap sample the correlation was stronger for ROC. 500 outer-loop samples were used to estimate confidence intervals (the dots represent an average over the reader index).

and rAUC) for AFROC-derived and ROC-derived figures of merit, and the corresponding differences (AFROC minus ROC) and 95% confidence intervals. The agreement measures follow the ordering Pearson>Spearman>Kendall>kappa. All agreement measures were small, and all AFROC-derived measures were slightly larger; however, none of the differences was significant. The absolute width of the confidence interval was approximately 0.38 for the difference of the correlation measures and kappa, and 0.14 for the difference of rAUC.

Under "Actual ratings, conventional AUCs", Table 3 lists AFROC and ROC areas for all readers and their averages and 95% confidence intervals in parentheses. Since FROC requires localisation, the AFROC figures of merit are smaller than the corresponding ROC figures of merit (even if the lesion is not correctly localised, a high rated mark is counted as a true positive in ROC but not as lesion localisation in FROC). The average AFROC area was 0.660 whereas the average ROC area was 0.771. Under "Correctness measure as ratings, rAUCs", Table 3 lists the relevance AUCs and their averages and confidence intervals, for AFROC- and ROC-derived

correctness measures. The observed reader-averaged rAUC was 0.598 (0.49 to 0.70) using the AFROC-derived correctness measure and 0.564 (0.45 to 0.66) using the ROC-derived correctness measure. Unlike the values under "Actual ratings" the AFROC-generated rAUCs tended to be larger than the ROC-generated rAUCs. For 16 of the 21 readers the AFROC correctness measure classified the clinical decisions better than the ROC correctness measure. The average difference in rAUCs (AFROC minus ROC) was 0.034 (−0.04 to 0.11), *i.e.* AFROC yielded slightly higher AUCs, but the difference was not significant at the 5% level.

Figure 3a shows trapezoidal AFROC, labelled "LLF", and ROC, labelled "TPF", curves for Reader 1. The areas under these curves are 0.607 for the AFROC and 0.742 for ROC, respectively. Figure 3b shows the rROC curves for this reader. In both plots AFROC-related values are shown with filled circles and solid lines, and ROC-related values are shown with filled triangles and dotted lines. The rROC data are noisy since they are derived from pseudovalues, *i.e.* differences of figures of merit interpreted as ratings, not actual ratings. The area under

Table 2. Summary of the results of the agreement analysis

Quantity	Pearson	Spearman	Kendall	Kappa	rAUC
θ	0.373 (0.10, 0.58)	0.358 (0.07, 0.57)	0.254 (0.05, 0.41)	0.244 (0.04, 0.43)	0.594 (0.50, 0.67)
A	0.351 (0.05, 0.58)	0.337 (0.06, 0.56)	0.239 (0.04, 0.41)	0.227 (0.00, 0.42)	0.564 (0.46, 0.65)
$\Delta = \theta - A$	0.021 (−0.17 to 0.22)	0.021 (−0.17 to 0.22)	0.016 (−0.12 to 0.16)	0.017 (−0.17 to 0.18)	0.030 (−0.04 to 0.10)
Width of CI for Δ	0.39	0.39	0.38	0.37	0.14

The results for the conventional measures are listed under Pearson, Spearman, Kendall and Kappa, and the new relevance receiver operating characteristic (ROC)-based results are listed under relevance area under the curve (rAUC). Quantities in parentheses are 95% confidence intervals (CIs). The quantity θ denotes the alternative free-response receiver operating characteristic (AFROC) figure of merit (as in jackknife AFROC analysis) and A denotes the AUC figure of merit (as in ROC analysis). The CIs were calculated using the method described in Figure 1.

Table 3.

Reader	Actual ratings, conventional AUCs		Correctness measure as ratings, rAUCs		
	θ	A	rAUC ^{θ}	rAUC ^A	rAUC ^{θ} – rAUC ^A
1	0.607	0.742	0.660	0.642	0.019
2	0.645	0.748	0.591	0.614	–0.022
3	0.675	0.802	0.553	0.487	0.066
4	0.729	0.798	0.542	0.541	0.001
5	0.578	0.730	0.626	0.662	–0.036
6	0.586	0.779	0.569	0.502	0.067
7	0.700	0.835	0.487	0.459	0.028
8	0.606	0.703	0.576	0.588	–0.012
9	0.725	0.841	0.574	0.481	0.093
10	0.619	0.712	0.697	0.616	0.081
11	0.622	0.793	0.562	0.585	–0.024
12	0.742	0.808	0.539	0.474	0.065
13	0.715	0.822	0.598	0.548	0.050
14	0.616	0.731	0.646	0.600	0.047
15	0.658	0.713	0.568	0.504	0.064
16	0.675	0.740	0.630	0.530	0.100
17	0.646	0.743	0.671	0.571	0.100
18	0.764	0.906	0.592	0.683	–0.092
19	0.675	0.818	0.673	0.652	0.021
20	0.702	0.749	0.614	0.560	0.054
21	0.574	0.674	0.592	0.540	0.052
Avg. (95% CI)	0.660 (0.59 to 0.73)	0.771 (0.71 to 0.83)	0.598 (0.49 to 0.70)	0.564 (0.45 to 0.66)	0.034 (–0.04 to 0.11)

θ , area under trapezoidal alternative free-response receiver operating characteristic (AFROC) curve; A, area under trapezoidal receiver operating characteristic (ROC) curve; rAUC, area under trapezoidal relevance ROC curve.

Under "Actual ratings, conventional AUCs" are listed AFROC (θ) and ROC (A) figures of merit for all readers. Under "Correctness measure as ratings, rAUCs" are listed the rAUCs for all readers; rAUC ^{θ} =rAUC using the θ figure of merit; rAUC^A=rAUC using the A figure of merit. The last row lists the averages and, in parenthesis, the bootstrap 95% confidence intervals. The rAUC differences were not significant.

the curve labelled "AFROC-derived" is 0.660 and that under the curve labelled "ROC-derived" is 0.642 for a difference of 0.019. This reader showed a slight improvement in rAUC for AFROC-derived correctness measure. Figure 3c is the (normalised) histogram of the correctness measure split between correct (light fill) and incorrect (dark fill) prospective interpretations. Figure 3d is the corresponding histogram using the ROC-derived correctness measure. rROC curves can be constructed from these histograms by imagining a vertical line to the far right of each figure, and progressively moving it towards the left. The area under the light-filled histogram to the right of this line is the relevance TPF and the corresponding area under the dark-filled histogram is the relevance FPF. As the line is moved to the left, the rROC curves shown in Figure 3b are swept out. The two histograms shown in Figure 3c,d yield the rROC shown in Figure 3b for AFROC-derived (solid line, filled circles) and ROC-derived (dashed line, filled triangles) correctness measures, respectively. Figure 4 applies to Reader 18, who showed the maximum decrease (0.092) in rAUC for the AFROC-derived correctness measure, also readily visible in Figure 4b. Figure 5 is analogous to Figure 3 except it applies to Reader 17, who showed the maximum improvement (0.1) in rAUC for the AFROC-derived correctness measure, readily visible in Figure 5b.

Because of the unfamiliar nature of the rROC curve, a detailed example may be helpful. It is helpful in the following to bear in mind that image numbers 40–79 were nodule free and the rest had at least one nodule and

that there were 54 C=1 images and 26 C=0 images. Table 4 illustrates, for Reader 18 and Figure 4, the formation of the rROC curve using the ROC-derived correctness measure. In columns 3–6 the numbers of images are displayed in four columns split between incorrect (C=0) and correct (C=1) prospective decisions, and, for each decision category, between nodule-free and nodule-present images. In each cell there is a number followed by an alphabetical label referring to a table legend. For example, the entry "16 (a)" means that there were 16 nodule-free C=1 images, and the label (a) refers to note (a) in the table legend, where the actual image numbers in curly brackets followed by the common rating in parenthesis, and any comments, if applicable, are shown. In this example, the image numbers are 41, 45, 52, 54, 60, 62, 63, 65, 67, 68, 69, 70, 71, 72, 76, 77, and each of them was rated 0, *i.e.* they were not marked.

In the top half of Table 4 the contributions of large correctness measures are listed [see Equation (5) for definitions of the correctness measures]. The correctness measures for images of the cell labelled 16 (a) were highest (just exceeding the threshold $\chi_k^A=0.166$), resulting in the triangle symbol on the *y*-axis in Figure 4b. Since there are 16 such images, relevance TPF=16/54=0.296. An unmarked nodule-free image represents the best possible laboratory decision, and since these decisions occurred on correct prospective outcome images (in the clinic these images were correctly reported as nodule free), it results in upward movement of the relevance operating point, which tends to increase the rAUC. Note the discreteness of the correctness measures: since the 16 unmarked

Figure 3. (a) Trapezoidal alternative free-response receiver operating characteristic (AFROC), labelled "LLF", and receiver operating characteristic (ROC), labelled "TPF", curves for Reader 1. (b) The relevance ROC curves for this reader. In both plots AFROC-related plots are shown with filled circles and solid lines, and ROC-related plots are shown with filled triangles and dotted lines. The relevance ROC data are noisy since they are derived from pseudovalues, *i.e.* differences in figures of merit interpreted as ratings, not actual ratings. This reader showed a slight improvement in relevance area under the curve (0.019) when using the AFROC-derived correctness measure. (c) The (normalised) histogram of the correctness measure split between correct (light fill) and incorrect (dark fill) prospective interpretations using the AFROC-derived correctness measure. (d) The corresponding histogram using the ROC-derived correctness measure. FPF, false-positive fraction; LLF, lesion localisation fraction; TPF, true-positive fraction.

images each had the same ROC rating (*i.e.* zero, for an unmarked image) they yielded the same correctness measure. In fact, because the ROC ratings scale allowed only five ratings, there were, for this reader, only eight distinct values of the correctness measures, yielding the eight operating points listed in the table.

The next highest correctness measure (just exceeding $\chi_k^4=0.140$) yielded the data shown in the row corresponding to operating point 2. Images 58 and 59 represent C=0 nodule-free images that were rated 1 (the minimum of the rating scale). These are relatively good laboratory decisions (not marking them would have been even better), but since they occur on C=0 images (in the clinic these images were falsely reported as containing nodules) this results in rightward movement along the rROC, which tends to reduce rAUC. Images 2, 4, 6, 7, 9, 10, 12, 14, 17 and 19 represent C=0 nodule-present images (these nodules were missed in the prospective interpretations). The nodule on image 2 was rated 4 (the highest on the rating scale). The

nodule on image 4 was missed but a nodule-free region on that image was rated 4, and by the highest rating assumptions the image ROC rating was 4. The nodule on image 6 was rated 4, and likewise for images 7, 9, 10, 12, 14, 17 and 19. All of these represent very good decisions in the ROC sense (the single "right for wrong reason" decision on image 4 counts as a right decision in ROC scoring), but since they occur on C=0 images, this results in further rightward movement along the rROC, which tends to reduce rAUC. The net value of relevance is $\text{FPF}=(2+10)/26=0.462$, listed under the column labelled relevance FPF. In the same row of the table, nodule-free C=1 images (these images were correctly interpreted prospectively as not containing nodules) 44, 49, 75, 78 were rated 1, which are relatively good decisions (unmarked images would have been even better) and this results in upward movement along the rROC, which tends to increase rAUC. The nodules on images 20, 21, 23, 24, 27, 28, 31, 32, 34, 35, 36, 37, 38 were rated 4, which is the best possible decision on such

Figure 4. This figure is analogous to Figure 3 except it applies to Reader 18, who showed the maximum decrease (0.092) in relevance area under the curve for the alternative free-response receiver operating characteristic (AFROC)-derived correctness measure, readily visible in (b). See explanation of Table 4 in the text. FPF, false-positive fraction; LLF, lesion localisation fraction; ROC, receiver operating characteristic; TPF, true-positive fraction.

images, and since these are $C=1$ images (the nodules were detected in the prospective reads) this results in further upward movement along the rROC, tending to increase rAUC. The net value of relevance $TPF = 0.296 + (4+13)/54 = 0.611$.

In the bottom half of Table 4 the contributions of low (or large negative value) correctness measures are listed, corresponding to incorrect laboratory decisions—the more incorrect the decision, the more negative the correctness measure. To understand the behaviour of the rROC point in this region it helps to consider the movement of the operating point relative to the (1,1) corner, which is reached when all images are included, and consider the cumulative effect of *not including* images. Not including image 29, a $C=1$ nodule-present image, on which the nodule was not marked, the worst possible laboratory decision, results in downward movement to relevance $TPF = 1 - 1/54 = 0.981$, which tends to decrease rAUC. Likewise, not including nodule-free $C=0$ images 46 and 57 (incorrectly interpreted in the clinic as nodule containing), which were rated 4 (incorrect laboratory decisions, but consistent with the incorrect clinical decisions), moves relevance FPF to the left, to $1 - 2/26 = 0.923$, which tends to increase rAUC. Not

including the unmarked nodule-present $C=0$ image 16 results in leftward movement to relevance $FPF = 1 - (2+1)/26 = 0.885$, which tends to increase rAUC. Table 5 is similar to Table 4 except that it applies to Reader 17 and the AFROC-derived correctness measure, and the corresponding plot is the one labelled “AFROC-derived” in Figure 5b. The big jump in relevance TPF is due to the cell labelled “30 (f)”, which represents a big block of nodule-free $C=1$ images that were not marked by this reader (best possible laboratory decision and consistent with the clinical decisions). [This table has more operating points because the AFROC figure of merit, Equation (1b), involves comparisons of *lesion* ratings—not the highest ratings of nodule-present images as with ROC figures of merit, Equation (1a)—and highest ratings on nodule-free images, and the number of lesions is greater than the number of nodule-present images.]

To summarise, correct laboratory decisions tend to occur in the upper half of Tables 4 and 5, and, when these occur on $C=1$ images, there is upward movement relative to (0,0), which increases rAUC; when these occur on $C=0$ images, there is rightward movement relative to (0,0), which decreases rAUC. Incorrect laboratory

Figure 5. This figure is analogous to Figure 3 except it applies to Reader 17, who showed the maximum improvement (0.1) in relevance area under the curve (rAUC) for alternative free-response receiver operating characteristic (AFROC)-derived correctness measure. The improvement in rAUC is readily visible in (b). See also explanation of Table 5 in the text. FPF, false-positive fraction; LLF, lesion localisation fraction; ROC, receiver operating characteristic; TPF, true-positive fraction.

decisions tend to occur in the lower half of Tables 4 and 5, and, when these occur on $C=1$ images, there is downward movement relative to (1,1), which decreases rAUC; when these occur on $C=0$ images, there is leftward movement relative to (1,1), which tends to increase rAUC. Both halves of Tables 4 and 5 can be summarised by noting that incorrect laboratory decisions that are consistent with incorrect prospective readings tend to increase rAUC, as do correct laboratory decisions that are consistent with correct prospective readings, and images for which the laboratory and prospective readings are inconsistent decrease rAUC. This gives full meaning to rAUC as a measure of the clinical relevance of the laboratory study.

The argument made in this section has analogues to the interpretation of the ROC ratings and movement along the ROC curve. A high ROC-rated abnormal image, which represents a correct decision, results in upward movement of the operating point relative to the origin, which tends to increase AUC. Likewise, a high ROC-rated normal image, which represents an incorrect decision, results in rightward movement of the operating point relative to the origin, which decreases AUC. A low ROC-rated normal image represents a correct decision,

removal of which image results in leftward movement of the operating point relative to (1,1), which tends to increase AUC. Finally, a low ROC-rated abnormal image represents an incorrect decision, removal of which image results in downward movement of the operating point relative to (1,1), which tends to decrease AUC. Note the symmetry between the origin and (1,1) in the ROC. The operating points can be inferred by moving the threshold from larger to smaller values and observing the addition of operating points, or, equivalently, by moving the threshold from small to larger values and observing the deletion of operating points.

Discussion

Unlike most traditional ROC studies, in which the emphasis is on comparing two modalities, the emphasis in the current single modality study was on comparing the clinical relevance of two methods of analysing observer performance data. Clinical relevance was defined loosely as the association between the laboratory figure of merit (e.g. the area under the ROC curve) and the clinical figure of merit (defined as the fraction of prospective

Table 4. The formation of the relevance receiver operating characteristic (rROC) curve using the ROC-derived correctness measure, for Reader 18

ROC-derived correctness measure		Incorrect prospective outcome (C=0) no. of relevance FPs		Correct prospective outcome (C=1) no. of relevance TPs		Relevance FPF	Relevance TPF
Operating point	χ_k^A	Nodule free	Nodule present	Nodule free	Nodule present		
1	0.166			16 (a)		0	0.296
2	0.140	2 (b)	10 (c)	4 (d)	13 (e)	0.462	0.611
3	0.115			9 (f)		0.462	0.778
4	-0.087		9 (g)		6 (h)	0.808	0.889
5	-0.29	2 (i)		5 (j)		0.885	0.981
6	-0.493		1 (k)			0.923	0.981
7	-1.25	2 (l)				1	0.981
8	-1.43				1 (m)	1	1

χ_k^A , threshold value of ROC-derived correctness measure yielding the indicated operating point; FP, false positive; TP, true positive; FPF, false-positive fraction; TPF, true-positive fraction; the corresponding relevance ROC plot is Figure 4b; specifically, the plot labelled "ROC derived".

The image counts are displayed in four columns split between incorrect (C=0) and correct (C=1) prospective decisions, and for each decision category, between nodule-free and nodule-present images. In each cell there is a number, representing the image count, followed by an alphabetic label referring to the footnotes below. See text and notes for explanation.

{image numbers} (common rating): comments, if applicable

(a) {41, 45, 52, 54, 60, 62, 63, 65, 67, 68, 69, 70, 71, 72, 76, 77} (0): none of these images was marked

(b) {58, 59} (1)

(c) {2, 4, 6, 7, 9, 10, 12, 14, 17, 19} (4): the highest rated mark on each of these images was rated 4

(d) {44, 49, 75, 78} (1)

(e) {20, 21, 23, 24, 27, 28, 31, 32, 34, 35, 36, 37, 38} (4): the highest rated mark on each of these images was rated 4

(f) {40, 42, 43, 51, 53, 55, 61, 64, 79} (2)

(g) {0, 1, 3, 5, 8, 11, 13, 15, 18} (3): the highest rated mark on each of these images was rated 3

(h) {22, 25, 26, 30, 33, 39} (3)

(i) {47, 56} (3)

(j) {48, 50, 66, 73, 74} (3)

(k) {16} (2): the highest rated mark on this image was rated 2

(l) {46, 57} (4)

(m) {29} (0): the lesion was not marked

interpretations that were judged correct by correlation with additional imaging). Different ways of measuring the association were explored. To the best of our knowledge a truly satisfactory method for measuring agreement between laboratory ROC decisions and clinical decisions has not been previously proposed; indeed, to our knowledge, no one has proposed the type of study that was conducted here. This is the reason why we tried various conventional methods and developed a new one, namely the rAUC measure that has the potential of being more sensitive to agreement at the individual image level than traditional measures. This expectation was confirmed by the smaller absolute width of the 95% confidence interval for the difference in rAUC measures *vs* the other measures (Table 2). The sensitivity to differences in clinical relevance is important as it allows optimal selection of methodology for an observer performance study.

While the results of this study are inconclusive, we believe that it is important to pursue this issue further, and we hope at least to have provided some tools for those seeking to investigate this issue. The motivation for measuring clinical relevance is that clinical trials are very expensive, and laboratory ROC or FROC studies provide less expensive ways of determining which modality is superior. However, if attention is not paid to the clinical relevance of the measurements, then the results might be misleading. There has been some controversy about the clinical relevance of the FROC paradigm relative to the ROC paradigm [24–27], and it has been argued that ROC

may be more clinically relevant than FROC in certain lesion detection tasks. The present work gives a specific quantitative definition to the rather loosely used term "clinical relevance", and shows that, at least in the current nodule-detection study, FROC may be more clinically relevant than ROC.

The general tendency, in our experience, is to regard ROC studies as the "closest to clinical as one can get". After all, one is using the same images that the radiologists interpret in the clinical domain and the task is realistic. Those who believe that it is axiomatic that ROC is clinically relevant would question the need for the study we conducted, and indeed may find the results both surprising and disturbing. We believe the small observed correlations and agreement values (Table 2) between ROC/AFROC and clinical decisions are likely to be the result of a combination of factors. The laboratory study used ratings, while the clinical decisions were binary. The retrospective study was conducted using single-view (posteroanterior) chest radiographs only. The original interpreting radiologist had access to both posteroanterior and lateral views, and sometimes a lesion, which may be ambiguous in just the posteroanterior view, may become more obvious when both views are provided. As another example, a suspected nodule may reveal itself as a healed rib fracture on the CT, leading to a clinical judgement that the patient has no nodules although the clinical judgement of the chest radiograph alone would be suspicion of a nodule. Not only did the laboratory readers not get the

Table 5. This table is similar to Table 4 except that it applies to Reader 17 and the alternative free-response receiver operating characteristic (AFROC)-derived correctness measure

AFROC-derived correctness measure		Incorrect prospective outcome (C=0) no. of relevance FPs		Correct prospective outcome (C=1) no. of relevance TPs		Relevance FPF	Relevance TPF
Operating point	χ_k^0	Nodule free	Nodule present	Nodule free	Nodule present		
1	0.673		3 (a)		4 (b)	0.115	0.074
2	0.572		3 (c)		5 (d)	0.231	0.167
3	0.160	2 (e)		30 (f)		0.308	0.722
4	0.090				2 (g)	0.308	0.759
5	0.040		2 (h)			0.385	0.759
6	0.002				1 (i)	0.385	0.778
7	-0.201		1 (j)			0.423	0.778
8	-0.403	1 (k)		2 (l)		0.462	0.815
9	-0.416	1 (m)				0.500	0.815
10	-0.492		11 (n)		8 (o)	0.923	0.963
11	-0.657			2 (p)		0.923	0.981
12	-1.09	2 (q)				1	1

The corresponding relevance receiver operating characteristic plot is Figure 5b, specifically the plot labelled "AFROC derived".

{image numbers} (common rating): comments, if applicable

- (a) {2, 9, 13} (4)
- (b) {27, 34, 35, 37} (4)
- (c) {3, 6, 19} (3)
- (d) {24, 26, 30, 33, 39} (3)
- (e) {58, 59} (0)
- (f) {41, 42, 43, 44, 45, 48, 49, 50, 52, 53, 54, 55, 60, 61, 62, 63, 65, 66, 67, 68, 69, 70, 71, 72, 74, 75, 76, 77, 78, 79} (0)
- (g) {25, 32} (4): but the other nodule was not marked
- (h) {1, 8} (3): but the other nodule was not marked
- (i) {31} (2): but the other nodule was not marked
- (j) {15} (4): but 3 other nodules were not marked
- (k) {57} (1)
- (l) {51, 64} (1)
- (m) {47} (2)
- (n) {0, 4, 5, 7, 10, 11, 12, 14, 16, 17, 18} (0)
- (o) {20, 21, 22, 23, 28, 29, 36, 38} (0)
- (p) {40, 73} (3)
- (q) {46, 56} (4)

benefit of the lateral view, they did not see the prior radiology reports, the older radiographs and the CTs that the prospective readers may have seen. Ideally, the same readers who participate in the prospective interpretations should participate in a subsequent laboratory FROC study, and the fact that we had different sets of readers is another plausible reason for the low observed agreement. The selection of difficult cases could have also contributed to the observed low agreement.

The issue of optimal case selection for a rROC study requires further work. At the design phase of this study it was considered desirable to achieve equal splitting of the images (1) between lesion-free and lesion-present images and (2) between correct and incorrect prospective interpretations. The first constraint is needed for optimal precision in the measurements of AUC under the ROC curves, and it was rather easy to satisfy (the splitting was 40–40). The second criterion is necessary for optimal estimation of rAUC. Therefore, we aimed for equal numbers of incorrect and correct prospective interpretations. This dictated the choice of a rather difficult image set, as otherwise most of the prospective decisions would be correct. The average AUC is actually in the range of what is considered optimal [1, 2, 28] for modality discrimination ROC studies (*i.e.* not too easy, not too difficult). In our study there were 54 correct interpretations

split between 20 lesion-present and 34 lesion-free images, and 26 incorrect interpretations split between 20 lesion-present and 6 lesion-free images. Obtaining incorrect prospective interpretation nodule-free images turned out to be difficult. It is rare at our institution that a case is referred to CT for a suspected nodule but no nodule is found. Only six images fitted this definition, and this is a limitation of the study. To avoid biasing in favour of FROC, and because such cases are relatively rare at our clinic, we deliberately avoided including any "right for wrong reason" interpretations, *i.e.* images that were prospectively interpreted as nodule containing but in which the wrong location was identified.

When comparing two modalities FROC tends to have greater statistical power than ROC, and the statistical power advantage increases with greater numbers of lesions per image [15, 29, 30]. In our study the average number of lesions per abnormal image was only 1.225. There are several reasons for not including images with too many nodules in our investigation, as follows. (1) With numerous pulmonary nodules, accurately identifying each one of them on the chest CT becomes challenging and for selection of the gold standard truth for our laboratory study we were trying to eliminate this possibility of error. (2) When multiple nodules were reported on the prospective studies that we compared, the

prospective report tended to be more vague, often identifying the presence of multiple nodules yet often without identifying the location of each one of them. (3) It is known [31] that, because of "satisfaction of search", when one nodule is found, another is more likely to be overlooked. (4) In clinical reality it makes little difference whether the chest radiograph identifies one, two or three pulmonary nodules, and all such prospective interpretations would be judged as correct. However, in the laboratory, the readers would be penalised for not marking all the nodules, which would decrease the correlation between the laboratory and clinical figures of merit. Once an unexpected non-calcified nodule is found on a chest radiograph it is investigated, whether it is solitary or whether more than one is identified, as the main differential of a solitary as well as several pulmonary nodules is between granulomata and malignancy. Even though one may argue that a solitary nodule may represent a primary lung malignancy, as compared with multiple nodules in a known cancer patient that would be more likely to represent metastases, the point is that patients with either the solitary malignancy or the multiple nodules will be investigated with additional imaging studies. The timing of the follow-up, investigation and imaging modality to be used in the investigation depends on clinical suspicion, age of patient and clinical context, but a patient like that will have additional imaging regardless of the number of nodules identified.

Even though the *number* of reported nodules may not be important (as long as at least one is reported), the *locations* of the reported nodules have an important bearing on patient diagnosis. A cluster of nodules is likely to be an infection. Multiple nodules scattered somewhat randomly but favouring the lower lobes tend to be metastatic cancer. One expects that FROC is clinically more relevant than ROC in situations in which the location of abnormalities is in itself clinically relevant. The location of one or a few pulmonary nodules on a chest radiograph is clinically relevant. Normal radiological practice calls for description of the location of the suspected nodule in order to guide attention on the CT scan and on subsequent conventional radiographs. Ultimately, the location of the nodule is of crucial importance for biopsy and therapy. In some clinical situations, however, the specific location of an abnormality may be irrelevant. This is particularly true for diffuse or widespread manifestations of disease. If a patient has not 1 or 2 pulmonary nodules but 20, for example, the prospective interpretation is not likely to discuss each one. This will create a more ROC- than FROC-like clinical report, in which the interpreting radiologist is likely to give an estimate of the number of nodules and perhaps describe one or two of them in more detail.

The number of images was relatively small ($N=80$) owing to a study limitation in which large numbers of readers were available but each reader was only available for approximately 1 h. In this study, the ROC rating was not obtained directly, rather it was inferred from the FROC ratings using the highest rating assumption. There is an ongoing debate about the validity of this approach, as the presence of multiple lesions may lead to a higher "true" ROC rating than any given FROC rating. A refinement of the study would be to acquire the ROC

rating independently, or other methods of inferring the ROC rating could be used [33]. In view of these limitations, the current study should be regarded as a pilot that demonstrates proof of concept, namely a method of measuring the laboratory/clinical correlation and a preliminary application to ROC and FROC using a small data set.

The rROC agreement measures described here have been validated in a recent simulation study [32]. This study focused on the free-response approach to acquiring location-specific data. Other figures of merit for FROC analysis [33–36] and other location-specific approaches [37–39] are available that could be similarly studied. Since the rROC method is somewhat unconventional, the code used to generate the results in this paper is available on the authors' website (www.devchakraborty.com). The code is written in the open-source R-programming language [40], which is freely available and popular with statisticians.

Conclusion

A new type of study aimed at quantifying the clinical relevance of an observer performance paradigm is described. The new rAUC measure was consistent with the traditional correlation measures and had greater precision at measuring the difference in clinical relevance between the FROC and ROC paradigms.

Acknowledgments

DPC is grateful to Dr Tianhu Lei for proofing the document and to Dr Hong-Jun Yoon for programming assistance and help with manuscript preparation.

References

1. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978;8:283–98.
2. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest Radiol* 1989;24:234–45.
3. Kundel HL, Berbaum KS, Dorfman DD, Gur D, Metz CE, Swenson RG. Receiver operating characteristic analysis in medical imaging. ICRU Report 79. Bethesda, MD: International Commission on Radiation Units and Measurements; 2008.
4. Wagner RF, Metz CE, Campbell G. Assessment of medical imaging systems and computer aids: a tutorial review. *Acad Radiol* 2007;14:723–48.
5. Dodd LE, Wagner RF, Armato SG, McNitt-Gray MF, Beiden S, Chan HP, 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;11:462–75.
6. Egan JP, Greenburg GZ, Schulman AI. Operating characteristics, signal detectability and the method of free response. *J Acoust Soc Am* 1961;33:993–1007.
7. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH. A free-response approach to the measurement and characterization of radiographic-observer performance. *J Appl Photogr Eng* 1978;4:166–71.
8. Chakraborty DP. New developments in observer performance methodology in medical imaging. *Semin Nucl Med* 2011;41:401–18.

9. Chakraborty DP, Yoon HJ, Mello-Thoms C. Spatial localization accuracy of radiologists in free-response studies: inferring perceptual FROC curves from mark-rating data. *Acad Radiol* 2007;14:4–18.
10. Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty LA, Ganott MA, et al. The “laboratory” effect: comparing radiologists’ performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* 2008;249:47–53.
11. Gur D, Bandos AI, Fuhrman CR, Klym AH, King JL, Rockette HE. The prevalence effect in a laboratory environment: changing the confidence ratings. *Acad Radiol* 2007;14:49–53.
12. Samei E, Badano A, Chakraborty D, Compton K, Cornelius C, Corrigan K, et al. Assessment of display performance for medical imaging systems: Executive summary of AAPM TG18 report. *Med Phys* 2005;32:1205–25.
13. Chakraborty DP. Recent developments in free-response methodology. In: Samei E, Krupinski E, eds. *The handbook of medical image perception and techniques*. Cambridge, UK: Cambridge University Press; 2010. pp. 216–39.
14. Chakraborty DP, Yoon HJ. Operating characteristics predicted by models for diagnostic tasks involving lesion localization. *Med Phys* 2008;35:435–45.
15. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: modeling, analysis and validation. *Med Phys* 2004;31:2313–30.
16. Chakraborty DP, Winter LHL. Free-response methodology: alternate analysis and a new observer-performance experiment. *Radiology* 1990;174:873–81.
17. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* 1947;18:50–60.
18. Wilcoxon F. Individual comparison by ranking methods. *Biometrics* 1945;1:80–3.
19. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–45.
20. Efron B, Tibshirani RJ. An introduction to the bootstrap. In: Cox DR, Hinkley DV, Reid N. *Monographs on statistics and applied probability*. Boca Raton, FL: Chapman & Hall; 1993.
21. Hanley JA, Hajian-Tilaki KO. Sampling variability of nonparametric estimates of the areas under receiver operator characteristic curves: An update. *Acad Radiol* 1997;4:49–58.
22. Larsen RJ, Marx ML. *An introduction to mathematical statistics and its applications*, 3rd edn. Upper Saddle River, NJ: Prentice-Hall Inc.; 2001.
23. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37–46.
24. Gur D, Rockette HE. Performance assessment of diagnostic systems under the FROC paradigm: experimental, analytical, and results interpretation issues. *Acad Radiol* 2008;15:1312–15.
25. Gur D, Rockette HE. Performance assessments of diagnostic systems under the FROC paradigm: experimental, analytical, and results interpretation issues. *Acad Radiol* 2009;16:770–1.
26. Chakraborty DP. Counterpoint to “performance assessment of diagnostic systems under the FROC paradigm” by Gur and Rockette. *Acad Radiol* 2009;16:507–10.
27. Paquerault S, Samuelson FW, Petrick N, Myers KJ. Comparing signal-based and case-based methodologies for CAD assessment in a detection task. *Proc SPIE* 2008;6917:691708–9.
28. Green DM, Swets JA. *Signal detection theory and psychophysics*. New York, NY: John Wiley & Sons; 1966.
29. Chakraborty DP, Yoon HJ. JAFROC analysis revisited: figure-of-merit considerations for human observer studies. *Proc SPIE* 2009;7263:72630T.
30. Chakraborty DP. Validation and statistical power comparison of methods for analysing free-response observer performance studies. *Acad Radiol* 2008;15:1554–66.
31. Berbaum KS, Franken EA, Dorfman DD, Rooholamini SA, Kathol MH, Barloon TJ, et al. Satisfaction of search in diagnostic radiology. *Invest Radiol* 1990;25:133–40.
32. Chakraborty DP. 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. *Phys Med Biol* 2012;57:2873–904.
33. Song T, Bandos AI, Rockette HE, Gur D. On comparing methods for discriminating between actually negative and actually positive subjects with FROC type data. *Med Phys* 2008;35:1547–58.
34. Bandos AI, Rockette HE, Song T, Gur D. Area under the free-response ROC curve (FROC) and a related summary index. *Biometrics* 2008;65:247–56.
35. Samuelson FW, Petrick N. Comparing image detection algorithms using resampling. In: *Proceedings of the 2006 IEEE International Symposium on Biomedical Imaging: From Nano to Micro*; 6–9 April 2006; Arlington, VA. New York, NY: IEEE; 2006. pp. 1312–15.
36. Samuelson FW, Petrick N, Paquerault S. Advantages and examples of resampling for CAD evaluation. In: *Proceedings of the 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*; 30 March–2 April 2007; Chicago, IL. New York, NY: IEEE; 2007. pp. 492–5.
37. Starr SJ, Metz CE, Lusted LB, Goodenough DJ. Visual detection and localization of radiographic images. *Radiology* 1975;116:533–8.
38. Swenson RG. Unified measurement of observer performance in detecting and localizing target objects on images. *Med Phys* 1996;23:1709–25.
39. Obuchowski NA, Lieber ML, Powell KA. Data analysis for detection and localization of multiple abnormalities with application to mammography. *Acad Radiol* 2000;7:516–25.
40. R Development Core Team. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2011.