Published in final edited form as:

Semin Nucl Med. 2011 November ; 41(6): 401–418. doi:10.1053/j.semnuclmed.2011.07.001.

New Developments in Observer Performance Methodology in Medical Imaging

Dev P. Chakraborty, PhD

Department of Radiology, University of Pittsburgh, 3520 Forbes Ave, Suite 109, Pittsburgh, PA 15261, 412-605-1553 (phone), 412-605-1554 (fax)

Dev P. Chakraborty: dpc10@pitt.edu

Abstract

A common task in medical imaging is assessing whether a new imaging system, or a variant of an existing one, is an improvement over an existing imaging technology. Imaging systems are generally quite complex, consisting of several components – e.g., image acquisition hardware, image processing and display hardware and software, and image interpretation by radiologists—each of which can affect performance. While it may appear odd to include the radiologist as a “component” of the imaging chain, since the radiologist’s decision determines subsequent patient care, the effect of the human interpretation has to be included. Physical measurements like modulation transfer function, signal to noise ratio, etc., are useful for characterizing the non-human parts of the imaging chain under idealized and often unrealistic conditions, such as uniform background phantoms, target objects with sharp edges, etc. Measuring the effect on performance of the entire imaging chain, including the radiologist, and using real clinical images, requires different methods that fall under the rubric of observer performance methods or “ROC analysis”. The purpose of this paper is to review recent developments in this field, particularly with respect to the free-response method.

OVERVIEW OF THE ROC PARADIGM

Over three decades have elapsed since a paper titled “Basic Principles of ROC Analysis” appeared in this journal [1]. Several excellent reviews have since appeared [1,2,3,4,5,6,7]. This paper reviews recent developments in this field. Receiver operating characteristic (ROC) analysis is a method for measuring performance of an observer in a binary classification task. The “receiver” in “ROC” originates from early military (ca 1940s) applications where the objective was to detect enemy aircraft and the detector was a radio receiver [11]. Based on analysis of the reflections of the radar pulses, a decision needed to be made: enemy aircraft present or absent. In the medical imaging context the radiologist is shown images that could be from non-diseased or diseased patients, the radiologist is “blinded”, of course, to the true diagnosis, and the radiologist’s task is to diagnose the patient as “non-diseased” or “diseased”. It is assumed that the experimenter (i.e., the person conducting and analyzing the ROC study) knows the true disease status of each patient. If the radiologist responds “diseased” to a diseased patient image, that event is termed a *true positive* (TP), and if the radiologist responds “diseased” to a non-diseased patient image, that event is termed a *false positive* (FP). False negatives (FN) and true negatives (TN) are the

© 2011 Elsevier Inc. All rights reserved.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

complements of true positive and false positive events, respectively. True positives and true negatives are correct decisions and false positives and false negatives are incorrect decisions. *The fundamental idea in observer performance methodology is to compute a figure of merit (FOM) that rewards correct decisions and penalizes incorrect decisions.* When comparing two imaging systems (commonly referred to as *modalities* or *treatments*) one computes a FOM for each modality and performs a statistical test to determine if the difference between the two FOMs is sufficiently large compared to its statistical variability to be unlikely to be due to chance. The degree of “unlikely” is quantified by the *p-value*.¹

CONFIDENCE LEVEL AND DECISION THRESHOLD

The concept of a *confidence level* is widely used in observer performance. It is a random (i.e., modality, reader and case dependent) scalar variable, which we refer to as a *z-sample*, that summarizes information regarding disease presence, or absence, that the observer derives from the image. If the confidence level is high (or *z* is more positive) then the observer is more confident that the image is diseased. If the confidence level is low (or *z* is more negative) then the observer is more confident that the image is non-diseased. The confidence level depends on imaging system characteristics, patient characteristics (including disease prevalence, lesion contrast, and background clutter) and radiologist characteristics. A high quality image in which expected anatomic structures are readily visible gives more confidence in the presence/absence of disease than a poor quality image. An easily visible lesion on an uncluttered background gives greater confidence in presence of disease than a less visible one. If the population is known to be high risk, then, all other factors being the same, the confidence level in presence of disease will tend to be higher. The confidence level also depends on the expertise of the radiologist. An expert radiologist can extract more confirmative evidence regarding presence of disease from an image with a subtle lesion than a novice, and consequently *z* is expected to be higher and the expert is more likely to call the case as diseased. The confidence level concept also applies to actually non-diseased cases, but here the expert extracts more confirmative information regarding *absence* of lesions than the novice, and consequently *z* is lower (or more negative) and the expert is more likely to call the image normal. To summarize, the confidence level is a random variable that depends on the radiologist, the case and the imaging system, high values of which correspond to greater confidence in the presence of disease.

True positive fraction (TPF) is the fraction of diseased patients (cases) that are correctly diagnosed as diseased. False positive fraction (FPF) is the fraction of non-diseased cases that are incorrectly diagnosed as diseased. The complements of TPF and FPF are the false negative fraction (FNF) and true negative fraction (TNF) respectively. True positive fraction is also known as *sensitivity* and true negative fraction is known as *specificity*. A second concept is that of *threshold*, which we denote by the Greek symbol ζ (pronounced zeta). It is also a scalar quantity but unlike the confidence level, which is a random variable, ζ is to be thought of as a modality and radiologist dependent deterministic quantity, not a random variable. In one model of decision-making, the observer compares the *z*-sample to ζ , and if $z \geq \zeta$, the case is diagnosed diseased and otherwise the case is diagnose non-diseased. This implies that TPF and FPF will be functions of ζ , with the nature of the function being determined by the statistical distribution of the *z*'s for diseased and non-diseased cases, respectively.

¹Henceforth the term *case* will be used to denote possibly multiple images from a patient, e.g. multiple views of a breast, two breasts per patient, multiple modality images of the same patient, etc.

What factors affect the choice of ζ ? Clearly, by adopting a lower ζ the observer is more likely to call a case diseased and sensitivity, $TPF(\zeta)$, will increase but specificity, $1 - FPF(\zeta)$, will decrease, i.e., $FPF(\zeta)$ will increase. Increased sensitivity is desirable but since false positive decisions lead to unnecessary recalls and diagnostic workups, decreased specificity is not desirable. Therefore, the radiologist must balance the benefit of more correct disease detections versus the cost of incorrectly calling non-diseased cases as diseased. The choice of threshold represents a compromise that is based, among other factors, on the radiologist's knowledge of the patient population, specifically the a-priori probability of disease (disease prevalence), the costs and benefits of the different types of decisions, and interactions with the radiologist's peers (standard good practice). In a low prevalence situation, where most patients are actually normal, the compromise results in a shift towards higher ζ , i.e., a stricter threshold, because otherwise the observer would be calling too many false positives. The threshold is also affected by the cost and/or invasiveness of the follow-up that results from a positive diagnosis. If the follow-up is costly or invasive, e.g., risky surgery, the radiologist is expected to adopt a stricter criterion, i.e., higher ζ . On the other hand, if the follow-up involves more imaging and perhaps a less risky biopsy, the radiologist will tend to adopt a more lenient threshold, i.e., lower ζ . For obvious reasons the choice of threshold is affected by malpractice considerations.

THE ROC CURVE AND RELATED FIGURES OF MERIT

As the threshold ζ decreases $TPF(\zeta)$ increases and $FPF(\zeta)$ also increases (the dependence on ζ , shown here explicitly, is usually suppressed). The ROC curve, see dotted line in Fig. 1, is the plot of TPF along the ordinate vs. FPF along the abscissa. A pair of values (FPF , TPF) defines an *operating point* on the ROC curve. As the threshold ζ is decreased from a large value (e.g., positive infinity) the operating point moves up the ROC curve from (0,0) and approaches (1,1) as the threshold ζ approaches a small value (e.g., negative infinity). The area under the ROC curve, denoted AUC, is a widely used figure of merit in ROC analysis. It takes into account the correlation of sensitivity and specificity and is independent of threshold, the costs and benefits of the decisions, and disease prevalence. A perfect observer is capable of operating at (0,1) where all diseased cases are called diseased ($TPF = 1$) while no non-diseased cases are called diseased ($FPF = 0$). The observer can operate along the y-axis by adjusting the threshold used to declare cases diseased – raising the threshold would lower the operating point. For this observer $AUC = 1$. An observer operating at chance level would decide a-priori the fraction (f) of cases to call diseased, and since the observer is not using image information, $TPF = FPF = f$. By varying f such an observer could operate along the chance diagonal and $AUC = 0.5$. An observer with $AUC = 0$ is actually equivalent to a perfect observer with $AUC = 1$ if one reverses the interpretations of the observer (diseased to non-diseased and vice versa – suggesting the observer has the disease and non-diseased classes reversed).²

Other figures of merit have been proposed besides AUC. A number of partial areas measures have been suggested. One [12] is the area under the ROC curve above a specified value $TPF = TPF_0$, divided by $1 - TPF_0$. The partial area under the ROC curve to the left of a given $FPF = FPF_0$, or between a range of FPF values, has also been suggested as a figure of merit [13]. The true positive fraction at given FPF , or the FPF at given TPF , are also plausible figures of merit (in the 2nd case smaller values are desirable). The reason for seeking alternate figures of merit is the desire to attain greater clinical relevance. For example, for high sensitivity (and attendant low specificity) tests, it makes sense to emphasize the high sensitivity portion

²The ROC curve shown in Fig. 1 is a *population* ROC curve, which can only be realized by having an infinite number of normal and abnormal cases. Such a curve does not show sampling variability and the operating point lies exactly on the curve. With finite size data sets the observed operating point will exhibit sampling variability and will generally not fall exactly on the estimated ROC curve.

of the ROC curve, i.e., the portion above $TPF = TPF_0$ and extending to (1,1). All of these figures of merit can be calculated from the parametrically fitted ROC curve to be described later.

It can be shown [14] that AUC is the probability that a randomly chosen diseased case yields greater confidence in presence of disease than a randomly chosen non-diseased case, a property not shared by non-AUC measures. This implies that AUC also equals the probability that a randomly selected diseased case will be judged more likely to be diseased when paired (i.e., shown together) with a randomly chosen non-diseased case (termed the two-alternative forced choice paradigm). These interpretations of AUC do not depend on any distributional assumptions, but do depend on the assumption that all images are independent.

As an example of the utility of the AUC measure consider the study [15] conducted to evaluate the effectiveness of screening mammography by estimating the variability in radiologists' ability to detect breast cancer. Fifty accredited mammography centers were randomly sampled from across the United States. One hundred eight (108) radiologists from these centers gave blinded interpretation to the same set of 79 randomly selected screening mammograms. Ground truth for these women had been established either by biopsy or by 2-year follow-up. The observed range of sensitivity (TPF) was at least 40% and the range of FPF was at least 45%, Fig. 2. [Mammography equipment physical measurements like dose are typically made to better than a few percent, which is dwarfed by the magnitude of the radiologist variability.] The study shows that a large part of the variability in sensitivity and specificity is due to the radiologists' variable thresholds for reporting disease. Once this is accounted for the variability of the AUC measure, representing the intrinsic variability in diagnostic abilities of the mammographers is smaller (11%).

PLOTTING ROC OPERATING POINTS

A study in which the observer is asked to classify each case as “non-diseased” or “diseased” yields a single operating point and requires the observer to maintain a single threshold ζ . The abscissa equals the fraction of non-diseased cases that are classified as diseased and the ordinate equals the fraction of diseased cases that are classified as diseased. Multiple operating points could be determined by repeating the study after inducing the observer to change the threshold ζ . This can be done for example by manipulating the rewards or costs to individual decisions, termed the “Yes/No” method. In practice, it is more efficient to ask the observer to simultaneously maintain several ordered thresholds (e.g., $\zeta_1 > \zeta_2 > \zeta_3 > \zeta_4$) and rate each case for confidence in presence of disease. The rating is an ordinal number representing a binned confidence level. For example, one could use a 5-point rating scale with higher numbers representing greater confidence in presence of disease, e.g., 1 = definitely disease free, 2 = probably disease free, 3 = ambivalent, 4 = probably diseased and 5 = definitely diseased. This would require the observer to simultaneously maintain 4 ordered thresholds $\zeta_1, \zeta_2, \zeta_3, \zeta_4$. The observer would rate a case as “5” if the z-sample exceeded ζ_4 , as “4” if the z-sample was between ζ_3 and ζ_4 , as “3” if the z-sample was between ζ_2 and ζ_3 , as “2” if the z-sample was between ζ_1 and ζ_2 , and as “1” if the z-sample was below ζ_1 . Table 1 shows a hypothetical 5-rating ROC dataset for 60 non-diseased cases and 50 diseased cases. It is evident that the diseased cases tended to get higher ratings, i.e., 4s and 5s, while the non-diseased cases tended to get lower ratings, i.e., 1s and 2s – a typical structure of an ROC dataset. This data yields four operating points, which are obtained by cumulating the numbers in the bins. For example, the lowest operating point, corresponding to the strictest criterion, is obtained by cumulating the counts with ratings ≥ 5 . This yields $FPF_5 = FPF(\zeta_5) = 1/60 = 0.017$ and $TPF_5 = TPF(\zeta_5) = 22/50 = 0.44$. The next operating point is obtained by cumulating counts with ratings ≥ 4 . This yields $FPF_4 = (1+2)/60 = 0.05$

and $TPF_4 = (22+12)/50 = 0.68$, and so on. The operating points are shown in the last two rows of Table 1. If one cumulates all the counts then the trivial operating point (1, 1) is reached (trivial in the sense that every ROC data set will give this point when all the bins are cumulated). If no counts are cumulated the (trivial) operating point is (0,0). An ROC study employing an R-point rating scale yields, in general, R-1 non-trivial operating points. Fewer than R-1 operating points may be obtained if some bins contain zeroes. [R = 2 corresponds to a Yes/No ROC, which yields one operating point.]

The confidence level scale used in Table 1, or for that matter any ROC rating scale, has the symmetry property that the extreme ratings, the “1s” and “5s”, represent high confidence decisions. The “5s” represent high confidence in presence of lesions and the “1s” represent high confidence in the absence of lesions. This property is not true for the free-response rating scale to be described later.

FITTING AN ROC CURVE TO RATINGS DATA

For a given modality and reader, how does one objectively fit a smooth curve to the operating points and calculate the area AUC under the smooth curve? The binormal model [16,17] involves two Gaussian distributions. One distribution models the statistical variability of the z-samples from non-diseased cases and the other models the statistical variability of the z-samples from diseased cases. Let us denote by $N(\mu, \sigma^2)$ the Gaussian distribution with mean μ and variance σ^2 . The symbol “ $\sim$ ” is to be interpreted as “is distributed as”. According to the binormal model, $z_n \sim N(0, 1)$ and $z_s \sim N(\mu, \sigma^2)$, where the subscripts n and s refer to non-diseased (n=noise) and diseased (s=signal) cases, respectively. The binormal parameters μ and σ , the area under the fitted ROC curve (denoted A_z since it is estimated using the binormal model assumption) and the thresholds $\zeta_1, \dots, \zeta_{R-1}$, can be estimated from the rating data by a procedure known as maximum likelihood. The method involves iteratively searching for parameter values until the likelihood of the data, given the parameters, is maximized. Sometimes the algorithm may fail to converge, or it may predict fits consisting of vertical and horizontal lines, and the corresponding datasets, termed *degenerate*, are often characterized by few interior data points, or points that do not span the ROC space adequately. Different methods have been developed for dealing with this type of problem [3,18,19,20]. An important assumption in the estimation procedure is that *the ratings are statistically independent*. If the images are from different patients, who are not biologically related, then independence is a reasonable assumption. However, ratings from multiple images originating from the same patient, e.g., mammograms of breasts of the same patient, or ratings from different regions of interest from the same image, are correlated. *Neglecting the independence requirement is a common error made by those new to this field.*

A related figure of merit is d' , the *detectability index*, which is the separation of two unit variance normal distributions that yields the same A_z as the binormal model [21]. Two such distributions will actually yield a different ROC curve than that predicted by the binormal model, but the areas under the two curves will be the same. The detectability index could be interpreted as a perceptual signal-to-noise ratio. For simple tasks it is expected to scale linearly with lesion contrast (by simple tasks we mean tasks where it is relatively easy to define noise and contrast, for example a circular disk superposed on random Gaussian noise). AUC and A_z do not have this property as they are constrained to be less than or equal to unity. For example, doubling contrast doubles d' but A_z will not double. [AUC vs. A_z : the distinction in terminology is meant to emphasize that A_z is the *binormal model estimate* of the *true* area AUC under the ROC curve.]

Example

The data shown in Table 1 was fitted using a web-based calculator for ROC curves [22]. The calculator yielded $a = 1.3204$, $b = 0.6075$, $\zeta_1 = 0.0077$, $\zeta_2 = 0.8963$, $\zeta_3 = 1.5157$ and $\zeta_4 = 2.3967$. The reported parameters a and b are related to μ and σ by $\mu = a/b$ and $\sigma = 1/b$. These yield $\mu = 2.17$ and $\sigma = 1.65$ and the predicted ROC curve is shown in Fig. 3 by the solid line. The software also estimates the area under the ROC curve $AUC = A_z = 0.87$ and its standard deviation $\sigma_{A_z} = 0.0378$. Assuming that AUC is normally distributed, an approximate 95% confidence interval for A_z is $0.87 \pm 1.96 (0.0378)$, i.e., (0.794, 0.946) (represented by the dashed and dotted curves, respectively) and the detectability parameter $d' = 1.59$. The straight line in Fig. 3 is called the chance diagonal, which represents chance-level performance ($AUC = 0.5$).

Other curve fitting methods—It may be noticed that the example data in Table 1 yields the ROC curves shown in Fig. 3 that have $b < 1$, i.e., $\sigma > 1$. This is typical of clinical ROC datasets, which usually imply that the diseased distribution of z is wider than the non-diseased distribution of z . The predicted ROC curve then has a “hook” just below (1,1) where it crosses the chance diagonal, implying worse than chance level performance, and it subsequently hooks upward to approach (1,1) with infinite slope. This is most evident for the lower 95% confidence interval band shown in Fig. 3 (the dotted line). The hook is a consequence of allowing $b \neq 1$. Alternate ROC models have been suggested [18,19,20,23] that do now allow the ROC curve to fall below the chance diagonal. ROC curves whose slope decreases monotonically as the operating point moves up the curve, and which cannot cross the chance diagonal, are termed “proper” ROC curves.

Non-parametric or empirical area under the curve—One may ask, is it even necessary to fit the data to a parametric model? If one connects adjacent operating points – including the “trivial” ones at the origin and (1,1) – with straight lines, this defines the trapezoidal ROC, the area under which is a non-parametric estimate of AUC (since normality assumptions are not made). This is favored by some statisticians who distrust normality assumptions. Another point in its favor is that the non-parametric estimate always exists. There are no problems with degenerate datasets where the maximum likelihood algorithm fails to converge. However, this can lead to a false sense of security. The trapezoidal area is quite sensitive to the location of the uppermost (or lowermost) operating point [24]. For example, if the uppermost point is well below (1,1) then the area from this point to (1,1) dominates AUC. A study has been conducted in which different non-Gaussian distributions (binomial, Poisson and gamma type distributions) were used to generate ROC data, which was then fitted by the binormal model [25]. It was shown that the deviation of the binormal fit from the true form was of no practical significance since it was obscured by the much larger variation attributable to the sampling of cases. Even if a large sampling of cases was available, the small numbers of rating categories that are used in human observer ROC studies do not permit seemingly very different models to be distinguished from each other. It can be appreciated that the underlying distributions can be non-Gaussian, but as long as they can be transformed to Gaussians by a strictly monotonic increasing transformation, the ROC curve is unchanged. This is because the ROC is determined by the ordering of the ratings, and any strictly monotonic increasing transformation of the ratings leaves their relative ordering unchanged. For this reason the parametric ROC models are often termed semi-parametric approaches. [This argument would not apply to the evaluation of computer aided detection algorithms, where the ratings are inherently floating point numbers, and if a large number of cases are available, closely spaced operating points defining the non-parametric ROC are available that extend from the origin to the upper right corner. In this situation it makes much sense to use the trapezoidal area under the ROC curve. A similar argument applies to any diagnostic test that produces a continuous output.]

Choosing a rating scale for ROC studies—For a 5-rating ROC study, as shown in Table 1, it is not possible to unambiguously calculate the operating point of the observer in the binary task of discriminating between non-diseased and diseased cases. One possibility would be to use the 3 and above ratings to define the operating point, but one might also have chosen 2 and above. A better strategy is to obtain the rating via two questions [26]. The 1st question is “is the case abnormal”? The binary (Yes/No) response to this question allows unambiguous calculation of the operating point. The 2nd question is: “what is your confidence in your first decision” and allow three responses, namely Low, Medium and High. The two question approach is equivalent to a 6-point rating scale. Some experts advocate continuous or quasi-continuous ratings, e.g., 0 to 100, in increments of one [27,28,29]. Others have argued that given too many ratings levels (e.g., 101) an observer essentially bins the data into a much smaller number of bins (e.g., 0, 20, 40, 60, 80, 100) and adds a zero-mean noise term to appear to be “spreading out the ratings” [30].

It is desirable to be able to relate the rating scale to the radiologists’ daily practice. This assures greater consistency –the maximum likelihood fitting algorithms assume that the thresholds are held constant for the duration of the ROC study. Depending on the clinical task, a natural rating scale may already exist. For example, the American College of Radiology developed the Breast Imaging Reporting and Data System (BI-RADS) in 1992 to standardize mammography reporting [31]. There are six assessment categories: category 0 indicates need for additional imaging; category 1 is a negative interpretation; category 2 is a benign finding; category 3 is probably benign, with short-interval follow-up suggested; category 4 is a suspicious abnormality for which biopsy should be considered; category 5 is highly suggestive of malignancy and appropriate action should be taken. The 4th edition of the BI-RADS manual [32] divides category 4 into three subcategories 4A, 4B and 4C and adds category 6 for a proven malignancy. Apart from categories 0 and 2, the categories form an ordered set with higher categories representing greater confidence in presence of cancer.

A large study has been reported in which BI-RADS category data was acquired for 469,512 screening mammograms from 308,634 women interpreted by 124 radiologists [33]. The ROC analyses used the assessments codes ordered as follows: 1, 2, 3 with no immediate work-up; 3 with immediate work-up; and 0, 4, and 5 reflecting increasing likelihood of cancer. It was found that the probability of cancer increased monotonically with the proposed ordering, see Table 2.

The use of the BI-RADS ratings shown in Table 2 has been criticized [34] in an Editorial titled “BI-RADS Data Should Not Be Used to Estimate ROC Curves”. They claim that the Barlow et al method argument confuses cancer yield with diagnostic confidence. They also state that BI-RADS categories 1 and 2 should not be separate entries of the confidence scale, because both indicate no suspicion for cancer.

The author agrees with the suggested ordering of the 2s as more likely to have cancer than the 1s. A 2-category means the radiologist found something to report, and the location of the finding becomes part of the clinical report. Even if they believe it is definitely normal, there is a finite probability that a category 2 finding will eventually turn out to be cancer. In contrast there are no findings associated with a 1-category report.

The problem with “where to put the 0s” arises only when one tries to analyze clinical BI-RADS data using the ROC method. In a laboratory study the radiologist would not be given the 0-category option. In analyzing a clinical study it is incumbent on the study designer to justify the choice of the rating scale adopted. Showing that the proposed ordering agrees with the probability of cancer is a valid justification – and in the author’s opinion, given the

very large sample size this was accomplished rather convincingly in the Barlow et al study, see Table 2. Smaller studies one would have difficulty matching this level of precision.

ROC DATA ANALYSIS

We have seen how to estimate the area under the ROC curve for a single modality and reader. If there are two modalities, then one could determine AUC for each modality and the modality with the greater AUC would be deemed superior. That is the general idea behind ROC analysis but a number of details need to be addressed. Does one use the same cases to assess the two modalities? If not then modality A could get an easier dataset, meaning easier cases were imaged with modality A than with modality B, which would bias the comparison in favor of modality A. As an example, a case set with predominantly fatty breasts would be easier to interpret compared to one with predominantly dense breasts, since any lesions that could be present would be easier to detect against the relatively uniform background of a fatty breast than against the cluttered background of a dense breast. A related advantage of matching the cases is that any modality difference would be easier to detect since it would not be obscured by differences between the two image sets. Sometimes it may not be possible to image the same cases in both modalities (with x-ray modalities additional radiation dose would be incurred, which raises ethical issues), in which case one has to *match* the cases (by sex, age, and other patient demographics) to minimize bias. The next question is does one need more than one reader? Readers are variable and could have different degrees of expertise in the modalities, which would introduce bias favoring the modality interpreted by the more expert reader. For this reason, it is necessary to have several representative readers interpret the same set of cases – reader matching, which also makes it easier to detect a modality difference. Again, it may not be possible to use the same readers, for example when expertise is limited to one modality, in which case matching should be done using other factors (e.g., years of experience, board-certification, number of cases read in the past year, etc) to minimize bias. A dataset obtained by having the same readers interpreting a common set of cases in all modalities is termed fully-crossed multiple-reader multiple-case (MRMC).

At the conclusion of a MRMC data acquisition one has an estimate of the area A_{ij} (we abbreviate AUC to A) under the ROC curve for each modality and reader combination, where i is the modality index and j the reader index. For example, A_{23} denotes the area under the ROC curve for reader 3 in modality 2. Consider the reader-averaged figure of merit for modality i denoted by $\overline{A}_i$ (in general the bar denotes the average over the index or indices replaced by the dot symbol or symbols). Denote by ΔA the observed difference of the reader averaged figure of merit between a pair of modalities, i.e., $\Delta A = \overline{A}_2 - \overline{A}_1$, termed the *observed effect size*. It might be tempting to declare the modality with the greatest average area as superior, but this would not have accounted for the variability of the estimates. An observed effect size of 0.10 may not be convincing (significant) if its variability is large, and by the same token, if the variability is small, an observed effect size of 0.02 may be highly significant. What do we mean by variability? If the study were repeated with different readers and cases, but still following the MRMC design, a different value would be observed for the observed effect size³. One can imagine repeating the MRMC study many times, each time with different readers and cases, computing the differences ΔA_r , where r is a replication index, and determining the variance of these numbers, denoted $\sigma_{\Delta A}^2$. It is not practical – or necessary - to repeat studies as just described to

³Due to reader inconsistency a different value would most likely be obtained even if the study were repeated with the same readers and cases.

determine $\sigma_{\Delta A}^2$. It can be estimated from the data using statistical techniques to be described shortly, and a p-value is used to summarize how large the observed effect size is relative to $\sigma_{\Delta A}$.

The p-value is the probability that an effect size at least as large in magnitude as the observed value could occur given that the true difference between the modalities was zero, which is termed the *null hypothesis* (NH) condition⁴. The qualifier “at least as large” is needed, as otherwise one expects *zero* probability of getting *exactly* the observed effect size. For example, if $p = 0.0129$, that means that the probability is 1.29% that chance would account for the observed effect size, or larger, given no difference between the modalities. Since this is a rather small probability, one can with some justification claim that the modalities are indeed different. By convention, the threshold probability for declaring significance is chosen to be 5% (this quantity – specified prior to the study – is formally termed α , the Type I error probability of the test, or the *size* of the test). If $p \leq \alpha$ one has shown a statistically significant difference at the α level; if $p > \alpha$ then one has not shown a significant difference. If $p > \alpha$ one cannot conclude that the true effect size is zero, i.e., it is incorrect to then claim that “there is no difference between the modalities”. The inability to reject the null hypothesis does not imply that the null hypothesis is true. The fallacy of such a conclusion becomes evident if one recognizes that the probability is infinitesimal that the true modality effect is *exactly* zero. Likewise, claiming that the p-value “approaches significance” (e.g., $p = 0.051$) amounts to drawing a statistically unfounded conclusion from the data. The rules of null hypothesis testing allow a strict binary conclusion: the null hypothesis is either rejected (and the alternative hypothesis is accepted) or the null hypothesis is not rejected. Note that p is a random (data dependent) variable while α is a fixed constant.

The Dorfman-Berbaum-Metz (DBM) MRMC significance testing method

For a given modality, the ratings assigned by the reader to the cases yield a single number, namely the area AUC (=A) under the ROC curve. For example, the ratings of 120 cases (i.e., 120 numbers) yield a single A value, which represents a 120-fold reduction in the “raw” numbers available for further analysis. The jackknife procedure is a method for teasing out the individual contributions of the cases, albeit on a different quantity (not A) termed a *pseudovalue*, defined by [35]

$$Y_{ijk} = KA_{ij} - (K - 1)A_{ij(k)}, \quad \text{Eqn. 6}$$

Y_{ijk} is the pseudovalue corresponding to modality i ($i = 1, 2, \dots, I$), reader j ($j = 1, 2, \dots, J$) and case k ($k = 1, 2, \dots, K$), where I is the number of modalities, J is the number of readers and K is the number of cases. A_{ij} is the area under the ROC curve for modality i and reader j , when all cases are included in the analysis. $A_{ij(k)}$ is the area under the ROC curve for modality i and reader j when case k is *excluded* from the analysis. The jackknife procedure is repeated for all modalities, readers and cases, yielding a 3-dimensional matrix of IJK pseudovalues.

The pseudovalues have the physical interpretation of the “correctness” of decisions on the cases. For example, if diseased case k receives a high rating (e.g., 5 = “definitely diseased”,

⁴This is the NH condition when the goal of the experiment is to show that the performances of the readers using the two modalities are different. Sometimes, one may be interested in showing that the two modalities are substantially equivalent, in which case this would not be the NH. In that case, one may set as NH $A_2 - A_1 = \delta$, where δ is a pre-specified (and clinically justified) constant, and as alternative hypothesis $A_2 - A_1 = 0$.

following the 5-rating example in Table 1), then removing the case will leave a data set with one less high-confidence correct decision case, which will cause $A_{ij(k)}$ to decrease because the remaining data set is, perhaps ever so slightly, “more difficult”, and since it appears with a negative sign, Y_{ijk} will increase. Likewise, removing a non-diseased case with a low rating (e.g., “1” = definitely disease free) will leave a data set with one fewer high-confidence correct decision case, and $A_{ij(k)}$ will decrease and the pseudo-value will increase. Conversely if a diseased case with a low rating (e.g., 1 = definitely disease free) is removed then the remaining data set has one fewer high-confidence incorrect decision case, and $A_{ij(k)}$ will increase because the remaining dataset is, perhaps ever so slightly, more easy, and the pseudo-value will decrease, and likewise if a non-diseased case with a high rating is removed. In other words, high confidence correct decisions, true positives or true negatives, will yield higher than average pseudo-values and high confidence incorrect decisions, false negatives or false positives, will yield lower than average pseudo-values.

Having recovered the case dependence, what does one do with it? Needed are a model for the pseudo-values and a method for estimating the parameters of the model. This allows one to determine if the modalities are significantly different, or formally, to calculate a p-value for rejecting the null hypothesis. The model needs to account for the variability and correlations resulting from the reader and case matching.

In the Dorfman, Berbaum and Metz (DBM) method [36] the ijk^{th} pseudo-value is modeled as a constant +
 the i^{th} modality contribution +
 the j^{th} reader contribution +
 the k^{th} case contribution +
 the ij^{th} modality-reader contribution +
 the ik^{th} modality-case contribution +
 the jk^{th} reader-case contribution +
 the ijk^{th} pure error contribution.

All terms except the constant and modality contribution are modeled as random effects, i.e., samples from zero-mean Gaussian distributions with specified variances, termed *variance components*. For example, a variance denoted by σ_R^2 corresponding to the reader factor models the variability of the readers, which arises because readers vary in their expertise. Each reader is represented by a random sample from the reader distribution. For example, if $\sigma_R^2=0.09$ (i.e., $\sigma = 0.3$) then an exceptional reader might contribute 0.6 units to the pseudo-value (i.e., two standard deviations above zero, which corresponds to the upper 2.5% percentile of the Gaussian distribution) while a lower than average performing reader might contribute -0.3 units to the pseudo-value. The variance component σ_C^2 corresponding to the case factor models the variability of the cases arising from the fact that some cases are easier to interpret than others. The variance component $\sigma_{\mu R}^2$ (μ denotes modality) corresponding to the modality-reader factor, allows for the possibility a reader's contribution to the pseudo-value is modality dependent (equivalently, the reader variability depends on the modality). The number of such contributions equals the number of modalities multiplied by the number of readers (IJ). Likewise the variance component $\sigma_{\mu C}^2$ corresponding to the modality-case factor allows for the possibility that the cases may be more variable in one modality than another, with IK contributions to the pseudo-values. A reader-case variance component σ_{RC}^2 allows for reader variability to vary between cases, with JK contributions to

the pseudovalues. Finally there is an error term, with variance component σ_e^2 , which models the residual variability, with IJK contributions to the pseudovalues. Note that modality is a fixed factor since one is not interested in a random sample from a population of modalities rather interest is in the specific modalities studied.

The magnitudes of the variance components (i.e., the variances of the Gaussian distributions) can be estimated using analysis of variance (ANOVA), which is a statistical technique designed for such additive linear models. The main interest is in determining whether the modalities are different. In terms of the pseudovalue model the equivalent question is whether the two levels of the modality factor are sufficiently separated (the difference is the modality effect) relative to the variability of the difference, to be unlikely to be due to chance. The procedure involves calculating a test statistic, namely a Fisher F-statistic. The p-value is the probability, under the null hypothesis, that the F-statistic is larger than the observed value. Consequently, a large value of the observed F-statistic would lead to a small p-value.

Statistical power and sample size estimation

Suppose one conducts a pilot study to investigate the effect of a new image processing algorithm. Prior to the study one has little or no idea of what effect size to expect, because no one has investigated the algorithm before and hence the need for the pilot study. At the conclusion of the analysis one has an observed effect size, which is the best estimate yet of what to expect from the new image processing algorithm. If the p-value from the study is greater than α , then the modality effect is not significant at the α -level. If the observed effect size is encouraging, then the next step is to perform statistical power analysis, i.e., estimate the numbers of readers and cases to achieve the desired statistical power. Statistical power is the probability of rejecting the null hypothesis when the alternative hypothesis (AH) $\Delta A_{true} \neq 0$ is true. The difference ΔA_{true} under the AH is referred to as the *true* effect size. In terms of the ANOVA analysis of the pseudovalues, the square of this quantity is proportional to a quantity known as the non-centrality parameter. Statistical power is the probability that a sample from a non-central F-distribution with this non-centrality parameter is greater than a critical value F_c , where the critical value is defined such that the corresponding probability for a central F-distribution (i.e., with non-centrality parameter equal to zero), is α .⁵ Note that the p-value is determined by the sampling distribution of the test statistic under the null hypothesis while statistical power is determined by the sampling distribution of the test statistic under the alternative hypothesis.

Since one is interested in the difference of modalities, one may suspect, as is true, that only the variance components that are modality dependent influence sample size. These are modality-case, modality-reader and the pure error term (which has modality dependence). For example, the j^{th} reader contribution could be large, due to large reader variance, but since it affects all modalities equally, there is no effect on the difference between the modalities, so reader variability in itself does not necessarily reduce statistical power, but the modality-reader term does. The variability of the reader-averaged observed effect size decreases as the number of readers and/or number of cases increases.⁶ This means that by

⁵The F-distribution depends on two quantities termed numerator and denominator degrees of freedom. The numerator degree of freedom (ndf) is always the number of modalities being compared minus 1 (i.e., $\text{ndf} = 1$ if there are two modalities). The denominator degree of freedom (ddf) depends in a complex manner on the variance components.

⁶This is a standard principle in statistics: the average over several readers has less variability than the variability of individual readers. This assumes that the readers are independent. Averaging multiple interpretations by the same reader will only attenuate the component of variability that causes reader inconsistency, or within-reader variability, which is distinct from, and generally smaller than, between-reader variability.

using sufficiently large numbers of readers and cases this quantity can be made sufficiently small relative to the true effect size, to achieve the desired statistical power.

Sample-size methodology estimates, from a pilot study with a relatively few number of readers and cases, the three relevant variance components, and uses these values to predict the number of readers and cases needed to achieve a given statistical power for the planned or pivotal study. Sample-size estimation is an important consideration at the planning stage of a study. An underpowered study (too few readers and/or cases) raises ethical issues since study patients are subjected to unnecessary imaging procedures for a study of questionable statistical strength. Conversely, an excessively overpowered study subjects unnecessarily large numbers of patients to imaging procedures and raises the cost of the study. Studies are typically designed for 80% power. The true effect size is unknown; indeed, if one knew it there would be no need to conduct an ROC study! Sample-size estimation involves making a decision regarding the *anticipated* or conjectured true effect size. To quote [37] “any calculation of power amounts to specification of the anticipated effect size”. Increasing ΔA_{true} will increase statistical power but may represent an unrealistic expectation of the true difference between the modalities. On the other hand, an unduly small ΔA_{true} may be clinically irrelevant besides requiring a very large sample-size to achieve desired power. The choice of clinically relevant effect size requires clinical input.

Two well-known sample-size estimation procedures for MRMC are the Obuchowski-Rockette [38,39,40] and the Hillis-Berbaum (HB) methods [41]. It has been shown that under some conditions the two are equivalent. The HB method works in conjunction with the Dorfman-Berbaum-Metz (DBM) method of analyzing MRMC data. It uses the variance components estimated by the DBM-MRMC method to predict the sample-size. DBM-MRMC analysis software is available from http://www-radiology.uchicago.edu/cgi-bin/roc_software.cgi and from <http://perception.radiology.uiowa.edu>. The HB method has been implemented in SAS software available on the University of Iowa website. A Windows and Macintosh implementation of the HB method is available on the author’s website (www.devchakraborty.com).

LOCATION SPECIFIC PARADIGMS

In the ROC paradigm, the reader assigns a single rating to each image and information regarding the location of suspected lesions is not utilized. When the disease manifests as diffuse abnormal characteristics in an organ then the single rating captures the available information and the ROC paradigm is appropriate. For example, interstitial lung disease, where the lung parenchyma exhibits an abnormal texture, can be characterized by a single rating and localization information is implicit, and stating “I found interstitial lung disease in the lung” is superfluous. However, when the disease is manifested by the presence of localized lesions within an organ, such as malignant nodules in the lung, not collecting location information can introduce ambiguity since the experimenter cannot rule out that the reader missed the lesion and mistook a suspicious non-diseased region for a lesion. In such tasks, marking the correct location informs the experimenter that the reader actually saw the lesion and, this leads to more precise scoring and a figure of merit that is better at rewarding correct decisions and penalizing incorrect decisions. Moreover, the location could be a clinically relevant variable, especially if it guides subsequent therapeutic interventions (e.g., biopsy or surgery). There are three location specific paradigms, the location ROC (LROC), the region of interest (ROI) and the free-response ROC (FROC) paradigms. While these differ in how the data is collected, the analysis of the data follows the same general principles outlined previously for ROC data – namely, one calculates a figure of merit, next an observed effect size (intra-modality difference in the figure of merit, averaged over

readers), and estimates its variability using standard statistical techniques. This leads to the p-value for rejecting the null hypothesis that the true effect size is zero.

Location ROC (LROC) paradigm

The location ROC (LROC) paradigm is concerned with lesion detection and localization in images that contain at most one target. For a given image, the radiologist provides a rating that the lesion is present and also marks the most suspicious location [42]. If the mark in a diseased image is sufficiently close to the actual lesion then the localization is classified as correct and otherwise, the localization is classified as incorrect. How to set acceptance radius is described later in connection with the free-response paradigm. The localizations on the non-diseased images are ignored for purposes of scoring and thus have no impact on observer performance.

Recall that the ROC curve plots $\text{TPF}(\zeta)$ along the y-axis and $\text{FPF}(\zeta)$ along the x-axis, with ζ the observer's decision threshold. The LROC curve differs from this in that the y-coordinate is instead the fraction of diseased images that have been correctly diagnosed as diseased and had the lesion correctly localized (CL). We denote this fraction as $\text{TPF}_{\text{CL}}(\zeta)$ while noting that the role of the decision threshold in determining true positives for LROC analysis is the same as for ROC analysis. When computed from a sample set of cases, $\text{TPF}_{\text{CL}}(\zeta)$ is an estimate of the joint probability that a diseased image will be both correctly diagnosed and localized.

As one figure of merit for LROC performance, the area under the LROC curve (A_L) takes values in the range $[0,1]$. The limiting value of $\text{TPF}_{\text{CL}}(\zeta)$ as $\zeta \rightarrow \infty$ is 0, so all LROC curves must contain the left-hand endpoint $(0,0)$. However, the LROC curve is not constrained to reach the point $(1,1)$ as $\zeta \rightarrow -\infty$. An LROC study will generally have some finite fraction of diseased images with incorrect localizations, even at the lowest threshold, and this constrains the ordinate for the uppermost endpoint of the LROC curve, given by the limit of $\text{TPF}_{\text{CL}}(\zeta)$ as $\zeta \rightarrow -\infty$, to less than one. The limit, referred to as the *probability of correct localization* (PCL), represents another figure of merit for LROC studies. PCL is an estimate of the probability that the lesion in a diseased image will be correctly localized, regardless of confidence level.

A latent-variable model [43] analysis, analogous to the binormal model for ROC curve fitting, is also available. In this model, the observer's rating and localization are based on the maximum of a perception measurement X made at all suspicious locations in an image. Suppose X_0 is the maximum of X over the non-lesion locations and let X_1 be the value of X at a lesion location. It is assumed that the presence or absence of a lesion does not affect X_0 . For a diseased image to be correctly diagnosed and localized requires that $X_1 > X_0$ and $X_1 > \zeta$. The data analysis involves assuming normal distributions for X_0 and X_1 . The model fits both ROC and LROC curves and the fits have attractive monotonicity properties insofar as the ROC curves are always "proper". The method requires that observers provide a rating and localization for each image. A subsequent modification [44] allows for incomplete localization data. The software implementing this modified fitting method for data from a single reader and one modality is available for download from www.philipfjudy.com/Home/lroc-software.

Data from an LROC study can also be analyzed non-parametrically. An LROC curve is generated directly from the rating data or from rating distributions created using kernel density estimation to smooth the data. Following up on previous work [45] on nonparametric ROC analysis, formulas for estimating A_L and its standard deviation have been derived [46]. To the best of the author's knowledge, stand-alone MRMC versions of either software are currently unavailable.

Region of interest (ROI) paradigm

In the region of interest (ROI) paradigm, the experimenter divides each image into Q (for quadrants) regions of interest [47]. The ROIs are selected according to clinical considerations but the value of Q needs to be the same for all images. For example, a chest image could be divided into a hilar region, left upper lobe, right upper lobe, etc. The radiologist's task is to rate each ROI for presence of lesion in that ROI. Therefore, the data consists of Q ratings per image. It is assumed that the truth is known for each ROI: it either contains or does not contain a lesion and lesions that overlap ROIs are not allowed. The figure of merit is the non-parametric area under the ROC curve calculated over all ROIs. The analysis takes correlations of the ratings of the ROIs on the same image into account. This type of data is often referred to as "clustered" by statisticians, in the sense that each patient contributes a cluster of values (the ratings of the ROIs), and not just a single value. The ratings of the items within a cluster can be correlated. Software for this method of data collection is available on <http://www.bio.ri.ccf.org/html/rocanalysis.html>.

FREE-RESPONSE (FROC) PARADIGM

In LROC the reader must localize something, even on an image which appears definitely normal, while in ROI the reader must rate each region of interest delineated by the investigator. In the free-response paradigm the observer is free to mark and rate as many suspicious regions as are considered reportable; moreover, the regions marked are at chosen by the reader, not the person conducting the study. The observers presumably mark and rate only those suspicious regions whose confidence levels exceed a certain minimum reporting threshold. The mark is the location of the suspicious region and the rating is the confidence level that the region harbors a lesion. The rating is similar to the ROC rating, e.g., 1–4, but with one important difference. The lowest rating, e.g., 1, *does not* represent high likelihood of the absence of a lesion at the indicated location (otherwise, the observer should fill up the whole image with marks rated 1). Rather, a "1" means the lowest non-zero confidence that there is a lesion at the indicated location. For example, a FROC rating scale could be designated as 1 = very low confidence, 2 = low confidence, 3 = moderate confidence, 4 = high confidence. The absence of any marks corresponds to the 1 = definitely disease-free ROC rating.

The unit of FROC data is the *mark-rating pair*. As part of the analysis, the experimenter decides whether a mark is close enough to a real lesion to qualify as lesion localization (LL) and otherwise the mark is classified as non-lesion localization (NL). This is referred to as *scoring the marks*. *What constitutes "close enough" is a clinical decision which should be decided based on the clinical application*. Two physicians do not need to agree on the exact center of a lesion in order to appropriately assess and treat it. The critical separation for a lesion localization event - termed *acceptance radius* - could be the diameter of the lesion, but constrained by appropriate lower and upper limits. For example, if the mark to lesion center distance were greater than the lesion diameter, it would be scored as non-lesion localization (NL). The lower limit on localization accuracy (e.g., 3 mm) is necessary, as otherwise "hand jitter" would cause small lesions to be incorrectly localized even though they were seen and correctly communicated. An unnecessarily large acceptance radius would result in clinically obvious non-lesion localizations being classified as lesion localizations. The acceptance radius should be similar for both modalities as otherwise there would be a bias favoring the modality with the larger value. A method of analyzing FROC data without explicitly scoring the marks has been described but needs further validation [48].

The FROC, AFROC2 and inferred ROC curves and operating points

Lesion localization fraction, $LLF(\zeta)$, is defined as the number of lesions that are correctly localized at threshold ζ divided by the total number of lesions in the dataset. This is the counterpart of $TPF(\zeta)$ in the ROC context except that $LLF(\zeta)$ is normalized to the total number of lesions N_L , while $TPF(\zeta)$ is normalized to the total number of diseased cases. The counterpart of $FPF(\zeta)$ is the number of non-lesion localizations divided by the number of images K , termed the non-lesion localization fraction, $NLF(\zeta)$. The FROC curve is defined as the plot of lesion localization fraction LLF vs. non-lesion localization fraction NLF [8,10]. FROC operating points can be calculated by a cumulation procedure analogous to that used to calculate ROC operating points [49].

If one assumes that the rating of the highest-rated mark on an image is its ROC-equivalent rating, termed the *highest rating assumption*, then one can infer ROC data from FROC data. If the image has no marks then its inferred rating is zero (any number smaller than the smallest explicit rating could be used without affecting the inferred ROC operating points). The AFROC2⁷ curve [50,51] is the plot of $LLF(\zeta)$ vs. $FPF(\zeta)$. Note that subject to the highest rating assumption, FPF is calculated over non-diseased cases only, identical to the manner in which FPF is calculated in the ROC context. Specifically, $FPF(\zeta)$ is the fraction of non-diseased cases that have one or more non-lesion localizations with a rating of ζ or higher.

An ROC curve can be inferred from FROC data, for example, using the highest rating assumption. On non-diseased images the highest rating is necessarily that of a NL, or zero, if there is no mark, but on diseased images, the highest rating could be a LL or an NL, whichever is rated higher, or zero, if there is no mark.

The FROC rating scale

The rating scale is at the discretion of the experimenter and should be selected in consultation with the radiologists. In mammography the BI-RADS [52] rating scale provides a natural scale, with the exception, as previously noted, that the 0-rating is not allowed in a laboratory study. Care should be taken to minimize confusion about the meaning of the rating. For example, a lesion visibility rating and a lesion malignancy rating are quite distinct, as benign lesions can be quite visible. In lesion detections studies it is common to use a malignancy rating, not a visibility rating. The 3-rating can be split as in [33] depending on the nature of the recommended follow up (short term vs. longer term). With multiple views, if the observer marks a location a pop-up menu should ask if the same lesion is visible on the other view and if so, the observer should mark the other view and the software should automatically assign the *same* rating to it. If the mark does not agree with the true lesion location in the 2nd view, then the observer is penalized with a non-lesion localization. An observer who marks both regions correctly receives the same credit as an observer who marks only one. [It is possible, but more complex, to design a scoring scheme where detection of the same lesion in both views is rewarded more than detection on a single view only. A clinician should be consulted on whether it is important for the same lesion to be localized in both views. This type of complication does not arise with true three dimensional imaging modalities like computed tomography or magnetic resonance imaging.]

⁷Notation: the x-axis of the original AFROC curve was defined as the number of images with one or more non-lesion localizations at a threshold divided by the total number of images, which counts non-lesion localizations on all images. The AFROC2 curve x-axis *excludes* non-lesion localizations on diseased images, so the relevant denominator is the number of non-diseased cases.

The raw FROC curve and its utility

Consider the marks arranged in two rows, one for non-lesion localizations (NLs) denoted by the dark shaded circles in Fig. 4, and one for lesion localizations (LLs) denoted by the light shaded circles. The marks are ordered with the ratings (z) increasing to the right. Starting with the threshold ζ on the extreme right at positive infinity, none of the ratings exceeds infinity, and the FROC operating point is (0,0). As one moves the threshold ζ slowly to the left the first mark crossed is a LL (light shaded circle) and crossing it yields the operating point $(0, 1/N_L)$ where N_L is the total number of lesions in the dataset. Continuing moving the threshold to the left, as one crosses each LL the operating point moves upwards by $1/N_L$ and as one crosses each NL the operating point moves to the right by $1/K$, where K is the total number of cases. In this way, one gets a saw-tooth-like curve, called the raw FROC curve, Fig. 5 (raw: because the data was not binned). Why is it useful? The raw FROC allows determination of whether the observer is fully using the rating scale. Ideally, the raw FROC curve should approach a plateau towards the upper end⁸.

The search model

The search-model is a statistical parameterization of concepts developed by Kundel and Nodine [53,54,55,56] that describe how experts interpret medical images. The Kundel-Nodine model has similarities to models developed in non-medical imaging contexts by others [57,58,59,60] and is consistent with eye-tracking recordings of the search patterns of radiologists. According to Kundel and Nodine, image perception involves a covert *search* stage and an overt *decision-making* stage. During the search stage, the radiologist quickly (within 300 ms from image onset) identifies suspicious regions. This is consistent with the view that based upon their experience the radiologist has built up a template of what a generic normal image should look like, and lesions in abnormal images are perceived as localized deviations (hot spots or suspicious regions) from this template. During the decision-making stage the fovea is directed towards each suspicious region in turn, the region and its surround are examined, and an overt decision is made whether or not to mark the region. In the simple model, the observer calculates a z -sample, and if it exceeds the minimum reporting threshold, the region is marked.

The search-model assumes that each image yields a random number of suspicious regions, termed *decision sites*. Decision sites are termed *noise sites* or *signal sites* if they correspond to normal anatomy or lesions, respectively. It is assumed that the number of noise-sites on an image is a sample from a Poisson distribution with mean λ . The number of signal-sites on a diseased image is assumed a sample from a binomial distribution with mean Lv and trial size L , where v is the probability that a lesion is a decision site (i.e., it is “found”) and L is the number of lesions in the image. The decision variable (z -sample) or confidence level from a noise-site is assumed sampled from $N(0,1)$ and that from a signal-site is assumed sampled from $N(\mu,1)$ ⁹. Fig. 6 illustrates the z -sampling model for noise sites and signal sites, respectively, for $\mu = 3.0$. The search model parameters can be estimated from ROC data using a maximum likelihood method [61].

⁸The number of points on the raw FROC curve equals the total number of marks. The binned curve plots a point only when the number in each of the NL and LL bins defined by a pair of neighboring cutoffs reaches a minimum value, typically 5. Then the number of FROC points will be considerably smaller than the number of marks.

⁹It is not necessary to assume that the two distributions have different variances. The ROC curves predicted by the model have binormal-like characteristics with b -parameter < 1 even though the sampling distribution is equal variance. This is because of the multiple noise sites samples per image, which results in a narrowing of the distribution of the highest z -samples on disease free images, which is equivalent $b < 1$ in the binormal model.

Operating characteristics

Expressions for operating characteristics (ROC, FROC, AFROC2 and LROC) predicted by the search model have been derived [62,63,64]. The point reached when the threshold is lowered to its lowest value, namely $\zeta = -\infty$, is denoted (x,y) . *Unlike conventional ROC curves, this point is not (1,1) rather (x,y) is always to the lower-left of (1,1).* The area under the ROC curve equals that under the continuous section extending from $(0,0)$ to (x,y) plus the area under the straight -line segment connecting (x,y) to $(1,1)$. The coordinates (x,y) depend on λ and v and the number of lesions on abnormal images. The reason (x,y) is not at $(1,1)$ is that the search-model allows for the possibility that some images have no decision sites and on them, there is literally nothing to report, and no z-samples are obtained. The probability¹⁰ that a non-diseased case has zero noise sites is $e^{-\lambda}$ and the probability that a diseased case with L lesions has zero noise sites *and* zero signal sites is $e^{-\lambda}(1-v)^L$. For $\lambda = 1$, $v = 0.7$ and $L = 1$, these translate to 0.368 and 0.110, respectively. In other words, about 37% of non-diseased cases and 11% of diseased cases have nothing to report or no z-samples. The absence of a z-sample is a stronger statement of perceived negativity of a case than the presence of a very low z-sample, since the latter could be called diseased at sufficiently low threshold but the former cannot. Suppose the observer is using a 101-point ROC scale, where zero represents certainty that there is no disease and 100 represents certainty that there is disease. If there is nothing to report, then the observer has to give the case the zero rating. On the remaining cases, there is something to report, and the observer is free to bin the cases using the bins 1 to 100. Cumulating the cases rated ≥ 1 will yield the operating point $x = \text{FPF}_{\text{MAX}} = 1 - 0.368 = 0.632$ and $y = \text{TPF}_{\text{MAX}} = 1 - 0.11 = 0.89$. As the cutoff is lowered the operating point approaches $(0.632, 0.89)$ from below, which is well short of $(1,1)$. At the highest achievable operating point, the observer has exhausted all cases with something to report. The point $(1,1)$ is not reached by the observer, rather it is trivially reached by the experimenter upon cumulating all cases, including the ones rated 0. The inaccessible section of the ROC curve is depicted Fig. 7 by a dotted straight line. With appropriate choices of parameters the uppermost operating point can be made practically indistinguishable from $(1,1)$. For example, for $\lambda = 10$, $v = 0.7$ and if every diseased case has one lesion ($L = 1$), the uppermost operating point¹¹ is $(0.99996, 0.999986)$ which is practically at $(1,1)$. The large value of λ ensures that practically every case has a noise site.

Another way of understanding the finite end-point property of the ROC is as follows. In conventional ROC models, every image yields a z sample. The z-samples are distributed according to the specified model (e.g., binormal) and while it is unlikely that unusually small samples (e.g., large negative values) will be observed, it is not impossible. With a million non-diseased images there will be about 32 z-samples¹² that are more negative than -4.0 , and at $\zeta = -4.0$ all but about 32 of the million non-diseased images are reported as diseased, and $\text{FPF} = 1 - 32 \times 10^{-6}$, which is very close to unity. In other words given enough cases the conventional ROC models allow the observer to operate arbitrarily close to $\text{FPF} = 1$ and a similar argument can be made for TPF. With the search model, however, a finite fraction of images yield no z-samples, and these can never be reported. For example, if $\lambda = 1$ then 0.368×10^6 non-diseased images have no z-samples, and $\text{FPF} = 1 - 0.368 = 0.632$, which cannot be made arbitrarily close to unity by increasing the number of cases.

¹⁰Explanation: on non-diseased cases, the only possible decision sites are at noise sites. Since the probability of observing n noise sites is $e^{-\lambda}\lambda^n/n!$ (the Poisson distribution) the probability of observing zero noise sites ($n=0$) is $e^{-\lambda}$. On disease-present cases the probability that no lesion is seen (zero signal sites) is $(1-v)^L$ - this follows from the fact that v is the probability that a lesion is a signal site and therefore $1-v$ is the probability that it is not a signal site, and the probability that none of the L lesions is a signal site is $(1-v)^L$. The probability that the diseased case does not have a noise site is $e^{-\lambda}$. Therefore, the probability that a diseased case has *no* sites is $e^{-\lambda}(1-v)^L$.

¹¹For $\lambda = 10$ and $v = 0.7$, $\exp(-\lambda) = 4.53999e^{-5}$, $e^{-\lambda}(1-v) = 1.36200e^{-5}$; FPF_{MAX} and TPF_{MAX} are the complements of these values.

¹²The area under the unit normal to the left of -4.0 is $3.16712e^{-5}$.

Example ROC, AFROC2 and FROC curves are shown in Fig. 7(a–d) for various choices of search model parameters and L (number of lesions per diseased case, assumed constant). Fig 7(a) is for $\mu = 2$, $\lambda = 1.5$, $v = 0.8$, $L = 1$. Fig 7(b) is for two lesions per diseased case, otherwise the parameters are the same as for Fig. 7(a). Note the effect on the ROC curve – AUC increases from 0.834 to 0.935 – but the AFROC2 and FROC curves are unaffected. This is because the y-axes of the AFROC2 and FROC (LLF) are normalized to the total number of lesions, so doubling the number of lesions has no effect on them. The ROC area increases because diseased cases with two lesions provide two opportunities for the observer to detect a lesion.¹³

Fig. 7(c) is for $\lambda = 5$, otherwise the parameters are the same as for Fig. 7(a). The ROC curve is seen to extend almost continuously to (1,1). Fig. 7(d) is $\mu = 0$, $\lambda = 1$, $v = 0.8$, $L = 1$. It illustrates an important point: observer performance in a lesion detection task has two components, namely search performance and classification performance. Search performance is determined by λ and v , which determine the location of the end-point (x,y). Search performance puts a lower limit on net ROC performance. If $\mu = 0$, then the ROC curve is that shown in Fig. 7 – namely, a straight line extending from the origin to (x,y) followed by the inaccessible section from (x,y) to (1,1). Consequently in Fig. 7(d) the value $AUC = 0.673$ is totally determined by search performance. In the limiting case of perfect search performance, all lesions are identified and no noise sites are identified. This puts the end-point at (0, 1). Having identified the lesions without the “baggage” of noise sites, there is no need to discriminate between signal and noise sites. Given perfect search performance any classifier, including a chance level classifier with $\mu = 0$, would achieve $AUC = 1$, simply because there is no need for a classifier to classify between the signal and the (non-existent) noise sites.

FROC Data analysis

The area under the AFROC2 curve is the most common figure of merit used in FROC analysis. Recall the AFROC2 curve is defined as the plot of lesion localization fraction (LLF) vs. false positive fraction (FPF), where the false positive rating of a non-diseased image is defined as the rating of the highest rated mark on the image, or negative infinity, if the image has no marks. Note that the only difference between the ROC plot and the AFROC2 plot is in the definitions of the respective y-axes, namely TPF vs. LLF. The area under the AFROC2 rewards lesion localizations and penalizes non-lesion localizations on normal images to different degrees, as determined by the assigned confidence levels. For example, a higher rated LL is rewarded more than a lower rated LL, and likewise a higher rated FP (the x-axis of the AFROC2 is false positive fraction) is penalized more than a lower rated FP.

An aspect of FROC analysis that has caused confusion is how to deal with the true negatives, i.e., non-diseased regions that were found by the search (equivalently “looked at” or “considered for marking” or “located”) but not marked. It has been recommended [65] that “Tip: In a lesion-level analysis, be sure to explain how you calculate the number of true-negative findings”. The editorial has unwittingly set an impossible task. The observable data in a free-response study are lesion and non-lesion localizations and their ratings. One does not know the number of location level “true negatives”, i.e., non-cancerous regions that were looked at but not marked. [The quotations are meant to emphasize the inherent ambiguity in applying ROC terminology to a search task; this is explained below.] To get at this number one might consider recording the radiologist’s eye-movements but, besides being

¹³If one included correlation effects then the independence of the area under the AFROC2 and lesion density would no longer be true, nevertheless, the dependence on lesion density is expected to be weaker than that of the ROC area, a second order effect.

impractical, even then one does not truly know, because peripherally found regions are not recorded by eye-trackers. Even asking the radiologist to point to the locations that were considered but not marked would not account for unconscious “true-negative” decisions.

While one does not know the total number of “true-negatives”, nevertheless it is possible to indirectly account for them. An observer who avoids marking non-diseased regions will also tend to have smaller number of case-level false positives, leading to smaller false positive fractions, which will increase the area under the AFROC2. This is illustrated in Fig. 8 where AFROC2 data points for two observers are shown. They have identical LLF values (0.3, 0.4, 0.5, 0.6) but the FPF values for the observer denoted by the triangle symbols are individually smaller (0.03, 0.08, 0.25, 0.6) than for the observer denoted by the circle symbols (0.05, 0.1, 0.3, 0.7). The corresponding trapezoidal areas under the AFROC2 curve are 0.611 and 0.575, respectively. Note that the observer with the smaller FPF values has the larger area under the AFROC2. This shows that the figure of merit is crediting observers for not marking non-diseased regions. A subtle point: an observer who avoids even looking at non-diseased regions will have an even larger area than an observer who looks at non-diseased regions but avoids marking them.

A substantial part of the observed area difference in Fig. 8 is due to the change in the contribution to the area from the line connecting the last explicit operating point to (1,1), which is the contribution from non-diseased cases with no marks. A non-diseased case with no marks is a perfect decision and it is important that the figure of merit credit the observer for such decisions. This contribution to the area under the AFROC2 would be missed by a partial area measure such as the area under the FROC curve [66,67,68,69] but not missed by the partial area measure suggested by [12] for ROC if applied to the AFROC2.

The use of the term “true-negatives” in the free-response context can create confusion. In ROC terminology a true-negative is a correct decision – the case was non-diseased and was interpreted as non-diseased, and more true-negatives are desirable. If applied blindly to the free-response context this would imply that an observer who looks at more non-diseased regions and does not mark them is better than an observer who looks at fewer non-diseased regions and does not mark them, when in fact the reverse is true. “True-negative” regions are potential non-lesion localizations. At sufficiently low threshold the observers would mark every region they looked at, in which case the first observer would generate more non-lesion localizations than the second observer. A non-diseased region that is not looked-at cannot ever be marked, no matter how low the threshold, so although both observers did not mark any non-diseased regions on this case, the first observer is potentially capable of marking more than the second observer, and on the average will generate more non-lesion marks and case-level false positives, and have a smaller area under the AFROC2.

JAFROC2

In jackknife alternative FROC2 (JAFROC2) analysis of FROC data the figure of merit is the area under the AFROC2 curve, currently estimated non-parametrically (i.e., using the trapezoidal rule), and denoted θ , and the significance testing is performed using the DBM-MRMC ANOVA method described earlier. A similar variance component model for the pseudovalues is used, the only difference being that the pseudovalues are calculated by jackknifing cases and computing the trapezoidal area θ under the AFROC2 curve each time. The pseudovalues are defined by

$$Y_{ijk} = K\theta_{ij} - (K - 1)\theta_{ij(k)}, \quad \text{Eqn. 7}$$

where θ_{ij} is the area under the AFROC2 for modality i and reader j , and $\theta_{ij(k)}$ is corresponding area with case k removed. Notice that in spite of the multitude of mark rating pairs that could occur on a case, *each case is represented by a single pseudovalue*. Since case is treated as a random factor, the analysis generalizes to the population of cases. In the author's opinion the idea, that because the FROC paradigm is "lesion-specific" therefore the analysis generalizes to the "population of lesions", is a misconception.

Sample size estimation proceeds along similar lines as for ROC. One needs the estimates of the modality-reader, modality-case and error variance components that are produced by the ANOVA. These are plugged into the Hillis-Berbaum calculator which yields the desired sample size estimates. The only real issue is what effect size to use. The area under the ROC has a well understood meaning and one can (arguably) agree that a difference $\Delta A = 0.03$ is a small effect size, $\Delta A = 0.05$ is a moderate effect size, and $\Delta A = 0.07$ is a large effect size, but *FROC sample size estimation requires one to estimate the equivalent changes in the area under the AFROC2*. Assuming the highest rating is used to infer the ROC rating, two expectations hold: (i) the area under the AFROC2 is generally smaller than the area under the ROC, and (ii) the change in the area under the AFROC2 is generally larger than the equivalent change in the area under the ROC. The first expectation follows from the definition of the y-axis of the AFROC2, namely lesion localization fraction (LLF). Whereas an abnormal image with a mark is always scored as a true positive, it contributes lesion localizations only if one or more lesions are correctly localized. The second expectation can be understood intuitively: the area under the AFROC2 can range from 0 (none of the lesions are marked, all non-diseased cases have at least one mark) to unity (all lesions are marked, no non-diseased cases are marked), while the area under the ROC curve ranges from 0.5 (chance level performance) to unity (perfect performance). Given its larger span it should not be surprising that the change in the AFROC2 area between two modalities tends to be larger than the corresponding change in the ROC area. [The fact that the effect size depends on the units can be seen to be trivially true: if instead of AUC one uses twice AUC, then the effect size will double. As another example, the effect size in d' units is also larger than the effect size in AUC units, particularly when AUC is close to unity.]

The problem in FROC sample size estimation is how to choose an effect size that corresponds to a given effect size in AUC units. This is accomplished in JAFROC2 software by the following method. One performs search model based maximum likelihood fits to the ROC curves, yielding μ_{ij} , λ_{ij} and v_{ij} , where i is the modality index and j is the reader index, and μ , λ and v are the search model parameters. One averages these over modalities and readers, denoted $\bar{\mu}_{..}$, $\bar{\lambda}_{..}$ and $\bar{v}_{..}$. [The values are averaged over modalities to obtain more stable estimates, bearing in mind that the pilot study probably returned a non-significant result, as otherwise sample size estimation for a larger study would be unnecessary.] The average parameters are taken as characterizing the null hypothesis modality and both AFROC2 and ROC area are computed using search model formulae. The alternative hypothesis modality is constructed by increasing $\bar{\mu}_{..}$ by a small amount, decreasing $\bar{\lambda}_{..}$ by a small amount and increasing $\bar{v}_{..}$ by a small amount, recalculating the ROC area, and repeating the process until the change in ROC area equals the desired amount, e.g., 0.05. The corresponding AFROC2 area minus the null hypothesis value yields the desired effect size in AFROC2 units. [The parameter choices for a given effect size are not unique.]

JAFROC1

The reader may have noticed that the area under the AFROC2 as defined previously does not utilize the non-lesion localizations on diseased cases. An alternate figure of merit, termed JAFROC1, used a different definition of the false positive fraction axis, namely by counting the number of highest rated non-lesion localization on every case, and dividing by

the total number of cases, in other words it uses the historical definition of the AFROC curve [50,51]. Since more cases are used to form the estimate of FPF, the values tend to be more stable and statistical power is higher [70] as compared to JAFROC2, where only non-lesion localization on non-diseased cases are counted towards FPF. When tested with a single reader dual modality search model based simulator JAFROC1 had correct null hypothesis behavior. However, the author has encountered a dataset where JAFROC1 analysis did not reach statistical significance while both JAFROC2 and ROC (using the highest rating) analysis did. This dataset is characterized by very few non-lesion localization marks on diseased cases relative to the non-diseased cases. This effect, which would traditionally be attributed to satisfaction of search [71], has been observed by the author on almost every human observer FROC dataset, is not modeled in the simulator used to test JAFROC1. Therefore the author made the recommendation to not use JAFROC1 unless the dataset has no non-diseased cases, in which case the need to account for asymmetry in FPF between non-diseased and diseased cases disappears, and one should use JAFROC1.

The JAFROC2 figure of merit has the drawback that it gives more importance to cases with more lesions. The denominator in the formula for the lesion localization fraction is the total number of lesions. For example, if 99 cases had 1 lesion each and one case had 99 lesions, the single case with 99 lesions would have as much influence on the figure of merit as the remaining 99 cases combined. This possibility is not taken into account in the pseudovalue analysis, where every case is assumed to be sampled from a common Gaussian distribution representing the case population. A method of weighting the lesions that essentially gives each case equal importance has been described [72] but has not yet been implemented in the software.

Other FROC figures of merit

In any observer performance paradigm, the choice of figure of merit is a key to proper analysis of the data. Just as several figures of merit have been proposed for ROC analysis [73], there are a few that have been proposed for FROC analysis. The area under the FROC curve to the left of a pre-selected value of the abscissa has been proposed as a suitable figure of merit [68]. Like the AFROC (since the two share the same y-axis) it rewards more lesion localizations. However, unlike the AFROC, it does not explicitly reward fewer marks on normal images and does not pass the accounting for “true negatives” criteria that was discussed previously. It is also susceptible to loss of statistical power since marks that contribute to the portion of the FROC curve to the right of the selected threshold do not contribute to the area under the selected portion of the FROC curve. Another full area index of performance has been proposed [74], namely the area under the empirical FROC curve penalized for the number of erroneous marks, rewarded for the fraction of detected abnormalities, and adjusted for the effect of the target size (or “acceptance radius”). The index was interpreted as a measure of average performance superiority over an artificial “guessing” free-response process.

Acknowledgments

The author was supported in part by grants from the Department of Health and Human Services, National Institutes of Health, R01-EB005243 and R01-EB008688. The author acknowledges contributions to the manuscript from Dr. Howard Gifford and helpful comments from Dr. Berkman Sahiner.

References

1. Metz CE. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*. 1978; 8:283–298. [PubMed: 112681]

2. Metz CE. ROC Methodology in Radiologic Imaging. *Investigative Radiology*. 1986; 21:720–733. [PubMed: 3095258]
3. Metz CE. Some Practical Issues of Experimental Design and Data Analysis in Radiological ROC studies. *Investigative Radiology*. 1989; 24:234–245. [PubMed: 2753640]
4. Zweig MH, Campbell G. Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. *Clin Chem*. 1993; 39:561–577. [PubMed: 8472349]
5. Kundel HL, Berbaum KS, Dorfman DD, Gur D, Metz CE, et al. Receiver Operating Characteristic Analysis in Medical Imaging. ICRU Report 79. 2008; 8(1)
6. Metz C. ROC analysis in medical imaging: a tutorial review of the literature. *Radiological Physics and Technology*. 2008; 1:2–12. [PubMed: 20821157]
7. Shiraishi J, Pesce LL, Metz CE, Doi K. Experimental Design and Data Analysis in Receiver Operating Characteristic Studies: Lessons Learned from Reports in Radiology from 1997 to 2006. *Radiology*. 2009; 253:822–830. [PubMed: 19864510]
8. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH. A Free-Response Approach to the Measurement and Characterization of Radiographic-Observer Performance. *J of Appl Photogr Eng*. 1978; 4:166–171.
9. Egan JP, Greenburg GZ, Schulman AI. Operating characteristics, signal detectability and the method of free response. *J Acoust Soc Am*. 1961; 33:993–1007.
10. Miller H. The FROC curve: a representation of the observer's performance for the method of free response. *The Journal of the Acoustical Society of America*. 1969; 46:1473–1476. [PubMed: 5361517]
11. Marcum, JI. A Statistical Theory of Target Detection by Pulsed Radar. 1947.
12. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology*. 1996; 201:745–750. [PubMed: 8939225]
13. McClish DK. Analyzing a Portion of the ROC Curve. *Medical Decision Making*. 1989; 9:190–195. [PubMed: 2668680]
14. Green, DM.; Swets, JA. *Signal Detection Theory and Psychophysics*. New York: John Wiley & Sons; 1966. p. 505
15. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample. *Archives of Internal Medicine*. 1996; 156:209–213. [PubMed: 8546556]
16. Dorfman DD, Alf E. Maximum-Likelihood Estimation of Parameters of Signal-Detection Theory and Determination of Confidence Intervals - Rating-Method Data. *Journal of Mathematical Psychology*. 1969; 6:487–496.
17. Metz CE, Herman B, Shen J. Maximum-Likelihood Estimation of Receiver Operating Characteristic (ROC) Curves from Continuously-Distributed Data. *Statistics in Medicine*. 1998; 17:1033–1053. [PubMed: 9612889]
18. Dorfman DD, Berbaum KS. A contaminated binormal model for ROC data: Part III. Initial evaluation with detection. *Acad Radiol*. 2000; 7:438–447. [PubMed: 10845403]
19. Dorfman DD, Berbaum KS, Metz CE, Lenth RV, Hanley JA, et al. Proper Receiving Operating Characteristic Analysis: The Bigamma model. *Acad Radiol*. 1997; 4:138–149. [PubMed: 9061087]
20. Metz CE, Pan X. Proper Binormal ROC Curves: Theory and Maximum-Likelihood Estimation. *J Math Psychol*. 1999; 43:1–33. [PubMed: 10069933]
21. Burgess AE. Comparison of receiver operating characteristic and forced choice observer performance measurement methods. *Med Phys*. 1995; 22:643–655. [PubMed: 7643805]
22. Eng, J. ROC analysis: web-based calculator for ROC curves. 2006.
23. Pesce LL, Metz CE. Reliable and Computationally Efficient Maximum-Likelihood Estimation of Proper Binormal ROC Curves. *Acad Radiol*. 2007; 14:814–829. [PubMed: 17574132]
24. McNeil BJ, Hanley JA. Statistical Approaches to the Analysis of Receiver Operating Characteristic (ROC) Curves. *Medical Decision Making*. 1984; 4:137–150. [PubMed: 6472062]
25. Hanley JA. The Place of Statistical Methods in Radiology (and in the Bigger Picture). *Journal of the American Medical Association*. 1988; 24:10–16.

26. Chakraborty, DP.; Krupinski, EA. Short Course: Statistical Methods in Bioengineering and ROC and FROC Analyses. Orlando, FL: 2011.
27. Rockette HE, Gur D, Metz CE. The Use of Continuous and Discrete Confidence Judgments in Receiver Operating Characteristic Studies of Diagnostic Imaging Techniques. *Investigative Radiology*. 1992; 27:169–172. [PubMed: 1601610]
28. Wagner RF, Beiden SV, Metz CE. Continuous versus Categorical Data for ROC Analysis: Some Quantitative Considerations. *Academic Radiology*. 2001; 8:328–334. [PubMed: 11293781]
29. Metz CE, Herman BA, Shen J-H. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*. 1998; 17:1033–1053. [PubMed: 9612889]
30. Berbaum KS, Dorfman DD, Franken EA, Caldwell RT. An Empirical Comparison of Discrete Ratings and Subjective Probability Ratings. *Academic Radiology*. 2002; 9:756–763. [PubMed: 12139089]
31. D'Orsi, CJ.; Bassett, LW.; Feig, SA.; Jackson, VP.; Kopans, DB., et al. Illustrated Breast Imaging Reporting and Data System. Reston, Va: American College of Radiology; 1998.
32. D'Orsi, CJ.; Bassett, LW.; Berg, WA. ACR BI-RADS-Mammography. Reston, Va: American College of Radiology; 2003.
33. Barlow WE, Chi C, Carney PA, Taplin SH, D'Orsi C, et al. Accuracy of Screening Mammography Interpretation by Characteristics of Radiologists. *Journal of the National Cancer Institute*. 2004; 96:1840–1850. [PubMed: 15601640]
34. Jiang Y, Metz CE. BI-RADS Data Should Not Be Used to Estimate ROC Curves¹. *Radiology*. 2010; 256:29–31. [PubMed: 20574083]
35. Efron, B. The Jackknife, the Bootstrap and Other Resampling Plans. Montpelier: Capital City Press; 1982.
36. Dorfman DD, Berbaum KS, Metz CE. ROC characteristic rating analysis: Generalization to the Population of Readers and Patients with the Jackknife method. *Invest Radiol*. 1992; 27:723–731. [PubMed: 1399456]
37. ICRU. Statistical Analysis and Power Estimation. *JOURNAL OF THE ICRU*. 2008:37–40.
38. Obuchowski NA. Multireader, Multimodality Receiver Operating Characteristic Curve Studies: Hypothesis testing and Sample Size Estimation Using an Analysis of Variance Approach with dependent Observations. *Acad Radiol*. 1995; 2:S22–S29. [PubMed: 9419702]
39. Obuchowski NA. Sample size calculations in studies of test accuracy. *Statistical Methods in Medical Research*. 1998; 7:371–392. [PubMed: 9871953]
40. Obuchowski NA. Sample Size Tables For Receiver Operating Characteristic Studies. *Am J Roentgenol*. 2000; 175:603–608. [PubMed: 10954438]
41. Hillis SL, Berbaum KS. Power Estimation for the Dorfman-Berbaum-Metz Method. *Acad Radiol*. 2004; 11:1260–1273. [PubMed: 15561573]
42. Starr SJ, Metz CE, Lusted LB, Goodenough DJ. Visual detection and localization of radiographic images. *Radiology*. 1975; 116:533–538. [PubMed: 1153755]
43. Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. *Med Phys*. 1996; 23:1709–1725. [PubMed: 8946368]
44. Swensson RG. Using Localization Data from Image Interpretations to Improve Estimates of Performance Accuracy. *Medical Decision Making*. 2000; 20:170–184. [PubMed: 10772355]
45. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*. 1975; 12:387–415.
46. Popescu LM. Nonparametric ROC and LROC analysis. *Medical Physics*. 2007; 35:1556–1564. [PubMed: 17555237]
47. Obuchowski NA, Lieber ML, Powell KA. Data Analysis for Detection and Localization of Multiple Abnormalities with Application to Mammography. *Acad Radiol*. 2000; 7:516–525. [PubMed: 10902960]
48. 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. [PubMed: 17178361]

49. Chakraborty DP. Recent developments in imaging system assessment methodology, FROC analysis and the search model. *Nucl Instr and Meth A*. 2010
50. Chakraborty DP. Maximum Likelihood analysis of free-response receiver operating characteristic (FROC) data. *Med Phys*. 1989; 16:561–568. [PubMed: 2770630]
51. Chakraborty DP, Winter LHL. Free-Response Methodology: Alternate Analysis and a New Observer-Performance Experiment. *Radiology*. 1990; 174:873–881. [PubMed: 2305073]
52. Geller BM, Barlow WE, Ballard-Barbash R, Ernster VL, Yankaskas BC, et al. Use of the American College of Radiology BI-RADS to Report on the Mammographic Evaluation of Women with Signs and Symptoms of Breast Disease1. *Radiology*. 2002; 222:536–542. [PubMed: 11818625]
53. Kundel HL, Nodine CF. A visual concept shapes image perception. *Radiology*. 1983; 146:363–368. [PubMed: 6849084]
54. Kundel HL, Nodine CF. Modeling visual search during mammogram viewing. *Proc SPIE*. 2004; 5372:110–115.
55. Nodine CF, Kundel HL. Using eye movements to study visual search and to improve tumor detection. *RadioGraphics*. 1987; 7:1241–1250. [PubMed: 3423330]
56. Kundel HL, Nodine CF, Conant EF, Weinstein SP. Holistic Component of Image Perception in Mammogram Interpretation: Gaze-tracking Study. *Radiology*. 2007; 242:396–402. [PubMed: 17255410]
57. Rock, I. *The logic of perception*. Cambridge, Massachusetts: The MIT press; 1983.
58. Rock, I. *Perception*. New York, NY: WH Freeman and Company; 1984.
59. Gregory, RL. *The Intelligent Eye*. New York, NY: McGraw-Hill; 1970.
60. Wolfe JM. Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*. 1994; 1:202–238.
61. Chakraborty, DP.; Svahn, T. Estimating the parameters of a model of visual search from ROC data: an alternate method for fitting proper ROC curves. Lake Buena Vista, FL: 2011.
62. Chakraborty DP. A search model and figure of merit for observer data acquired according to the free-response paradigm. *Phys Med Biol*. 2006; 51:3449–3462. [PubMed: 16825742]
63. Chakraborty DP. ROC Curves predicted by a model of visual search. *Phys Med Biol*. 2006; 51:3463–3482. [PubMed: 16825743]
64. Chakraborty DP, Yoon HJ. Operating characteristics predicted by models for diagnostic tasks involving lesion localization. *Med Phys*. 2008; 35:435–445. [PubMed: 18383663]
65. Levine D, Bankier AA, Halpern EF. Submissions to Radiology: Our Top 10 List of Statistical Errors1. *Radiology*. 2009; 253:288–290.
66. Paquerault S, Samuelson FW, Myers KJ, Smith RC. Non-localization and localization ROC analyses using clinically based scoring. *Proc SPIE Medical Imaging*. 2009; 2009:7263.
67. Paquerault S, Samuelson FW, Petrick N, Myers KJ. Comparing Signal-based and Case-based Methodologies for CAD Assessment in a Detection Task. *Proc SPIE Medical Imaging*. 2008:6917.
68. Samuelson, FW.; Petrick, N. Comparing image detection algorithms using resampling. *IEEE International Symposium on Biomedical Imaging: From Nano to Micro*; 2006. p. 1312-1315.
69. Samuelson, FW.; Petrick, N.; Paquerault, S. Advantages and examples of resampling for CAD evaluation. *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*; 2007. p. 492-495.
70. Chakraborty DP. Validation and Statistical Power Comparison of Methods for Analyzing Free-response Observer Performance Studies. *Acad Radiol*. 2008; 15:1554–1566. [PubMed: 19000872]
71. Berbaum KS, Franken EA, Dorfman DD, Rooholamini SA, Kathol MH, et al. Satisfaction of Search in Diagnostic Radiology. *Invest Radiol*. 1990; 25:133–140. [PubMed: 2312249]
72. Chakraborty DP, Yoon HJ. JAFROC analysis revisited: figure-of-merit considerations for human observer studies. *Proc SPIE Medical Imaging: Image Perception, Observer Performance, and Technology Assessment*. 2009; 7263:72630T.
73. Pepe, MS. *The statistical evaluation of medical tests for classification and prediction*. New York: Oxford University Press; 2003.

74. Bandos AI, Rockette HE, Song T, Gur D. Area under the Free-Response ROC Curve (FROC) and a Related Summary Index. *Biometrics*. 2008; 65:247–256. [PubMed: 18479482]

Fig. 1.

An ROC curve (dotted line) showing the variation of true positive fraction (TPF) vs. false positive fraction (FPF). This curve was generated using the binormal model, to be described below, with parameters $a = 1.5$ and $b = 0.6$. The operating point (0.1587, 0.8159) shown by the filled circle corresponds to $\zeta = 1$. As ζ approaches $-\infty$, the operating point approaches (1,1) and as ζ approaches $+\infty$, the operating point approaches (0,0). The area under the dotted curve is $AUC = 0.9$. The thick line is the operating characteristic for a perfect observer who is capable of operating at (0,1) by declaring every diseased case as diseased while not declaring any non-diseased case as diseased.

Fig. 2.

This figure is adapted with permission from the Beam et al study. It shows the ROC operating points of 108 mammographers. Wide variability in both sensitivity (40%) and specificity (45%) are evident. When fitted to ROC curves, the variability in AUC is substantially smaller (11%).

Fig. 3. ROC curves fitted by maximum likelihood estimation to the data shown in Table 1. The dashed curve is the upper 95% confidence range and the dotted curve is the lower 95% confidence range.

Fig. 4.

This figure illustrates the formation of the raw FROC curve. The light shaded circles represent lesion localization marks and the dark shaded circles are non-lesion localization marks, and the marks have been sorted by their z -samples (confidence levels) increasing towards the right. Starting at the very right, as the threshold ζ is lowered, each crossing of a light shaded circle results in an upward movement of the operating point by $1/N_L$, where N_L is the total number of lesions in the dataset, and each crossing of a dark shaded circle results in a rightward movement of the operating point by $1/K$, where K is the total number of cases.

Fig. 5.

Raw FROC curves for three readers for the same image set. The reader depicted in (a) is making good use of the rating scale, as evident by the leveling out of the raw FROC curve. The curves for readers (b) and (c) are not leveling out implying that these readers could detect more lesions by reporting subtler suspicious regions. Consequently, their ability to detect lesions will be underestimated. Moreover, their ratings are concentrated in the high confidence region at the expense of lack of coverage of the low confidence region. [Data courtesy of Tony Svahn.]

Fig. 6.

The probability density functions (pdfs) of noise and signal site z -samples according to the search model for $\mu = 3.0$. The two distributions are unit variance Gaussians, with means 0.0 (noise sites) and 3.0 (signal sites), respectively. If $\lambda = 1.5$ then the number of noise sites is Poisson distributed with mean 1.5 noise sites per image. If $v = 0.8$ then the number of signal sites per abnormal image is binomial distributed with 80% probability of success.

Fig 7.

Search model predicted ROC, FROC and AFROC2 curves for (a) $\mu = 2, \lambda = 1.5, v = 0.8, L = 1$ and (b) $\mu = 2, \lambda = 1.5, v = 0.8, L = 2$, (c) $\mu = 2, \lambda = 5, v = 0.8, L = 1$ and (d) $\mu = 0, \lambda = 1, v = 0.8, L = 1$. The ROC is the plot of TPF vs. FPF, the FROC is the plot of LLF vs. NLF and the AFROC2 is the plot of LLF vs. FPF. The areas under the ROC and AFROC2 curves for the four panels are (a) 0.834, 0.745, (b) 0.935, 0.745, (c) 0.766, 0.616 and (d) 0.673, 0.543, respectively. In (c) the FROC curve extends to NLF = 5.0.

Fig. 8.

This figure illustrates how the area under the AFROC2 curve accounts for “true negatives” in an FROC study. Assuming both observers find the same suspicious regions, the observer indicated by the triangle symbols makes fewer marks on non-diseased regions, i.e., has more “true-negatives”, than the one indicated by the circles. The smaller FPF values results in an increased area under the trapezoidal AFROC2 curve. Thus the observer with more “true-negatives” is rewarded with the larger area. If the triangle symbol observer avoids finding some of the non-diseased suspicious regions found by the second (circle) observer, then the FPF values will be even smaller and the area difference will be further accentuated, even though the triangle symbol observer has fewer “true-negatives”. This shows that the term “true-negatives” can be misleading in the FROC context.

Table 1

This table shows a sample ROC dataset for 60 non-diseased cases (NOR) and 50 diseased cases (ABN) and illustrates the procedure used to calculate the operating points. Ratings data was acquired on a 5-point scale ($R = 5$) with 5 representing maximum confidence in presence of disease. For example, 19 out of 60 non-diseased cases were rated 2 and 6 diseased cases were also rated 2. The last two rows list the operating points. The lowest operating point is (0.017, 0.440) and the highest is (0.5, 0.9). The operating point at (1, 1) is a trivial one in the sense that any ROC dataset will yield this point. An R-point rating scale generally leads to R-1 non-trivial operating points.

	Total	Bin counts				
		Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
NOR	60	30	19	8	2	1
ABN	50	5	6	5	12	22
ROC	Operating points					
		Bins ≥ 5	Bins ≥ 4	Bins ≥ 3	Bins ≥ 2	Bins ≥ 1
FPF		0.017	0.050	0.183	0.500	1.000
TPF		0.440	0.680	0.780	0.900	1.000

Table 2

The ordering of the BI-RADS ratings shown in the first column correlates with the cancer-rate shown in the last column. Data from [33], used with permission.

BI-RADS assessment	Total number of mammograms	Mammograms in women without breast cancer (%)	Mammograms in women with breast cancer (%)	Cancer rate per 1,000 screening mammograms
1: Normal	356,030	355,734 (76.2)	296 (12.3)	0.83
2: Benign finding	56,614	56,533 (12.1)	81 (3.4)	1.43
3: Probably benign with a recommendation for Normal or Short-term follow-up	8692	8,627 (1.8)	65 (2.7)	7.48
3+: Probably benign with a recommendation for immediate work-up	3094	3,049 (0.7)	45 (1.9)	14.54
0: Need additional imaging evaluation	42,823	41,442 (8.9)	1,381 (57.5)	32.25
4: Suspicious abnormality, biopsy should be considered	2022	1,687 (0.4)	335 (13.9)	165.68
5: Highly suggestive of malignancy	237	38 (0.0)	199 (8.3)	839.66