The Value of Observer Performance Studies in Dose Optimization: A Focus on Free-Response Receiver Operating Characteristic Methods*

John D. Thompson[1,2], David J. Manning[3], and Peter Hogg[1]

[1]University of Salford, Salford, United Kingdom; [2]University Hospitals of Morecambe Bay NHS Foundation Trust, Barrow-in-Furness, United Kingdom; and [3]Lancaster University, Lancaster, United Kingdom

Receiver operating characteristic (ROC) analysis has been successfully used in radiology to help determine the combined success of system and observer. There is great value in these methods for assessing new and existing techniques to see if diagnostic accuracy can be improved. Within all aspects of radiology there should be compliance with the as-low-as-reasonably-achievable principle, which requires optimization of the diagnostic suitability of the image. Physical measures of image quality have long been used in the assessment of system performance, but these alone are not sufficient to assess diagnostic capability. It is imperative that the observer be included in any assessment of diagnostic performance. The free-response ROC paradigm has been developed as a statistically powerful advancement of traditional ROC analysis that allows a precise interpretation of complex images by adding location information to the level of observer confidence. The following review of free-response ROC methodology will explain how observer performance methods can be valuable in image optimization, including examples of how these have already been successful in hybrid imaging.

Key Words: FROC; optimization; observer performance

J Nucl Med Technol 2013; 41:57–64
DOI: 10.2967/jnmt.112.116566

In radiology there is a requirement to produce images of adequate diagnostic quality while minimizing the radiation burden. As part of this optimization process, it is accepted that physical measures of image quality, such as signal-to-noise ratio, contrast resolution, and spatial resolution, can be useful to determine system performance. Despite their value in determining image quality, these types of measurements do not allow quantification of the diagnostic value of an image. This poses an important problem—how good does an image need to be in order to answer the clinical question? To solve this problem there is a need to involve the observer, with a clearly defined diagnostic objective against which the observer's decision outcome can be measured. There are various approaches to including observer opinion in measures of diagnostic performance, and one of them is receiver operating characteristic (ROC) analysis. ROC analysis, originally developed for the analysis of RADAR signal detection in the 1950s, was introduced to radiology in the 1960s (*1,2*), where it has been widely used to measure the diagnostic accuracy of imaging techniques in which the observer is considered an integral part of the system (*3–5*). Measures of observer performance are valuable when overall system performance is being assessed, with the potential to provide data complementary to physical measures of signal-to-noise ratio, contrast resolution, and spatial resolution that can be obtained from the image. Only when these data have been obtained can the ultimate success of a diagnostic test be established (*6*). Knowing that the procedures and methods associated with physical measures of image quality are well reported, with an emphasis on ROC, we will concentrate in this paper on outlining the value of observer performance and its relationship with dose and image quality in the optimization process.

OBSERVER PERFORMANCE FOR ASSESSMENT OF SYSTEM PERFORMANCE

When a new imaging technique is being assessed, it can be difficult to determine whether there is any advantage over existing methods (*7*) because such differences are often quite small. Image quality assessment via a quality assurance program can sometimes be a good predictor of diagnostic performance, but values of signal-to-noise ratio, contrast resolution, or spatial resolution do not necessarily predict success in the visual search, which is where a diagnosis is made. Indeed, the diagnostic process may not improve linearly with physics-based improvements in image quality when it is possible that a range of image qualities can provide the radiologist with the opportunity to achieve the correct diagnosis. Therefore, it is useful to

Received Oct. 30, 2012; revision accepted Feb. 28, 2013.
For correspondence or reprints contact: John Thompson, Nuclear Medicine Department, Furness General Hospital, Barrow-in-Furness, Dalton Lane, Cumbria, U.K. LA14 4LF.
*NOTE: FOR CE CREDIT, YOU CAN ACCESS THIS ACTIVITY THROUGH THE SNMMI WEB SITE (http://www.snmmi.org/ce_online) THROUGH JUNE 2015.
Published online Apr. 26, 2013.

determine the thresholds at which the image is no longer diagnostic, or at which no further information is gained. Despite the usefulness of physical measures of image quality, it is not possible to truly characterize system performance if the observer is not included in the analysis. As an example, consider the comparative detection of bone metastases by whole-body scintigraphy (WBS) and skeletal radiography. In skeletal radiography, signal-to-noise ratio and modulation transfer function are vastly superior to WBS. Nevertheless, WBS is more successful in allowing observer detection of metastases, largely because of the superior sensitivity of WBS to radiography in the detection of bone mineral turnover (*8,9*). So despite inferior measures of image quality, an ROC study would find WBS superior to skeletal radiography for observer detection. If a robust method of characterizing system performance is required, then observer performance studies, using ROC methods, are seen as a partial solution (*7*).

ROC ANALYSIS

To aid understanding of ROC, a summary of terms synonymous with this observer performance method can be found in Table 1. An observer decision in a diagnostic test will have 1 of only 4 possible outcomes: true-positive, true-negative, false-positive, or false-negative. Correctly classifying patients as being positive (true-positive) or negative (true-negative) for disease requires decision thresholds (*10*). However, there can be many variations of decision threshold for a disease type (e.g., normal, benign, possibly malignant, or definitely malignant), each with the potential to yield a different estimate of sensitivity and specificity (*10*). The classification of disease is highly dependent on the decision threshold, complicating matters if the comparison of tests or observers is a key objective (*10*). Typical measures of test accuracy (sensitivity and specificity) are produced via a binary decision (yes or no) with only a single threshold accounted for. Therefore sensitivity and specificity measurements appear limited in their definition of test accuracy in comparison to ROC analysis (*4,11*). Indeed, measures of single sensitivity and specificity pairs alone are inadequate since they do not account for the system's ability to distinguish between actually negative patients and actually positive patients (*11*). The success and popularity of ROC methods therefore arise from the ability of the analysis to provide an excellent description of diagnostic accuracy for a full range of sensitivity and specificity of a particular test (*11*), with ROC methods well established for comparing diagnostic tests. In ROC analysis, diagnostic accuracy is a comparison of an index of detectability, such as the area under the ROC curve (AUC) (*12*), and concerns the level at which diagnostic tests show agreement (*11*). An ROC analysis not only is concerned with the correct classification of disease but also requires a rating to classify the observer's confidence in determining signal from noise. Ratings can be performed with a variety of confidence scales, either discrete or quasi continuous.

ROC CURVES AND AUC

Typically, an ROC curve (Fig. 1) is produced for each imaging modality or diagnostic test and is a typical measure of test accuracy (*1,5,12*). The curves are a plot of true-positive fraction against false-positive fraction for a full range of threshold values (*1,12,13*), thus providing a solution to the problematic single-threshold values associated with measures of sensitivity and specificity (*11*). However, the ROC curve is a display not only of observer performance skill but also of the physical limitations of the image, such as noise (*11*).

ROC curves can be constructed only for diagnostic tests that have been exposed to known healthy and diseased subjects. ROC curves are constructed by calculating false-positive fraction/true-positive fraction pairs of figures at each decision threshold level, as dictated by the rating (confidence) scale in operation (*14*). The operating points (0,0) and (1,1) are included in all ROC curves. The number of rating scale thresholds determines the number of nontrivial operating points on the ROC curve. This process, however, plots only an empiric curve (straight lines), and curve fitting is often applied. The theoretic ROC curve is smooth and continuous and allows extrapolation of results. Curve fitting is performed by nonparametric or parametric methods.

The AUC, or A_Z, is a standard measure of diagnostic accuracy, reflecting combined observer performance and modality performance (*3,13*). ROC curves can provide investigators and their readership with a valuable visual assessment of diagnostic accuracy, and empiric estimations have shown that AUC is equivalent to the Mann–Whitney statistic (*13*).

TABLE 1
Terminology Associated with ROC and FROC

Term	Meaning
Conspicuity	Target visibility within structured surrounding
Case	Image
Modality	Variation (technique, image acquisition, imaging modality, dose)
Lesion localization	Correct marking of lesion
Nonlesion localization	Incorrect marking of lesion
Truth	True disease status of image
Mark-rating pair	Localization decision and confidence (rating) score
Multireader multicase	Study design using multiple readers to account for reader variability; results can be generalized to populations of readers and cases

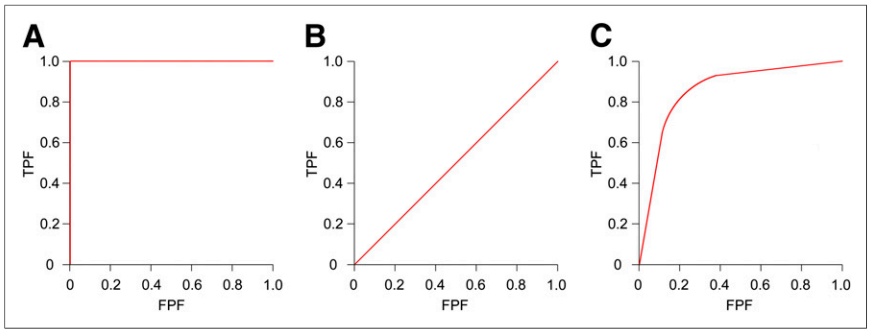

FIGURE 1. Typical ROC curve appearances: perfect test (A), chance diagonal (B), and good test (C). TPF = true-positive fraction; FPF = false-positive fraction.

For visual assessment of ROC curves to be worthwhile, it is important to understand the typical graphical appearances. A perfect test would show a curve as described by Figure 1A, a line running from point 0,0 to 0,1 and then to 1,1, with an AUC of 1.0. At the other extreme is the chance diagonal (Fig. 1B) running from point 0,0 to 1,1. This curve represents a test that is no better than random guessing (just as likely to be correct as incorrect). Furthermore, any curve sitting in the lower portion of the graph, below the chance diagonal, describes a test that is more often incorrect. The expected curve shape of a good test is described by Figure 1C. Tests with a high diagnostic accuracy are considered to have an AUC of around 0.9, and those with moderate accuracy reside in the region of 0.75 (*15*).

Measurements of full AUC represent the probability that diseased and nondiseased are correctly classified (*13*). These are the most commonly used measures of accuracy, but on occasion there is interest in only a certain portion of the ROC curve, known as partial area (*15*). The partial area is frequently defined as the area between 2 false-positive rate points on the ROC curve (*16*). Partial area can be valuable when false-negative results can have high significance (*15*) (screening for cancer) and when a near-perfect sensitivity is a requirement of the test. The converse is also true for tests that can incur potentially harmful treatment after a false-positive outcome (*15*). Considering these examples, it can be understood that evaluating full AUCs can in some instances be misleading and thus lead to poorer patient outcomes than expected (*16*).

ROC: CONFIDENCE SCALES

A confidence scale allows an observer to provide a numeric rating to an image based on the likelihood that disease is present (*3*). Confidence scales are generally of 2 formats—discrete (ordinal) or quasi continuous (Table 2)—and there is no optimal scale that suits all studies. It had been conventional to collect confidence scores using a discrete scale of 5–7 categories (*17,18*), with quasi continuous (101-point, 0–100) scales suggested as a solution to degenerate datasets (poorly constructed ROC curves) (*18,19*). The scale must suit the study design, but in some situations a quasi continuous scale may provide a more precise measure of diagnostic accuracy (*18*).

When using a discrete scale, the observer is frequently provided with a series of statements or percentage points to select from when scoring an image (Table 2). This style has been used to allow observers to rate their relative confidence that an abnormality is present (*5*). If observers are confident that no abnormality is present, they do not score the image and a default score of zero is applied.

Quasi continuous scales have recently become popular and in some instances have been applied through the use of a slider-bar style of confidence scale. These scales are classed as 101-point, 100 possible confidence ratings and zero for unscored images. Consequently, these scales can be effectively used to measure observer confidence in terms of percentage certainty of disease presence (i.e., the observer is 60% confident an abnormality is present).

However, there has been concern that observers cannot use such a scale effectively as they may be unable to distinguish between a large range of rating points (*17,19*) or have poor precision in using the scale (*19*). Furthermore, rebinning continuous data to a discrete scale of 1–5, 1–10, or 1–20 equal bins has not been found to generate significantly different results (*19*). Eleven-point or 21-point scales are considered adequate for reliable ROC curve fitting (*19*). The overriding value of multicategory rating scales is the ability to acquire information at a variety of thresholds, with a higher rating always indicating a higher level of suspicion. The selection of an appropriate rating scale is an important consideration when an ROC study is planned.

TABLE 2
Examples of Terminology Used in Discrete Rating Scales to Help Observers Classify Their Confidence and Use of Quasi Continuous Rating Scales

Observer	ROC confidence scale
Discrete	
Chesters (*12*)	0–5 (definitely not signal to definitely signal)
Langoltz (*43*)	0–5 (definitely benign to definitely malignant)
Obuchowski (*44*)	0–4 (normal to malignant)
Quasi continuous	
Chakraborty (*3*)	0–100 (101-point scale)
Rockette et al. (*17*)	0–100 (101-point scale)
Wagner et al. (*18*)	0–100 (101-point scale)
Hadjiiski et al. (*19*)	0–100 (101-point scale, rebinned to 0–5, 0–10, 0–20)

THE VISUAL SEARCH: DEVELOPMENT OF LOCATION-BASED ANALYSIS

According to the radiologic search model, image viewing begins with a global analysis of information collected by the observer's peripheral vision. Suggestive areas are then granted further interpretation, during which the radiologists must decide whether they are seeing signal (lesion) or noise (no lesion). The visual search has often been focused on detection, but this is complicated in radiology since the number of signals (lesions) is unknown (*20*). An observer's evaluation of an image can be divided into 3 phases—detection, recognition, and interpretation (*21*)—with the concern of misdiagnosis if the visual search is inadequate (*12*). These are considerations in image evaluation, and one of the drawbacks of conventional ROC methods is the inability to take into account location information and multiple lesions. Conventional ROC analysis would allow an observer to misinterpret an image but still arrive at what could be considered the correct answer (*21*). For example, a low-conspicuity lesion could be overlooked and a lesion mimic seen and scored, with the observer deemed to score a true-positive result. This type of assumption, as described in Figure 2, is clearly unacceptable, and location-based methods have been developed to overcome this problem.

Incorporating localization information into the analysis is a valuable way of increasing the statistical power of the test in comparison to conventional ROC studies. Location ROC (LROC) and free-response ROC (FROC) are the 2 increasingly popular variations of analysis that allow localization information to be incorporated.

LROC ANALYSIS

LROC was the first location method to be developed that could predict observer performance as a result of both detection and localization (*22*). This was a significant step in overcoming the failings of ROC to deal with multiple targets and location. LROC requires observers to localize and score a single region in an image that they deem to be the most suggestive (*23*). The LROC curve is then described as the ability to detect and correctly localize actual targets in an image. The first realization of this analysis was described by Swensson in 1996 (*24*). In this type of detection task, experimental control is strongly influential since the method requires that all images contain either one lesion or no lesions (*1,3*). This task does not necessarily reflect a true detection task, as a radiologist would not terminate the search of an image after the localization of a solitary lesion. In a clinical setting, the visual search would continue until the radiologist was satisfied that the image did not contain any further noteworthy abnormalities.

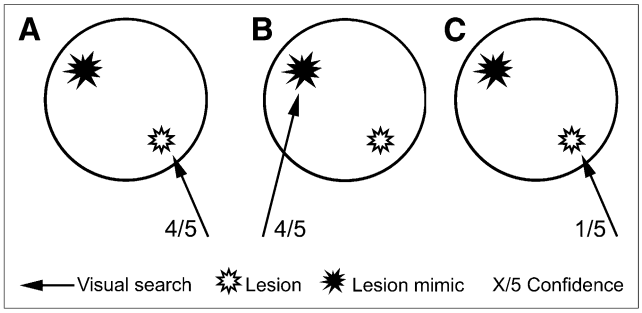

FIGURE 2. Potential for error in visual search task in conventional ROC methods. (A) One observer finds lesion (true) and scores confidence 4. (B) Another observer finds lesion mimic (false) and scores confidence 4. (C) Third observer finds lesion and scores confidence 1. Observers A and B are treated as equals by ROC analysis; observer C is penalized for having lower confidence. LROC and FROC account for this error type, and observer B would be penalized with false-positive (nonlesion localization).

FROC ANALYSIS

The FROC paradigm overcomes the limitations of LROC, allowing the localization of multiple abnormal areas within a single image (*3,25*). The observers can localize any area within the image that they deem to be abnormal, thus allowing both correct and incorrect localizations to be made on a single image (*3*). This paradigm is currently considered the most representative of the clinical situation. However, precisely localizing multiple lesions within an image can be a challenging perceptual task (*12*).

ACCEPTANCE RADII

The localization must be within what is considered a clinically acceptable distance from the true lesion (*3*). Those localizations within this acceptable distance are known as lesion localizations, and those outside are known as nonlesion localizations (*26*). Each localization is accompanied by a confidence score leading to the production of mark-rating pairs for each observer decision.

The clinically acceptable distance that determines mark-rating pairs as lesion or nonlesion localization is known as an acceptance radius or proximity criterion (*20*). The acceptance radius will allow a slight error in localization, with the radial size (from the center of the lesion) a predetermined value appropriate to the study. Different sizes of acceptance radius have an impact on the figure of merit (FOM) calculated at analysis, therefore showing a different measure of observer performance. Less strict radii can lead to an inflated FOM (*27*). The size of the acceptance radius should be a careful consideration in study design, with a significant effect on the classification of mark-rating pairs as either lesion or nonlesion localization (*25*). This disease process should have a strong influence on this decision. A potential error that can occur with an inappropriate acceptance radius is shown in Figure 3.

Since all observer decisions (lesion and nonlesion localizations) are accounted for when FROC methods are used, the observer is penalized for localizations of mimic lesions in addition to being rewarded for localizations of true lesions. Furthermore, overlooked lesions also contribute to a decrease in diagnostic accuracy.

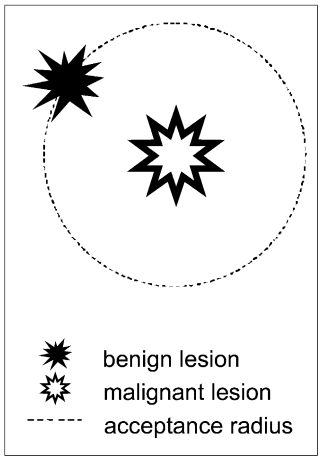

FIGURE 3. Size of acceptance radius can influence classification of mark-rating pairs. If acceptance radius of low-conspicuity malignant lesion overlaps high-conspicuity benign lesion, perceptual hit could occur (*25*), being classified incorrectly as lesion localization.

ANALYZING FROC DATA

Alternative FROC (AFROC) analysis can be used to analyze the data acquired from FROC studies, assuming independence between lesion localization and nonlesion localization mark-rating pairs while using only the highest nonlesion localization score if there are multiples on the same image (*3*). This highest-scoring nonlesion localization is used to create an equivalent ROC rating that can calculate the false-positive fraction (*26*). The AFROC curve, produced as a result of this analysis, is a plot of lesion localization fraction against false-positive fraction, with the area under the AFROC curve frequently used to define lesion detectability (*12,25*).

Jackknife AFROC (JAFROC) is a statistically powerful analysis method for FROC data that uses the area under the AFROC curve as the FOM (*7*). The JAFROC FOM, showing superior statistical power to ROC AUC (*33*), defines the probability that lesion localization marks on abnormal images are rated more highly than nonlesion localization marks on normal images (*26*). Nonlesion localization marks on abnormal images are ignored. The JAFROC FOM compares lesion ratings and ROC equivalent ratings of normal images (*26*). For studies with no normal images, the JAFROC-1 FOM can be used to analyze data for which the most highly rated nonlesion localization on each image is included in the analysis (*26*). Software and instruction on performing JAFROC analysis can be found online (*28*). For FROC studies, allowing multiple nonlesion localization decisions is thought to yield increased statistical power over conventional ROC methods (*3*).

STATISTICAL POWER FOR FROC STUDIES

Observer performance studies must yield good statistical power. An adequate sample size can give confidence in the reliability of the conclusions, whereas an underpowered study can cast doubts over the outcomes (*29*). Typically, large numbers of observers can be required for multireader, multicase studies, in which the aim is to reduce the potential impact of interobserver variation and increase statistical power. In some instances, the effect of interobserver variation has been found to be at least as significant as the difference in modality performance (*30*). Nevertheless, it is often desirable to keep the numbers of observers and cases to a minimum such that the time commitment of the observers and the costs of the study can also be kept to a minimum (*29*).

The importance of statistical significance should encourage investigators to perform a sample size calculation during the developmental stages of an observer performance study.

Dorfman, Berbaum, and Metz developed a multireader, multicase method for analyzing data acquired using a jackknife method (*31*). This method can account for significant differences in observer performance as a result of a change in imaging modality while also indicating that the observed effect may not be the same for all observers (*31*). This form of statistical analysis uses the Wilcoxon statistic (*31*) and can be successfully used for readers interpreting the same set of images obtained by 2 or more modalities (3).

An alternative method, which uses Dorfman–Berbaum–Metz multireader, multicase methods, is JAFROC. Suited to data produced from mark-rating pairs, this analysis enables the investigator to perform a sample size calculation based on the data used in analysis. The desired effect size (P value, i.e., $P < 0.05$) will contribute to determining the number of observers and cases required for optimal statistical analysis.

The P value represents the probability of finding a difference between 2 tests in a population of cases that contains no difference (*12*). For a statistical power calculation, the sources of variation (observers and cases) must be considered such that a meaningful conclusion can be drawn. Variance figures for observers, treatments, and cases are produced in JAFROC analysis, and these can then be entered into the sample size calculator for an accurate estimate of the number of observers and cases required for optimal statistical power. The assumptions used in this calculator (available at www.devchakraborty.com/downloads) are based on the type of sample used (random observers/cases or not) and the desired effect size and statistical power. In observer performance studies, it has been conventional to aim for high statistical power (0.8/80%) (*32*) and an effect size of $P < 0.05$ (probability of 5% that the significant difference is due to chance).

As an example, for a multireader, multicase free-response study to be analyzed using the JAFROC FOM, it is possible to perform a calculation, based on the number of observers completing the proposed study and the desired effect size (i.e., $P < 0.05$), to reveal the number of cases required for optimal statistical analysis. Although the statistical power of free-response studies increases with more lesions per image, investigators performing phantom simulations or adding simulated lesions to clinical images must be wary of exceeding what would be clinically realistic.

The increasing popularity of JAFROC FOM is in part due to the increased statistical power of observer studies

analyzed with AFROC analysis compared with traditional ROC analysis. The increased power is thought to be due to the consideration of location information (*3*).

A table of optimal sample sizes has been produced (*15*) as an alternative method for estimating adequate sample size, which can be useful for study planning. There are also several other points that must be considered during study development. It is important to have the correct case mix, reflecting the clinical presentation and prevalence of the disease process under investigation. It is also relevant to inform the observer of the range of lesions (size, shape, density, and average number) before beginning the study. Furthermore, the conspicuity of lesions and the difficulty in localization need to be tightly controlled such that acceptable numbers of nonlesion localization marks are made on normal images (*32*). For optimal statistical power, it is also desirable to present the observers with an equal ratio of normal and abnormal images that have been classified by a gold standard. The sample of observers enrolled into the study should also be representative of the population as a whole.

FROC STUDY EXAMPLE: DOSE REDUCTION IN HYBRID IMAGING

The following example will describe how FROC methods can be used in image optimization and dose reduction. Consider the optimization of the CT component of a hybrid scanner. With an aim to reduce dose, there is always a concern that the image will deteriorate to a level that is not clinically acceptable because dose reduction reduces the diagnostic performance of the observer. The impact of radiation dose on image noise is well documented, but the impact of increasing image noise on an observer's ability to arrive at the correct diagnosis, over a range of cases, is less clear. Before a clinical study, it may be desirable to perform a phantom study to simulate the effect of a reduced dose on image quality and lesion conspicuity under controlled conditions, where lesion size and distribution are known exactly.

The CT component of hybrid systems can be used to aid accurate anatomic localization of lesions (suspected or incidental) for many examinations. The high inherent contrast between lung parenchyma and lung lesion within the thorax allows a low radiation dose (tube current) to be used to provide images of acceptable quality. Therefore, it would be interesting to compare a high-dose (low-noise) acquisition with a low-dose (high-noise) acquisition for accurate localization of lesions within the thorax. Evidence in the literature of this type of optimization in hybrid imaging is sparse.

A statistical power calculation determines that approximately 220 random cases would be suitable for a sample of 5 observers. At this stage, the research team must select a suitable confidence scale and decide on a suitable image-viewing regime. It has been suggested that 40 min is a suitable length of time for each image observation session (*33*) and viewing conditions should be consistent for all observers.

Images for the high-dose and low-dose CT acquisition should be acquired such that the same lesions are shown in the same position (case-matched) and then displayed in a randomized order to avoid case memory. After observation, JAFROC analysis could be performed to produce AFROC plots and JAFROC FOM values for each modality (high-dose and low-dose). If P is greater than 0.05, the null hypothesis (no statistical difference) must be accepted. In the case of image optimization, this result would be desirable, showing that the diagnostic performance (lesion detection) at a low dose is equal to that at a high dose.

The suitability of FROC methods for dose optimization in hybrid imaging has recently been described. To analyze the FROC method, CT acquisition parameters suitable for use in SPECT/CT were evaluated (*34*). Looking at variations of pitch and amperage, it was found that the lowest-dose (highest pitch/lowest amperage) CT protocol allowed lesion detectability, within an anthropomorphic chest phantom, at confidence equal to that of higher-dose protocols. A natural extension of this work led to the evaluation of the 4 clinically available amperage settings on the Infinia Hawkeye 4 SPECT/CT scanner (GE Healthcare). This work also found that lesion detectability was equal at the 4 dose settings (*35*). In both of these phantom simulations, there was evidence to suggest potential dose savings in patients.

In addition to dose optimization, FROC methods have been used in hybrid imaging to assess the lesion detectability of different SPECT/CT systems (*36*). This phantom simulation showed that the CT attenuation correction images as would be acquired for myocardial perfusion imaging were of vastly different qualities, thus allowing significantly different lesion detection rates (*36*).

FROC STUDY EXAMPLE: MODALITY COMPARISON

ROC and FROC methods have been more frequently used to compare directly the success of 2 or more tests. In nuclear medicine the evolution of PET/CT has seen conventional techniques, such as WBS, replaced as the gold standard in some instances. A typical example is shown in Figure 4, comparing a ^{99m}Tc-methylene diphosphonate bone scan with an ^{18}F attenuation-corrected bone scan. Many more lesions can be identified in the ^{18}F images than in the WBS images. However, visually it is difficult to quantify the advantage that ^{18}F holds over WBS, as must be done to evaluate the benefit of the new technique. If no statistical advantage is identified, there may be a reluctance to explore the new technique. The free-response method would be well suited to an evaluation of this type, where the observer could accurately localize all suggestive areas of the image to define lesion detection performance over a range of cases for the 2 modalities. A statistical evaluation would then reveal any advantage held by the ^{18}F technique.

DRAWBACKS OF ROC AND FROC

The advantages of FROC methods for evaluation of diagnostic accuracy are clear, but it is not a one-size-fits-all solution to observer performance. Conventional ROC methods

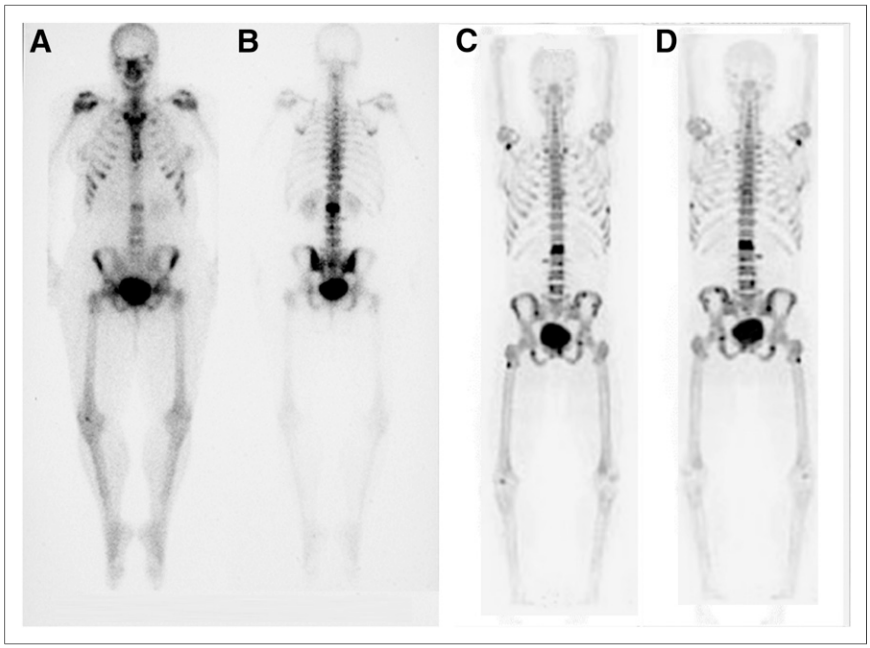

FIGURE 4. Comparison of ^{18}F (A and B) and ^{99m}Tc-methylene diphosphonate (C and D). Free-response study would have greater power than conventional ROC analysis in defining difference in diagnostic accuracy because of requirement to accurately localize lesions. (Reprinted from (*45*).)

still have a role in observer performance, particularly when a diffuse disease rather than focal disease is the central issue. For diffuse disease, classifying an image as normal or abnormal using ROC methods is acceptable, with FROC methods best saved for focal or multiple focal diseases (*23*). FROC methods can also suffer by not producing a clinically relevant answer. Although accurate localization is a good test of observer skill, it does not necessarily inform a clinician about the need for further diagnostic work-up (*25*), and even though the free-response method is the closest solution to clinical that is currently available, it is still not the real thing (*31*). Furthermore, it has been suggested that some of the information provided by either the ROC or the FROC paradigm can be irrelevant to the clinical question.

The free-response method has developed significantly over time, with methods of analysis changing and statistical validation ongoing as the paradigm evolves. Research in this area will continue to address these issues as the observer performance community strives to make these methods even more reliable.

ALTERNATIVE MEASURES OF OBSERVER PERFORMANCE

The 2-alternative forced choice procedure presents observers with pairs of images, one containing a signal (lesion) and the other no signal (*12*), with the observer forced to decide which image contains the signal. In this situation, observers can be highly sensitive to changes in image presentation during a side-by-side review (*37*). This method, similar to ROC, does not require the observer to keep a consistent decision threshold throughout the test (*12*). However, the 2-alternative forced choice does not provide information on the trade-off between true-positive and false-positive rates (*12*) and requires a greater number of observations to achieve the same accuracy. An excellent example of the value of the 2-alternative forced choice has been described by Good et al. (*38*), where radiologists successfully selected the high-quality format (4,000-pixel resolution as apposed to 2,000) of the same image when a difference in image quality had been thought to be imperceptible. The ROC AUC has been shown to equal the percentage correct in the 2-alternative forced choice (*12*), but one should also be mindful that FROC techniques provide greater statistical power than ROC.

The κ-method is based on a 2×2 square result (true-positive, false-positive, true-negative, and false-negative) and is used to measure the level of agreement between observers above what would be expected by chance. The κ-method is based on a binary observer decision that produces results of true-positive, false-positive, true-negative, and false-negative, which can also be used to calculate sensitivity and specificity. In theory, a binary decision can be statistically better than an ROC decision when information provided by ROC/FROC methods is irrelevant.

CONCLUSION

Observer performance assessment, in particular the FROC paradigm, can be highly valuable in the assessment of system performance and can be applied in the pursuit of image optimization and dose reduction. The increasing availability of ROC/FROC software programs to aid both the capture (*39,40*) and the analysis (*28,41,42*) of observer performance are making these techniques easier to implement. Combined with previous extensive research and guidance on performing observer studies, these techniques are providing great potential to optimize practice within nuclear medicine.

DISCLOSURE

No potential conflict of interest relevant to this article was reported.

REFERENCES

1. Vining DJ, Gladish GW. Receiver operating characteristic curves: a basic understanding. *Radiographics.* 1992;12:1147–1154.
2. Wang S, Chang CI, Yang SC, et al. 3d ROC analysis for medical imaging diagnosis. *Conf Proc IEEE Eng Med Biol Soc.* 2005;7:7545–7548.
3. Chakraborty D. Statistical power in observer-performance studies: comparison of the receiver operating characteristic and free-response methods in tasks involving localization. *Acad Radiol.* 2002;9:147–156.
4. Obuchowski NA, Beiden SV, Berbaum KS, et al. Multireader, multicase receiver operating characteristic analysis: an empirical comparison of five methods. *Acad Radiol.* 2004;11:980–995.
5. Gur D, Rockette HE, Bandos AI. 'Binary' and 'non-binary' detection tasks: are current performance measures optimal. *Acad Radiol.* 2007;14:871–876.
6. Manning DJ. Evaluation of diagnostic performance in radiography. *Radiography.* 1998;4:49–60.
7. Chakraborty DP. Recent developments in imaging system assessment methodology, FROC analysis and the search model. *Nucl Instrum Methods Phys Res A.* 2011;648(suppl 1):S297–S301.
8. Chisholm GD, Stone AR, Beynon LL, Merrick MV. The bone scan as a tumour marker in prostatic carcinoma. *Eur Urol.* 1982;8:257–260.
9. Imbriaco M, Larson SM, Yeung HW, et al. A new parameter for measuring metastatic bone involvement by prostate cancer: the bone scan index. *Clin Cancer Res.* 1998;4:1765–1772.
10. Obuchowski NA. Fundamentals of clinical research for radiologists: ROC analysis. *AJR.* 2005;184:364–372.
11. Metz CE. Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems. *J Am Coll Radiol.* 2006;3:413–422.
12. Chesters MS. Human visual perception and ROC methodology in medical imaging. *Phys Med Biol.* 1992;37:1433–1476.
13. Zou KH, Tempany CM, Fielding JR, Silverman SG. Original smooth receiver operating characteristic curve estimation from continuous data: statistical methods for analyzing the predictive value of spiral CT of ureteral stones. *Acad Radiol.* 1998;5:680–687.
14. Tourassi G. Receiver operating characteristic analysis: basic concepts and practical applications. In: Samei E, Krupinski E, eds. *The Handbook of Medical Imaging Perception and Techniques.* New York, NY: Cambridge University Press; 2010:187–203.
15. Obuchowski NA. Sample size tables for receiver operating characteristic studies. *AJR.* 2000;175:603–608.
16. Park SH, Goo JM, Jo CH. Receiver operating characteristic (ROC) curve: practical review for radiologists. *Korean J Radiol.* 2004;5:11–18.
17. Rockette HE, Gur D. Selection of a rating scale in receiver operating characteristic studies: some remaining issues. *Acad Radiol.* 2008;15:245–248.
18. Wagner RF, Beiden SV, Metz CE. Continuous versus categorical data for ROC analysis: some quantitative considerations. *Acad Radiol.* 2001;8:328–334.
19. Hadjiiski L, Chan H-P, Sahiner B, Helvie MA, Roubidoux MA. Quasi-continuous and discrete confidence rating scales for observer performance studies: effect on ROC analysis. *Acad Radiol.* 2007;14:38–48.
20. 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.
21. Hendee WR. The perception of visual information. *Radiographics.* 1987;7:1213–1219.
22. Starr SJ, Metz CE, Lusted LB, Goodenough DJ. Visual detection and localization of radiographic images. *Radiology.* 1975;116:533–538.
23. Chakraborty DP. Counterpoint to 'performance assessment of diagnostic systems under the FROC paradigm' by Gur and Rockette. *Acad Radiol.* 2009;16:507–510.
24. Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. *Med Phys.* 1996;23:1709–1725.
25. Gur D, Rockette HE. Performance assessments of diagnostic systems under the FROC paradigm: experimental, analytical, and results interpretation issues. *Acad Radiol.* 2008;15:1312–1315.
26. Chakraborty DP. Validation and statistical power comparison of methods for analyzing free-response observer performance studies. *Acad Radiol.* 2008;15:1554–1566.
27. Gur D, Bandos A, Klym A, et al. Agreement of the order of overall performance levels under different reading paradigms. *Acad Radiol.* 2008;15:1567–1573.
28. JAFROC 4.0.1 analysis software. devchakraborty.com downloads page. http://www.devchakraborty.com/downloads.html. Accessed April 9, 2013.
29. Chakraborty DP. How many readers and cases does one need to conduct an ROC study? *Acad Radiol.* 2011;18:127–128.
30. Obuchowski NA. Reducing the number of reader interpretations in MRMC studies. *Acad Radiol.* 2009;16:209–217.
31. Rockette HE, Campbell WL, Britton CA, Holbert JM, King JL, Gur D. Empiric assessment of parameters that affect the design of multireader receiver operating characteristic studies. *Acad Radiol.* 1999;6:723–729.
32. Chakraborty DP. Recent developments in FROC methodology. In: Samei E, Krupinski E, eds. *The Handbook of Medical Imaging Perception and Techniques.* New York, NY: Cambridge University Press; 2010:187–203.
33. Manning DJ, Bunting S, Leach J. An ROC evaluation of six systems for chest radiography. *Radiography.* 1999;5:201–209.
34. Thompson J, Hogg P, Szczepura K, Manning D. Analysis of CT acquisition parameters suitable for use in SPECT/CT: a free-response receiver operating characteristic study. *Radiography.* 2012;18:238–243.
35. Thompson J, Higham S, Hogg P, Manning D, Szczepura K. Accurate localization of incidental findings on the computed tomography attenuation correction image: the influence of tube current variation. *Nucl Med Commun.* 2013;34:180–184.
36. Thompson J, Szczepura K, Manning D, Hogg P. Lesion detection in the CT attenuation correction image of 5 different low resolution SPECT/CT systems: a multi-centre study [abstract]. *Nucl Med Commun.* 2012;33:548.
37. Gur D, Rubin DA, Kart BH, et al. Forced choice and ordinal discrete rating assessment of image quality: a comparison. *J Digit Imaging.* 1997;10:103–107.
38. Good WF, Gur D, Feist JH, et al. Subjective and objective assessment of image quality: a comparison. *J Digit Imaging.* 1994;7:77–78.
39. Thompson J, Hogg P, Thompson S, Manning D, Szczepura K. ROCView: prototype software for data collection in jackknife alternative free-response receiver operating characteristic analysis. *Br J Radiol.* 2012;85:1320–1326.
40. Svahn T, Andersson I, Chakraborty D, et al. The diagnostic accuracy of dual-view digital mammography, single-view breast tomosynthesis and a dual-view combination of breast tomosynthesis and digital mammography in a free-response observer performance study. *Radiat Prot Dosimetry.* 2010;139:113–117.
41. DBM MRMC software program. The Medical Image Perception Laboratory Web site. http://perception.radiology.uiowa.edu/Software/ReceiverOperatingCharacteristicROC/DBMMRMC/tabid/116/Default.aspx. Accessed April 9, 2013.
42. Metz ROC software general description. Metz ROC software Web site. http://metz-roc.uchicago.edu/MetzROC/software. Accessed April 9, 2013.
43. Langlotz CP. Fundamental measures of diagnostic examination performance: usefulness for clinical decision making and research. *Radiology.* 2003;228:3–9.
44. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. *Radiology.* 2003;229:3–8.
45. Bridges RL, Wiley CR, Christian JC, Strohm AP. An introduction to $Na^{18}F$ bone scintigraphy: basic principles, advanced imaging concepts. and case examples. *J Nucl Med Technol.* 2007;35:64–76.