Some Methodological Questions Concerning Receiver Operating Characteristic (ROC) Analysis as a Method For Assessing Image Quality in Radiology

Michael B. Harrington

This paper raises five methodological questions concerning receiver operating characteristic (ROC) analysis: (1) can the ROC "confidence criterion" be applied in a valid, reliable way?; (2) can ROC deal with ambiguous findings?; (3) can ROC deal effectively with false-negative findings?; (4) are ROC curves susceptible to valid statistical testing?; and (5) are ROC results useful in choosing among alternative imaging modalities? A review of the evidence leads to six conclusions. First, using ROC, all radiological findings must be unambiguously scored as true-positive, true-negative, false-positive, or false-negative, often forcing arbitrary, procrustean choices on readers and evaluators. Second, ROC requires radiologists to report findings by confidence level on a consistent, reliable basis throughout a ROC experiment: something that seems unrealistic, given what is known about human performance in almost all perceptual tasks of comparable complexity. Third, as gathered during the typical experiment, ROC data are probably nominal, but treated as if ordinal (or even interval) data, leading to distorted results. Fourth, ROC does not deal effectively with false-negatives, despite their importance. Fifth, there is no satisfactory method for statistically testing the significance of observed differences between two ROC curves if they are based on nominal data. Finally, the artificial tasks required of radiologists in a ROC evaluation limit the usefulness of ROC results in choosing among the imaging modalities.

Adapted and reprinted with permission from Society of Photo-optical Engineers.

KEY WORDS: image evaluation, digital radiography, image quality studies, receiver operating characteristic, expert opinion usage.

THE USE OF RECEIVER operating characteristic (ROC) analysis and its variants to evaluate the quality of medical images is steadily increasing.[1-6] Today, ROC is the principal technique using human observers for testing competing medical imaging modalities. ROC analysis seeks to elicit expert radiological opinion concerning the quality of radiological images in an objective, replicable way. In one of its most important applications, ROC analysis is used to determine which of two or more imaging modalities produces superior images. It does this by measuring the ability of observers to identify pathologies using each modality. If observers obtain better ROC curves (Fig 1) with one modality than with others, the imaging modality used to get the better curve (ie, the one nearer the upper left hand corner of Fig 1) produces the better images.

SEPARATING THE OBSERVER FROM THE OBSERVED USING ROC

ROC seeks to separate the inherent ability of an imaging modality to portray various states of disease and health from the balance struck (consciously or otherwise) by radiologists between false-positive (FP) and false-negative (FN) errors due to variations in reporting "style."[7-9] "Inherent ability" can be measured objectively if the human observer is excluded. One can measure a number of physical characteristics of the imaging system that contribute to image quality (Table 1). But the relationship between measurable characteristics of an imaging system and a radiologist's ability to identify pathologies in the images it produces is poorly understood.[10-12] Changes in the former do not affect the latter in a simple way. Reporting style is known to vary among individuals, but it is one thing to acknowledge these differences, another to use them to appraise image quality. ROC separates inherent ability of the system from radiologist reporting style, and deals only with the latter. In Table 1, ROC analysis is used in the kinds of studies shown in the bottom three rows.

AN OVERVIEW OF ROC METHODOLOGY

A typical ROC image evaluation might take the following approach. Two imaging modalities are compared using ROC as the measure of image quality. A sample of cases is selected, and imaged in both modalities. Each radiologist using a given imaging modality examines images

From the Civil Systems Division, The MITRE Corp, McLean, VA.

Address reprint requests to Michael B. Harrington, PhD, Civil Systems Division, 7525 Colshire Dr, McLean, VA 22102-3481.

Adapted and reprinted with permission from Society of Photo-optical Instrumentation Engineers.

0897-1889/90/0304-0008$03.00/0

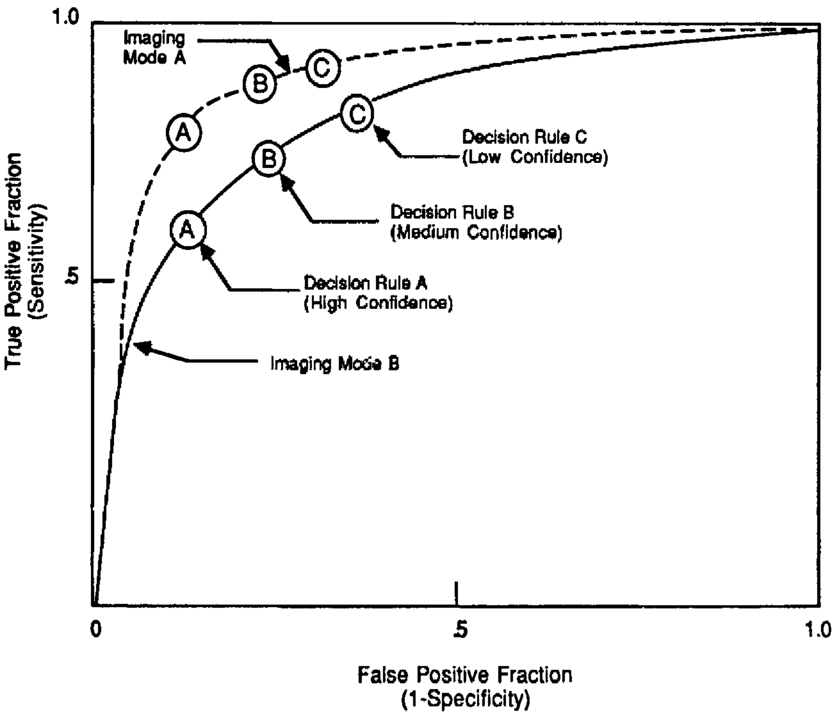

Fig 1. Illustrative ROC curves comparing two hypothetical diagnostic image devices.

from patients who are disease-free and from patients who have pathological conditions or whose images have been altered to contain an artificial or "phantom" pathology. Radiologists read half the cases using one modality, the other half using the other modality. No case is read twice by the same radiologist.

Each radiologist reports his or her findings at each of several levels of confidence (or its functional equivalent phrased in terms such as "definitely positive," "probably positive," and the like). Reports at three levels of confidence are common but as many as seven have been used. The results are scored for accuracy in some authoritative way, such as through pathology reports, surgical results, alternative imaging techniques, or consensus panels with access to additional nonradiological information on the patients' conditions. The results obtained with each modality are arrayed as shown in Table 2. The true-positive fraction (TPF) and false-positive fraction (FPF) can be derived from Table 2 and plotted as shown in Fig 1. Metz[7,13,14] provides a lucid discussion of the methodology.

Table 1. Some Major Methods for Assessing Image Quality

Method	Examples of Variables Measured
No observer used	
Pixel density	Pixels per unit area of the image
Gray scale differentiation	Gray scale values per pixel available between white and black
Performance of image capture mechanism	Resolution, contrast sensitivity, gray scale distortion, stability from image to image, signal transfer function
Pixel compression comparison	Differences between compressed and uncompressed pixel value distributions for test images
Display system performance	Spatial resolution, signal/noise ratio, dynamic range, size of display area, refresh rate, flicker
Observer used	
Geometric pattern detection	Accuracy in detecting subtle geometric patterns inserted into images with backgrounds of noise
Phantom image detection	Accuracy in distinguishing "pathologies" inserted into images from otherwise normal patients
Real pathology detection	Accuracy in distinguishing pathologies in images selected from real patient cases

SOME METHODOLOGICAL QUESTIONS CONCERNING ROC ANALYSIS

The need ROC purports to fill is clear. Whether ROC can fulfill this need at its present stage of development is still at issue. The following ques-

Table 2. Basic Data on Reporting Accuracy Collected During an ROC Experiment Using Radiologists Viewing Images

	Number of Positive Findings Reported by Test Radiologists	Number of Negative Findings Reported by Test Radiologists	Total
Correct positive findings	TP	FN	ToTP
Correct negative findings	FP	TN	ToTN
Totals	Total positives reported	Total negatives reported	Total findings

Abbreviations: TP, true-positive; TN, true-negative; ToTP, total-true-positives; ToTN, total-true-negatives.

tions are pertinent: (1) can the ROC "confidence criterion" be applied in a valid, reliable way?; (2) can ROC deal with ambiguous findings?; (3) can ROC deal effectively with FN findings?; (4) are ROC curves susceptible to valid statistical testing?; and (5) are ROC results, as currently produced, useful in choosing among alternative imaging modalities? These questions are most important for the ROC studies summarized in the bottom two rows of Table 1. Each of these questions is discussed below.

Can the Confidence Criterion be Applied in a Valid, Reliable Way When Reporting Findings?

ROC studies in the literature define the confidence criterion in various ways. Radiologists may be asked to report likelihood of a pathological finding, or the degree to which a finding is believed to exist. At bottom, all such formulations require reporting how much confidence the radiologist has that a finding actually exists, based on the image(s) examined. But confidence is not a simple concept. Radiology is perceptually complex, and the ingredients in a given level of radiologist confidence is equally complex. Radiologists vary in terms of such obvious characteristics as training, experience, and skill. They also vary in terms of other characteristics that have been widely acknowledged, but whose impact on accuracy and confidence is poorly understood. These include perceptual skills; conceptual constructs concerning pathologies and their radiographic appearance; implicit weighting schemes used in evaluating radiographic evidence; the amount of experience with diseases (common and rare) encountered in day-to-day practice; attitudes toward the role of radiological reports in the total spectrum of diagnostic techniques available; reactions to environmental conditions during a reading session; ability to perform under various levels of fatigue; and probably many others.[11,12,15,16] All can affect confidence in findings; all operate in imperfectly understood ways, both individually and collectively.

Confidence is also affected by the kind of perception required. The literature distinguishes three progressively inclusive kinds of perception: (1) detection—is a visual stimulus seen by the observer? (2) localization—can its location be placed correctly by the observer in the field of view? (3) recognition—can the stimulus be named and described correctly by the observer? There is much experimental data on observer performance on the first two perceptual tasks.[7-9] But a radiologist reviewing images in an ROC evaluation employs the third type of perception, the most complex of the three. Experimental results involving recognition are far less complete, but do confirm that there are many ingredients besides image quality contributing to the radiologist's ability to identify medically significant findings correctly.[11,12] Moreover, knowledge about factors contributing to the confidence they have in this ability is sketchy.

Against this background, the confidence criterion (in most of its ROC forms) can be questioned on four grounds: (1) its meaning can shift in an uncontrolled way both within and among radiologists; (2) there is no method for assuring that the criterion will be applied reliably, even if uniformly understood; (3) there is little recognition by ROC of the need to define "confidence" precisely; and (4) there seems to be no single, unidimensional scale for measuring the results of its use. Each is discussed below.

Can "Confidence" Be Limited to a Single Operational Definition?

Factors producing a given confidence level for a particular finding can vary widely. Over the past 8 years, for example, the MITRE Corporation has conducted image quality evaluations involving over 1,000 cases (nearly 2,500 individual images) chiefly comparing plain films against digitized images.[28] These cases were reviewed by consensus panels of at least two and as many as four Board-certified, faculty radiologists working together. The possible meaning of "confidence" in the mind of a working radiologist was subject to recurring discussion during these evaluations. Drawn from these discussions, the following possible meanings of a "high" confidence report by a radiologist are (1) that the image depicts the finding clearly (ie, the image quality is good), (2) the finding is a pathology that is clearly apparent whatever the quality of the image (ie, the finding is obvious) (3) the finding is common among patients of this sex, age, history, and presenting complaint (ie, the finding is both suggested and supported by extraimage information), (4) the finding could be one of two or more things, one far more common in patients like the present one (ie, the finding's existence is judged against its estimated likelihood of occurrence), (5) the finding is serious, if actually present, and should be verified in other ways (ie, the existence of the finding is judged against its potential risk to the patient), (6) the finding is a pathology seen many times and, therefore, is readily recognized even in images of mediocre quality (ie, the finding is familiar), (7) the finding is one which, if not reported, might be a factor in a subsequent malpractice suit (ie, the finding is good defensive medicine), (8) though not especially well depicted, the finding is usually present when a related, well-depicted finding is evident (ie, the finding is part of a common syndrome in patients like the present one), (9) the pathology is apparent on a familiar viewing medium though it might not be on a new medium (ie, the finding is apparent, in part, because the viewing medium is familiar), (10) the finding might be missed by others, but I can spot the tough ones (ie, the confidence of the radiologist concerning his or her professional skill is high).

These variations, which do not exhaust the possibilities, may crop up individually or in combination from image to image, case to case, and radiologist to radiologist. Yet only the first interpretation of "confidence" indicates that the reader reported a finding with high confidence because he or she saw a pathology on a high quality image. Which definition (or definitions) are in the radiologist's mind when reporting each finding? Which should be? Surprisingly, this subject receives little attention in the ROC literature. The assumption often seems to be that each radiologist will understand the definition of confidence properly and in the same way.

How Reliably Can the Confidence Criterion Be Applied?

"Confidence" has no objective, unvarying scale to which the radiologist can refer from time to time to be sure his or her thought processes have not drifted. Even if each radiologist starts an ROC experiment with the same definition in mind, variation can be expected. Variation can occur from one time to another in the same radiologist, from one time of day to another, from early in the reading session to late, from one finding to another, from one image to another, from one case to another, from one kind of pathology to another, and from one run of cases during a given session (eg, predominantly difficult) to another (eg, mostly easy). Needed is a method for periodic revalidation of confidence levels during a ROC study. This revalidation would help make sure readers are employing the same definition in a consistent fashion. Most users of ROC simply seem to assume "confidence" can be applied reliably.[9,17-20] A few acknowledge that no method of standardizing subjective criteria is yet available.[13,25]

Must "Confidence" be Defined Precisely?

Some ROC advocates argue that "confidence" is valuable because it subsumes the above variations in a single measure. Nearly any professionally defensible confidence criterion can be used, it is suggested, so long as it is the criterion a radiologist would ordinarily use. After all, an ROC curve simply plots two proportions against one another—the TPF against the FPF in Fig 1. Any point on the ROC curve is the combination of these results obtained by readers when they employ a given subjective criterion. It is not crucial that these criteria be well defined, say some ROC advocates. It is sufficient that they are different from one another.

This argument seems to allow the radiologist to use factors only tangentially related (or even unrelated) to image quality when arriving at the confidence associated with a finding. But if ROC experiments focus on the quality of the images, then surely this must be the focus of the data they gather. Reports based on other factors may reveal differences among radiologists; they reveal little about differences in image quality.

Is "Confidence" Measurable on a Usable Scale?

Measurement theory distinguishes four kinds of scales (Table 3): (1) nominal, which uses numbers or other symbols to classify phenomena into mutually exclusive categories; (2) ordinal, which ranks phenomena in terms of a single characteristic; (3) interval, which ranks phenomena but also measure the differences among the ranks; and (4) ratio, which measures phenomena in interval terms but has a true zero point at the scale's origin. The elaborate mathematical analyses of ROC curves developed in recent years presume that ROC curves are based on data of at least ordinal quality.[17-21] Is this presumption realistic?

At a minimum, ordinal scales must (1) establish mutually exclusive subclasses within some class of quantities based on a single property of interest; (2) rank each relative to the others using this single property; and (3) preserve the relationship among subclass members in terms of the property being scaled. "Confidence," as usually employed in a ROC study, is suspect in terms of all three requirements. Do all radiologist reports at a given confidence level fall in the same subclass on a confidence scale defined in terms of a single property of interest (eg, quality of image)? This seems unlikely; each radiologist is free to think of "confidence" in his or her own way. Are confidence levels ranked according to a single dimension? This also seems unlikely; there are numerous possible dimensions in the mind of the reporting radiologist. Are reports from different radiologists at a given confidence level the same in terms of the entity being scaled? Again, "confidence" is likely to have a slightly different meaning from one radiologist to another, unless stringent experimental controls are in place, controls that do not accompany the typical ROC study. Thus, it seems likely that ROC produces "confidence" data that are nominal at best, that is, data at the most primitive level of measurement. Confidence scores, as typically gathered in a ROC evaluation, may be simply different from one another, not points on a single scale.

Dealing with Ambiguous Findings

ROC deals only in findings that can be scored unambiguously. But what about findings that are not clearly positive or negative? Included under this heading are observed phenomena that may or may not be the early stage of a disease, may or may not be an extreme but otherwise normal aspect of human anatomy, and the like. Such "intermediate, indeterminate or uninterpretable diagnostic test results"[27] are common in radiology and in other medical fields. More important, cases with such findings are often selected for image quality studies precisely because they contain subtle findings posing a significant challenge. In the MITRE studies cited earlier, such findings constituted nearly a quarter of the total reported.[28] Such findings often engendered considerable debate among consensus panel radiologists concerning the proper classification (ie, should a given finding be classified as positive,

Table 3. A Comparison of Possible Measurement Scales for Use With the ROC Confidence Variable

	Measurement Scales		
Major Characteristics	Nominal	Ordinal	Interval
Purpose	To classify subjects	To rank subjects	To measure subject's dimensions
Unit of measure	Integers or qualitative categories	Integers or qualitative categories	Real numbers
Defining relations	Equivalence only	Equivalence, greater than	Equivalence greater than, ratio of any two intervals
Intervals on scale measured	Scale expresses only mutual exclusivity	Scale expresses rank order	Scale measures intervals
Examples of appropriate statistics	Mode, frequency	Median, percentile	Mean, standard deviation
Probability distributions applicable for statistical purposes	Some nonparametric distributions	Most nonparametric distributions	All parametric distributions
Arithmetic operations applicable to differences on scale	None	None	Usually all

Data from Siegel.[29]

Table 4. Revised Data on Reporting Accuracy Collected During an Image Quality Evaluation Using Radiologists Viewing Images

	Number of Positives Reported by Test Radiologists	Number of N-Ns Reported by Test Radiologists	Number of Negatives Reported by Test Radiologists	Total
Positive findings	TP	FN-N	FN	ToTP
Non-positive non-negative findings	FP	TN-N	FN	ToTN-N
Negative findings	FP	FN-N	TN	ToTN
Totals	Total positives reported	Total N-Ns reported	Total negatives reported	Total findings

Abbreviations: N-N, nonpositive-nonnegative finding.

negative, or simply eliminated from the study due to an inability to reach agreement?).

How does ROC deal with such findings? At its present state, ROC can deal with only four categories of results (ie, true-positive[TP], false-positive [FP], true-negative [TN] and false-negative [FN]). Ambiguous findings must be forced into these categories. But since ROC studies do not establish rules for doing this (or even acknowledge that it must be done), both readers and scorers are left to their own devices. Imprecision, misinterpretation and arbitrary judgments creep into the evaluation. What is needed, say some students of the problem, is an expansion of the basic ROC table from four to nine cells as shown in Table 4.[27] This expansion would allow findings to be scored properly. But if adopted, this scheme would require significant modifications to ROC. The ROC methodology cannot deal with results scored in nine categories in its present form.

Dealing with False-Negative Findings

The ROC curve expresses accuracy in detecting pathologies, juxtaposing the TPF against the FPF. But what about FNs? In some respects, an imaging system's ability to avoid misses should be of greatest interest. A system that portrays a pathology badly is usually preferable to one that does not portray the pathology at all. It would be desirable, therefore, to plot TNF against FNF. Figure 2 illustrates the concept.

But such a curve cannot be plotted accurately, because the confidence levels reported for FNs in an ROC evaluation refer to the wrong subject. A radiologist cannot record a confidence level for a finding if unaware of the finding. Nor can the proper confidence level associated with a missed finding be deduced accurately from the other data reported by the radiologist. For example, suppose an exam is reported as normal with medium confidence, but a pathology actually

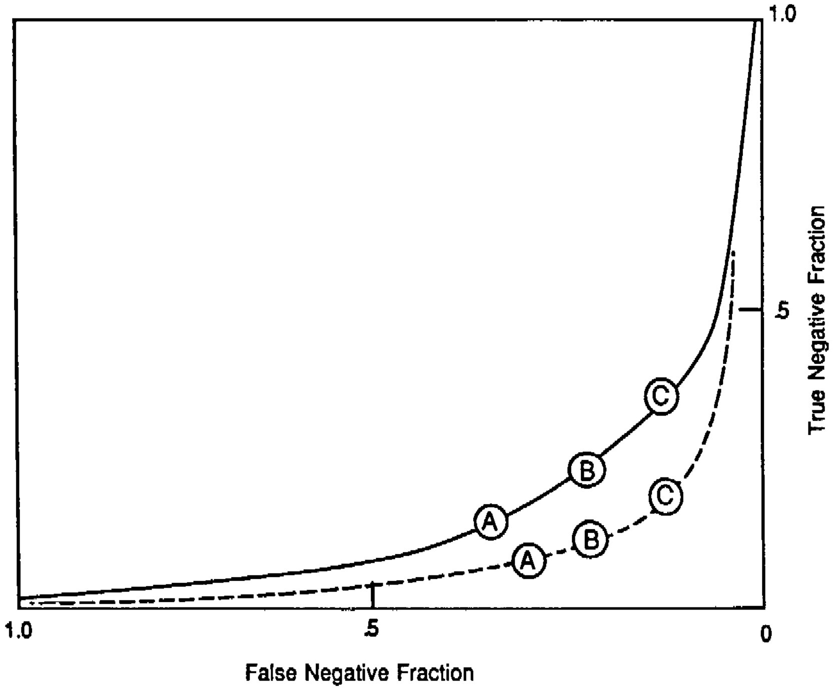

Fig 2. A hypothetical ROC curve for negative findings.

exists. In such a case, the missed pathology is scored as an FN. Is it valid to attribute "medium confidence" to the missed finding? The answer is almost certainly no. Under ROC reporting rules, the radiologist reports confidence levels only for a finding actually seen, or for a finding of normality. But seeing nothing with a given confidence level is not the same, for image quality purposes, as seeing something with that confidence level. As a result, ROC analysis is largely silent (or misleading) on one of the most important aspects of an imaging system's performance—the ability to avoid misses.

Testing the Difference Statistically

ROC aspires to testing the difference statistically between two imaging modalities, and should aspire to making statistical inferences to broader situations. Statistical testing requires setting up testable hypotheses, selecting tests, establishing power and significance criteria, and determining sample sizes. For the most part, studies in the ROC literature do one of three things: (1) acknowledge that there is no satisfactory test available; (2) compare the area under each ROC curve and compute related quantitative indices, as if the data points comprising the curves were scalable values; or (3) Ignore this issue.[19-22] There seems to be no comprehensive, statistically defensible test available for comparing the overall shape of two or more ROC curves. Since no test is available, there is also no precise way to determine the sample size or power and significance criteria needed to produce satisfactory test results.

DOES ROC IMPROVE CHOICE AMONG IMAGE MODALITIES?

ROC requires radiologists to read cases in a highly artificial way. Instead of reading cases as they ordinarily would, radiologists must use contrived decision criteria not used in day-to-day practice. This means ROC itself significantly alters radiologists' normal behavior, reducing the predictive power of its results. Moreover, acknowledging (as ROC correctly does) that decision making style varies among large numbers of radiologists is one thing. Simulating this by varying confidence levels within small numbers of radiologists is another. There is little reason and no evidence to believe the latter reveals much about the former. Therefore, ROC results are an uncertain guide to what could be expected from an imaging modality in ordinary practice on a large scale.

METHODOLOGICAL WEAKNESSES OF ROC ANALYSIS AT PRESENT

To summarize, six methodological weaknesses in ROC must be strengthened or compensated for if ROC is to be fully useful in comparing imaging modalities. First, ROC requires that radiologists report findings by confidence level on a consistent, reliable basis throughout an ROC experiment. This requirement seems highly unrealistic, given what is known about the vicissitudes of human performance in almost all perceptual tasks of similar complexity. Second, as gathered during the typical experiment, ROC data are nominal, but treated as if ordinal (or even interval) data arrayed on a unidimensional, normally distributed scale. Thus, what ROC curves actually portray in empirical terms is often largely unknown. Third, using ROC, all radiological findings must be unambiguously scored as TP, TN, FP, or FN. This forces arbitrary, frequently procrustean choices on reader and evaluator alike. Fourth, ROC does not deal effectively with FN results despite their importance. Fifth, there is no satisfactory method for testing statistically the significance of observed differences between two ROC curves if they do not employ ordinal (or interval) data. Sixth, the artificial reporting procedures ROC uses to simulate differences among radiologists are likely to fail at two key tasks—simulating differences in reader style, and predicting the probable performance of an imaging modality in day to day practice. Partly for these reasons, new approaches to evaluating image quality using human observers are being developed.[30,31]

NEXT STEPS

Remedying the methodological weaknesses noted above is challenging. Until remedies are available, ROC remains less than adequate for evaluating alternative imaging modalities. The main requirements for workable alternatives are apparent, however. Image quality evaluations using radiologists should (1) avoid uncontrolled "confidence level"reporting by radiologists, but should include instead a structured diversity of radiologists to incorporate the "reporting style" variable; (2) use the full range of scoring outcomes (shown in Table 4) so that a modality's strengths and weaknesses in portraying all diagnostic outcomes will be revealed; (3) employ established non-parametric statistical methods dealing with the entire range of diagnostic outcomes, not just the FPF and TPF, methods for which power, significance and sample size calculations can be made precisely; and (4) incorporate the physical characteristics measurements summarized in Table 1 to the maximum extent practicable.

REFERENCES

1. Chakraborty DP, Breatnach ES, Yester MV: Digital and conventional chest imaging: A modified ROC study of observer performance using simulated nodules. Radiology, 158:35-39, 1986

2. Getty DJ, Pickett RM, D'Orsi CJ, et al: Enhanced interpretation of diagnostic images. Invest Radiol 23:240-253, 1988

3. Kelsey CA, Mettler FA, Jr: ROC analysis can reveal best diagnostic method. Diagn Imag 11:155-161, 1989

4. Marioka C, Brown K, Hayrapetlan A, et al: ROC Analysis of Chest Radiographs Using Computed Radiography and Conventional Analog Films. The UCLA PACS Modules and Related Project, Medical Imaging Division, Department of Radiological Sciences, UCLA, 34-37, February 1989

5. McMahon H, Vyborny CT, Metz CE, et al: Digital chest radiography: Effect on diagnostic accuracy of hard copy, conventional video, and reversed gray scale video formats. Raiology 168:669-673, 1989

6. Seeley G, Newell JD: Use of Psychophysical Principles in the Design of a Total Digital Radiology Department. Radiologic Clinics of North America, Philadelphia, PA, Saunders, 23:341-348, 1985

7. Metz CE: Basic principles of ROC analysis. Semin Nucl Med VIII:283-298, 1978

8. Swets JA, Tanner WP: A decision-making theory of visual detection. Psychol Rev 61:401, 1954

9. Swets JA: ROC analysis applied to the evaluation of medical imaging Techniques. Invest Radiol 14:109-121, 1979

10. Kundel HL, Revesz G: The evaluation of radiographic techniques by observer tests: Pitfalls, problems and procedures. Invest Radiol 9:166-173, 1974

11. Swenson RG, Hessel SJ, Herman PG: Omissions in radiology: Faulty research or stringent reporting criteria? Radiology, 123:563-567, 1977

12. Truddenham WJ: Visual search, image organization and reader error in studies of the psychophysiology of roentgen image perception. Radiology 78:694-704, 1978

13. Metz CE: ROC methodology in radiologic imaging. Invest Radiol 26:120-123, 1986

14. Metz CE: Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 24:234-245, 1989

15. Metz C, Wang P, Kronman HB, et al: A new approach for testing the significance of differences between ROC curves measured from correlated data, in Deconink F (ed) Information Processing In Medical Imaging. Hague, The Netherlands, Nijhoff, 1984, 432-445

16. Kundel HL: Peripheral vision, structured noise and film reader error. Radiology, 114:269-273, 1975

17. Swets JA: Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. Psychol Bull 99:181-198, 1986

18. Swets JA: Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. Psychol Bull 99:100-117, 1988

19. Swets JA: Measuring the Accuracy of Diagnostic Systems. Science 1285-1293, 1988

20. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29-36, 1982

21. Hanley JA, McNeil BJ: Statistical approaches to analysis of receiver operating characteristic (ROC) curves. Med Decis Making 4:137-149, 1984

22. Revesz G, Kundel HL, Bonitatibus M, et al: The effect of verification on the assessment of imaging techniques. Invest Radiol 18:194-198, 1983

23. Roberts FS: Measurement Theory: With Applications to Decision-Making, Utility, and the Social Sciences, Reading, MA, Addison-Wesley, 1979 pp 4-34

24. Hanley JA, McNeil BJ: A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148:839-843, 1983

25. Hanley JA: Receiver operating characteristic (ROC) methodology: The state of the art. Crit Rev Diagn Imaging 29:307-335, 1989

26. Metz CE, Kronman HB: Statistical significance tests for binormal ROC curves. Math Psychol 22:218-243, 1980

27. Simel DL, Feussner JR, DeLong ER, et al: Intermediate, indeterminate, and uninterpretable diagnostic test results, Med Decis Making 7:107-114, 1987

28. Harrington MB, Miller KD: MTR-85W236, Results of the image quality evaluation conducted in conjunction with the 1984-1985 teleradiology field trial, McLean, VA: The MITRE Corporation, 1985

29. Siegel S: Non-parametric Statistics for the Behavioral Sciences. New York, NY, McGraw-Hill, 1956

30. Chakraborty DP: Free-response methodology: Alternative analysis and a new observer-performance experiment. Med Phys 174:873–881, 1990

31. Chakraborty DP: Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data. Med Phys 16:561-568, 1989