Observer studies involving detection and localization: Modeling, analysis, and validation[a)]

Dev P. Chakraborty[b)] and Kevin S. Berbaum
Department of Radiology, University of Pittsburgh, 3520 5th Avenue, Suite 300, Pittsburgh, Pennsylvania 15213 and Department of Radiology, University of Iowa, Iowa City, Iowa 52242

(Received 28 January 2004; revised 11 May 2004; accepted for publication 14 May 2004)

Although the receiver operating characteristic (ROC) paradigm is the accepted method for evaluation of diagnostic imaging systems, it has some serious shortcomings inasmuch as it is restricted to one observer report per image. By contrast the free-response ROC (FROC) paradigm and associated analysis method allows the observer to report multiple abnormalities within each imaging study, and uses the location of reported abnormalities to improve the measurement. Because the ROC method cannot accommodate multiple responses or use location information, its statistical power will suffer. The FROC paradigm/analysis has not enjoyed widespread acceptance because of concern about whether responses made to the same diagnostic study can be treated as independent. We propose a new jackknife FROC analysis method (JAFROC) that does not make the independence assumption. The new analysis method combines elements of FROC and the Dorfman–Berbaum–Metz (DBM) methods. To compare JAFROC to an earlier free-response analysis method (specifically the alternative free-response, or AFROC method), and to the DBM method, which uses conventional ROC scoring, we developed a model for generating simulated FROC data. The simulation model is based on an eye-movement model of how experts evaluate images. It allowed us to examine null hypothesis (NH) behavior and statistical power of the different methods. We found that AFROC analysis did not pass the NH test, being unduly conservative. Both the JAFROC method and the DBM method passed the NH test, but JAFROC had more statistical power than the DBM method. The results of this comparison suggest that future studies of diagnostic performance may enjoy improved statistical power or reduced sample size requirements through the use of the JAFROC method.
[DOI: 10.1118/1.1769352]

Key words: observer performance, ROC analysis, FROC analysis, localization, statistical power

I. INTRODUCTION

The receiver operating characteristic (ROC) method of assessing imaging system performance using experiments with observers has enjoyed widespread use. A recent review[1] describes the advantages of ROC analysis of ratings obtained when readers read the same cases in two or more modalities, as in the widely used Dorfman–Berbaum–Metz (DBM) method.[2] Such multi-reader multi-case (MRMC) methods are desirable for optimal statistical power and to generalize to both populations of cases and readers. In spite of its wide use it is not widely appreciated that the ROC method is not applicable when the decision task involves more than a simple determination of whether the patient is diseased or normal. In addition to *detecting* an abnormal condition, the radiologist often needs to *locate* specific image regions that are suspicious for disease. The additional location information cannot be used by ROC analysis and this neglect of location information may lead to a loss of statistical power, so that actual differences between modalities may be overlooked.[3]

Currently there are three ways to collect and analyze detection and localization data, each requiring a different task of the observer. In the localization ROC (LROC) paradigm[4] the observer provides one rating per case and indicates the location of the most suspicious region. Statistical analysis of data collected according to the LROC paradigm was made possible by Swensson.[5] In the region-of-interest (ROI) paradigm[6–8] the experimenter divides the image into regions and asks the observer to rate each for abnormality. The number of ratings per case equals the number of regions (in contrast to the ROC and LROC paradigms where the observer gives a single rating). Statistical analysis of data collected according to the ROI paradigm is described in Refs. 6–8. In the free-response ROC (FROC) paradigm[9] the observer searches each image for suspicious regions and assigns a rating to each identified (i.e., marked) region. Unlike the other methods, the number of ratings/marks is completely determined by the observer—on a particular case the observer may make zero, one, or more marks. The scoring of FROC data requires the specification of a criterion for correct localization. If the localization mark is within an investigator-specified distance of a lesion center—defined as the acceptance radius—it is classified as a true positive (TP) and all other events are classified as false positives (FP). Analysis of FROC data is possible by the classical FROC analysis[10] and the alternative free-response ROC (AFROC) method.[11] The latter reduces FROC data to pseudo-ROC data that can be analyzed by tools developed for ROC analysis [see the Medical Image Perception Society web-site,

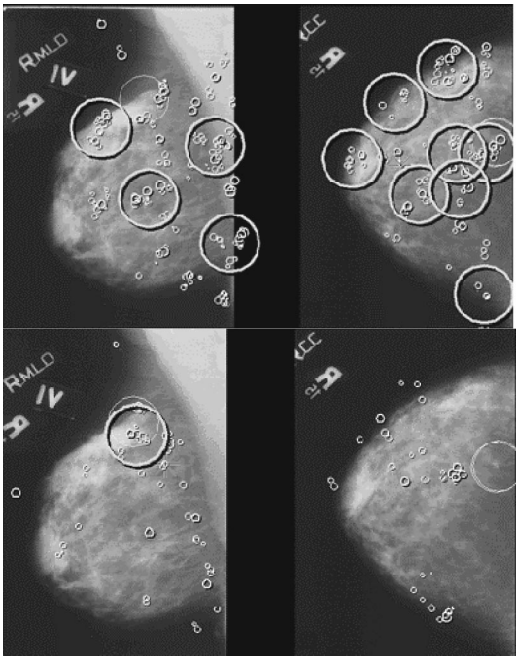

FIG. 1. Eye movement recordings for a two-view mammogram soft-copy display for two readers, an inexperienced reader (upper two panels) and an expert reader (lower two panels). Individual fixations are indicated by the small circles and clustered fixations are indicated by the larger high-contrast circles. A mass lesion visible on both views is indicated by the larger low-contrast circles. Note that the inexperienced reader makes decisions at many more locations than does the experienced reader. The large circles correspond to the fixation clusters in the model.

www.radiology.arizona.edu/krupinski/mips/rocprog.html for a compilation of available ROC-analysis software].

None of these approaches is entirely satisfactory. LROC analysis assumes independence between the signal and most suspicious noise location on the image, is limited to 0 or 1 lesion per image, and forces a location response even when the observer considers the image to be normal. AFROC analysis assumes independence between the multiple ratings obtained on the same case, which has drawn justifiable criticism.[12,5] Additionally, in AFROC scoring the noise property of each image is summarized by the highest rated false-positive event on that image and all other FP responses on that image are neglected. This neglect may lead to loss of statistical power.[6] A further vulnerability of AFROC analysis is that each lesion is assigned equal weight—so that a case with multiple lesions has a greater influence in determining the signal properties of the task than a case with only one lesion. To the extent that the interest is in making inferences about the population of cases, not the population of lesions, unequal weighting per case is undesirable.

While it does take into account intra-image correlations, the ROI approach has its own drawbacks.[13] Its most significant problem, in our opinion, is that it imposes on the reader a reading paradigm that is fundamentally different from that used in clinical practice. This can nullify the search strategy employed by radiologists, which can be directly measured by eye-movement studies.[14,15] Figure 1 shows eye movement recordings for a two-view mammogram soft-copy display for two observers, an inexperienced observer (upper two panels) and an expert observer (lower two panels). Individual fixations, defined as locations where the observer's dwell time exceeded 100 ms, are indicated by the small circles. Clustered fixations with a total dwell time exceeding 1 s are indicated by the larger high-contrast circles. It is believed[16] that the observer makes diagnostic decisions (to report or not to report) at these locations. In Fig. 1 the larger low-contrast circles indicate mass-lesions visible on both views. Notice that the inexperienced observer makes decisions at many more locations than does the experienced observer. The mechanism by which an observer generates the locations that subsequently receive prolonged dwell is not well understood, but clearly, the search strategy employed by the two observers is qualitatively different and experience-dependent. By imposing an investigator-specified definition of the regions where decisions are made, the ROI paradigm would tend to suppress the differences between the expert and the novice observer.

How does one decide which method (FROC, LROC, or ROI) to use? To compare evaluation tools against each other (in effect to *evaluate the evaluation tool*) one needs the language of statistical hypothesis testing. There are two states of truth regarding the modalities being investigated: they could be identical, termed the null hypothesis (NH), or they could be different, which is termed the alternative hypothesis (AH). The purpose of the evaluation tool is to correctly distinguish between these two possibilities. Incorrect rejection of the NH is termed a type I error (with probability α, if the statistical test is valid), and the incorrect rejection of the AH is termed a type II error (with probability β). The complement of the type II error probability ($1-\beta$) is termed the statistical power of the method. The method yielding greater power at the same α is defined to be superior, as it yields a greater probability of detecting a true difference between imaging systems without falsely declaring identical modalities to be different. For a specific set of data, say 100 normal and 100 abnormal cases, each read by 5 observers, the analysis method calculates a test statistic and compares it to a critical value. For a two-sided test if the magnitude of the test statistic exceeds the critical value, the method declares the modalities to be different, and otherwise it declares them not to differ. With respect to Fig. 2, the test-statistics under the two hypotheses are modeled by unit-normal distributions, one labeled NH and the other labeled AH, that are separated by d, the power parameter. The critical value for a 5% two-sided statistical hypothesis test ($\alpha=0.05$) is 1.96 and this value is represented by the arrow in Fig. 2. This means that the area under the NH distribution and lying above the critical value is 0.025 (i.e., half of 5%), and this region is shown as the cross-hatched region in Fig. 2. (The reason for the factor of half is that we are considering a two-sided test, so that values of the test statistic less than -1.96 will also lead to rejection of the NH.) The corresponding area under the AH distribution, lying above the critical value (and, for a two-sided test, an infinitesimal area lying below -1.96) is $1-\beta$, which is the statistical power, shown as the shaded area in Fig. 2, and is approximately 0.7 in this case. Note that as d increases the

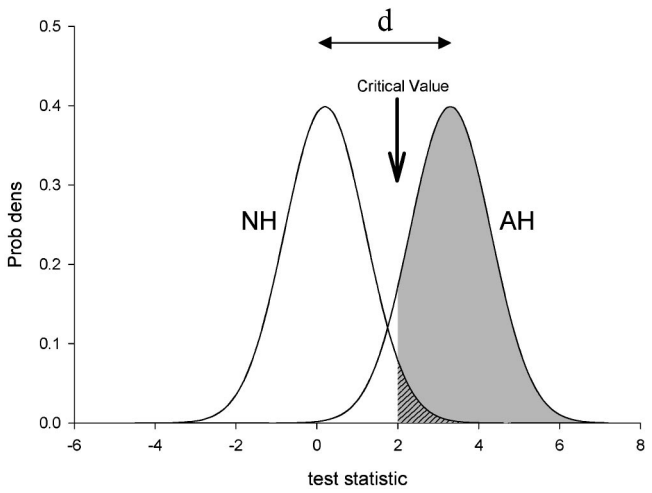

FIG. 2. The concept of statistical power is illustrated. The variations of the test-statistics are modeled by unit-normal distributions, one labeled NH and the other labeled AH, separated by d, the power parameter. The critical value for a 5% two-sided test ($\alpha=0.05$) is 1.96 and this value is represented by the arrow. For a two-sided test the acceptance region for the NH is from -1.96 to $+1.96$. The area under the NH distribution and lying above the critical value is 0.025 (cross-hatched area). Together with a similar area below -1.96 this yields the expected 5% value for the type I error. The area under the AH distribution lying outside the acceptance range (including an infinitesimal quantity below -1.96), is $1-\beta$, the statistical power, which in this case is approximately 0.7 (shaded area).

statistical power ($1-\beta$) increases rapidly and nonlinearly. Note also that the experimental determination of power requires many independent ROC experiments to be conducted, some with the modalities known to be identical and some with the modalities known to be different. Each of these experiments requires cases with known truth state, and willing and able observers. Each study requires a fresh sample of cases and new observers. Clearly it is futile to attempt to establish the power of a method with clinical experiments. The only viable approach is to use simulated data, and needed is a model for the data.

One can readily simulate ROC data by sampling from two normal distributions, and extensive simulation-based validations of ROC and DBM analysis have appeared.[17–21] However, due to the lack of a simulation model for generating data from the above-described experimental paradigms (FROC, LROC, or ROI), none of the statistical methods for analysis of data from these paradigms has been validated in this way. We describe a model for simulating data such as would be observed in FROC studies. We propose a jackknife method for the analysis of the ratings and location data that overcomes some of the problems with the AFROC method. Finally, we use the simulation model to perform a preliminary assessment of the new method.

II. METHODS

A. The simulation model

The purpose of the following is to describe a model for simulating ratings data such as are observed in FROC observer experiments. At the outset we note that in what follows the terms *signal*, *target*, and *lesion* are used synonymous.

1. Noise and signal sites

The biggest conceptual problem in scoring the marks generated by an observer in a free-response experiment is not how to score marks in signal-containing regions (*signal sites*), but how to score the marks made in the normal regions. Obviously the observer thought those sites might contain lesions. We use the term *noise sites* to designate those sites considered by the observer as potential lesion sites. If the observer finds one of two abnormalities, we know that sensitivity is 50%. However, if two abnormalities are reported at locations known to be normal, what is the specificity? How many sites are there in an image in which abnormalities might occur (but do not)? This is the same problem one faces in developing a valid simulation model for the FROC data. Our simulation model relies on perception theory to address this issue.

"Sometimes the eye pauses and returns one or more times or even stops for a few seconds. This looks like a decision-making activity"—Hu, Kundel, Nodine, Krupinski, Toto[22]

In Fig. 1 we show some eye position recordings and the locations of fixations (indicated by the small circles) and fixation clusters (indicated by the larger circles). Analysis of eye-position data usually involves grouping of fixations into clusters because the human visual system often prefers to use a series of closely spaced, shorter fixations rather than a single long fixation to sample visual information from a region within the visual field.[23,24] Nodine and Kundel differentiated scanning or discovery fixation clusters from reflective clusters by the longer duration of the latter. Greater gaze time dwell on a small region may indicate that the observer has recognized an image element that is sufficiently suspicious so as to require further perceptual analysis and a decision about whether the region contains an abnormality. This analysis of gaze dwell time tells us how many attention-units of analysis the radiologist uses in inspecting an image.

According to Nodine and Kundel's (1987) model of visual search and detection the observer does not examine all regions of the image in the same manner with the same information processing. Rather, when the image is first displayed, a global processing (lasting about 100 ms) identifies some locations as requiring closer inspection, and these locations receive clustered fixations. In this way, the observer rapidly eliminates much of the image as unlikely to contain lesions. The efficiency with which the observer can do this preliminary filtering depends on their expertise, the specific image, and the task. [Although initially non-fixated locations can be processed by peripheral vision, so long as detection requires that the signal be examined using the higher resolution fovea, the sites identified by peripheral vision must still receive a clustered fixation for detection. Of course, some visual tasks do not require foreal vision; we restrict the FROC simulation model to tasks that do.]

A simple example may help to explain why all regions on an image do not qualify as potential target sites, specifically, why the number of noise sites cannot be calculated using the area of the image divided by the average lesion area. Such a

calculation would make sense if the task involves detection of lesions in a uniform background, such as common quality control phantoms used by medical physicists, if by ''area'' one means the projected area of the phantom, not the total film area. The American College of Radiology phantom used in mammography involves low contrast objects on a uniform acrylic base.[25] Trained readers of this phantom know that no objects can occur outside the image of the acrylic base, the so-called ''black-film'' region. In other words, by using *a priori* knowledge and training, the expert phantom readers can reduce the number of sites that are possible noise sites, compared to an observer that has not received such *a priori* knowledge and training. This example shows how the number of noise sites can depend on the expertise of the reader, and is not simply equal to the image area divided by the lesion area. With clinical images the situation is similar, although the nature of the expertise is perhaps more mysterious to the nonexpert. We believe that just as the expert phantom reader eliminates the black-film regions of the quality control image from perceptual analysis, the expert mammographer eliminates certain areas inside the breast. In Fig. 1 these are the locations that received fixations but not clustered fixations.

One may wonder if the novice who has fixation clusters at several locations, only one of which would qualify as a noise site for an expert, and who, like the expert, decides to not mark any of them (true-negative decisions) is getting credit for more TN events, but the expert is not—so the analysis would appear to be biased against the expert. That is definitely not the case. As we shall see in the following, all other conditions being equal, the novice who looks at more locations than the expert would generate more FPs, and this will be reflected in the FROC analysis.

The total number of signal and noise sites, denoted by T, is a parameter of the FROC simulation model. They represent the sites that resulted from the global processing which identified the locations that received subsequent clustered fixations. Note that the number of noise sites ($T-$number of lesions) is always expected to be greater than or equal to the actual number of false positives on the image, since some of the noise sites may not be marked by the observer. The T-parameter is expected to depend on the case (difficult cases require more careful examination and T will be larger), the reader's expertise level (experts will tend to have smaller values of T) and the task (one needs to examine the image more closely if one is seeking to detect small lesions, and T will be larger). *None of the methods for analyzing FROC data requires knowledge of T*. However, the value of T does affect the simulation model for generating FROC data, which is needed to evaluate the methods.

On the assumption that each sufficiently long cluster of fixations indicates that a decision (to report or not to report, i.e., overt of covert) was made, eye-position recording technology can provide an approximate physical meaning of the T parameter of the FROC model. Other methods exist for measuring where and how many decisions were made on an image and the correspondence of fixation cluster times to decisions may be less certain for some tasks and types of abnormality. Indeed, the notion that a certain length of gaze dwell time indicates that a decision was made is itself largely a matter of definition. An operational meaning of T is useful because it allows us to simulate a process that we know must occur within the observer—the observer must be able to reduce the infinite number of possible sites to a finite number. With advances in our ability to measure perceptual search, a better understanding of the number of decisions made by an observer may become available.

Based on the assumption that the T parameter can be measured in eye-movement studies, such as Fig. 1, we could base the entire simulation on measured eye-movement data. While such a procedure would be optimal, it is beyond the scope of the present work. Therefore, we made some simplifying assumptions.

a. Value of T. We neglected the possible case dependence of T, so that for a given reader we could regard T as a constant. To model the reader dependence, and to determine if the results were sensitive to the T parameter, we used a range of values of T. Specifically, in the simulations T was set at either 3 (expert readers), 5 (average readers), or 20 (lay readers).

b. Precision of the observer marks. We assumed an infinitely precise observer and infinite precision in the scoring. In other words if a signal was detected the observer marked the corresponding signal site with infinite precision, so that there is no ambiguity in the scoring, and a mark intended for a signal site was always be recorded as a TP event. In practice, if the mark is not close to a lesion site and the acceptance radius is small, it might be classified as a FP. Likewise, we assumed that the probability of random true positives (by chance the observer's mark happens to be close to an actual lesion, and is scored as a TP, even though he did not actually detect the lesion) was zero.

c. Measurability of responses. All signal and noise site responses are measurable: sites that are marked receive an explicit (observer-specified) rating, and unmarked sites receive the default (investigator-specified) rating. In a 4-rating FROC study, where the explicit responses are 1, 2, 3, and 4, and assuming that the higher numerical ratings correspond to greater suspicion for the presence of lesion, the default rating could be 0.

2. Signal and noise detection events

From the case selection and verification process, the experimenter knows the signal sites, but the observer is blinded to this knowledge. The noise sites are determined by the observer, the image, and the task, and are not under the experimenter's control. Associated with each of the noise and signal sites there is a scalar *decision variable* (DV), with the property that higher values correspond to increased suspicion for lesion presence at that site. The observer marks all locations with DVs exceeding the cut-off value, and each such occurrence is termed a *detection event*. If the DV of a signal exceeded the cutoff one has a *signal detection event* and otherwise one has a *noise detection event*. This describes the situation for a FROC observer who marks and rates as ''1'' all locations where the DV exceeds the cut-off. By employing multiple cut-offs, the model can be readily extended to

the FROC observer using an N-point scale (typically N is 3 or 4 in FROC experiments). While it is not implemented in our current software, algorithms designed for the analysis of ROC data acquired on a continuous scale can be adapted to FROC studies. Hybrid data, which are discrete below some cut-point and continuous above it, can also be analyzed readily by current MRMC software and such extensions would seem to pose no fundamental problems for the methods proposed in this manuscript.

3. The ratings of the observer-generated marks

The ratings simulation model has been described previously.[3,26] To accommodate multiple responses per image in a FROC experiment one must use a multi-valued decision variable (DV). The DV is expected to have both *case* dependence (patient-to-patient variations) and *location* dependence (for a given patient it varies from location-to-location). The concepts of case and location sampling need some clarification. Consider a step-wedge phantom with holes of varying diameters. This is often used to measure a contrast-detail curve and it is well known that due to increased photon flux, smaller holes can be better visualized in the thinner parts of the phantom. This is an illustration of location sampling, whereby visibility of an object depends upon its location. Now imagine that images are acquired of the step-wedge phantom with superposed sheets of Lucite™, of varying thickness, that extend over all the steps of the step-wedge phantom. It is clear that the holes will be relatively easier to visualize if the Lucite thickness is small, and conversely, when the Lucite thickness is large, the holes will be harder to visualize. This is an illustration of case sampling, with the thin Lucite sheet images representing ''easy'' cases and conversely, the thick Lucite sheet images represent ''hard'' cases.

The FROC decision variable model has case-sampling and location-sampling components, and for each component one has to distinguish between noise and signal locations, leading to the four distributions shown in Fig. 3. For an abnormal case one samples all four distributions, and for a normal case one only samples the two noise distributions.

a. Case sampling. Shown at the top and labeled ''CASE'' is a bivariate Gaussian distribution describing the case sampling. Samples from this distribution correspond to the different Lucite sheets in the above-given example. There are two distributions, corresponding to noise and signal sites, as normal and abnormal regions can be present simultaneously in the same image (as in Fig. 1). A particular abnormal case realizes two decision variable outcomes (labeled ξ, ψ) from this distribution, representing the noise and signal DV, respectively, for that case. The arrow labeled ''ρ_{SN}'' at the top of Fig. 2 indicates the noise-signal correlation, which is a parameter of the bivariate Gaussian distribution. The correlation parameter is necessary as the noise and signal regions belong to the same case. A large positive value of ρ_{SN} (subject to $|\rho_{SN}|<1$) causes the two case samples to be positively correlated and, similarly, a large negative value causes them to be anticorrelated. The noise distribution is centered at 0 (by convention) and the signal distribution is centered at μ and the corresponding standard deviations are σ_{CN} and σ_{CS}, respectively (CN=case noise, CS=case signal).

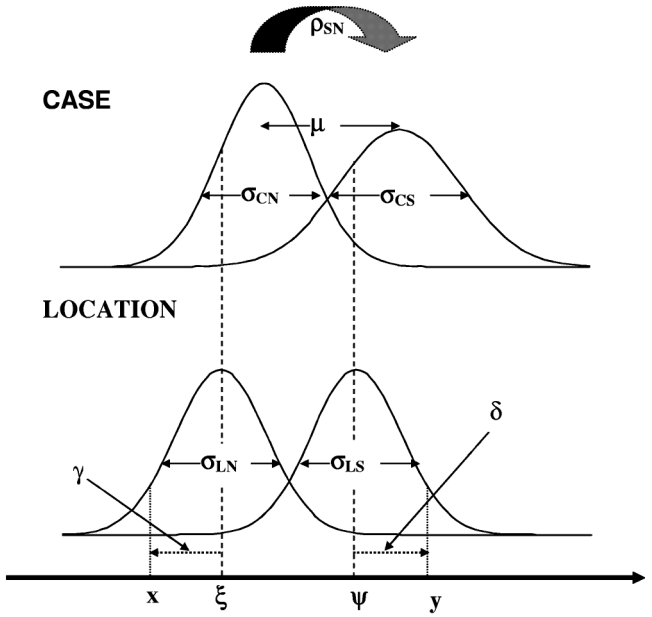

FIG. 3. The FROC decision variable sampling model in which the decision variable is expressed as the sum of case and location dependent terms. The upper distributions represent the case-sampling, which yields two samples per case (ξ,ψ) representing the normal and abnormal region contributions, respectively, from each case. These are drawn from a bivariate-normal distribution with correlation ρ_{SN} (as depicted by the arrow at the top). The lower two independent, univariate-normal distributions represent the location samples, which yield the samples $(\gamma_1,\gamma_2,\ldots,\gamma_N;\delta_1,\delta_2,\ldots,\delta_S)$ representing the N noise samples from the normal region and S signal samples from the abnormal region. In our notation $T=N+S$.

b. Location sampling. Centered on the ξ and ψ samples are two independent Gaussian distributions labeled ''LOCATION'' from which the decision variables γ, δ for the different locations are sampled T times for each case. Sampling from these distributions correspond to the varying step-wedge thickness in the example. Noise locations are sampled from the LN (location noise) distribution, with standard deviation σ_{LN}, and signal locations are sampled from the LS (location signal) distribution, with standard deviation σ_{LS}. In our example, samples from the LN distribution would correspond to locations of the step-wedge phantom that do not contain holes. These regions present opportunities for false positives. Likewise signal sites are sampled from the LS distribution, corresponding to the holes. Notice that centering the LN and LS distributions on the ξ, ψ samples causes the net DV (x,y) for each signal or noise site to be the sum of the corersponding samples from the case and location distributions. The fact that the multiple location samples from the same case are made to share a common case-component means that they will be correlated. If the case component is large relative to the location component, the correlation will be large. A detailed mathematical description of the detection model is given in Appendix A.

c. Relation of the model to FROC and AFROC analyses.

TABLE I. This is a summary of symbols, acronyms, and simulation parameters used in the manuscript.

	Symbol	Description	Simulation values
General	TP, FP, TN, FN	True positive, false positive, true negative, false negative	
	HN	Highest noise	
	DV	Decision variable	
	ML	Maximum likelihood	
	ILF	Incorrect localization fraction, see Sec. III F	
Detection	μ	Mean of case-signal distribution, see Fig. 3	0.75, 1.5, 2.5
	σ_{CS}, σ_{CN}	Standard deviation of case noise or signal samples	$\sigma_{CS}=\sigma_{CN}=0, \pm5, \pm10$
	ρ_{SN}	Correlation between case noise and case signal samples	$0, \pm0.5, \pm0.9$
	σ_{LN}, σ_{LS}	Standard deviation of location samples	$\sigma_{LN}=\sigma_{LS}=1$
	IICS	Intra-image correlation structure, see Table II	NN, N, Z, P, PP
	T	Total number of sites per image	$T=3, 5, 20$
	S	Number of signal sites per abnormal image	1
	NN, N, Z, P, PP	Names of intra-image correlation structures defined in Table II	
Simulations	BINS	Number of discrete ratings bins	4, 20
	N_{TRIALS}	Number of NH trials=Number of AH trials	2000
	N_T	Total number of normal images	200
	N_N	Number of normal images	100
	N_A	Number of abnormal images	100
	CRCS	Case-reader correlation structure[a]	HL
	$\Delta\mu$	Separation between modalities	0.25
	N_R	Number of readers	5
Statistical	NH, AH	Null hypothesis, alternate hypothesis	
	PV	Pseudovalues	
	α	Critical p value that is used for testing a NH, typically fixed at 5%	0.05
	P_{NH}	Observed (empirical) p value	
	β	Statistical power	
	d	Normalized separation of NH and AH distributions	

[a]Reference 19.

The above-described simulation model is basically a FROC model. That is, it predicts the DV corresponding to all T sites of a case. Some of these sites correspond to signals and the rest are noise sites. Application of the observer's lowest cutoff filters out the sites that fall below this value. The remaining sites and their ratings are the simulated FROC data for this case. The distinction between AFROC and classical FROC is in how one chooses to analyze this data. In classical FROC analysis one uses all the noise responses. In AFROC analysis one uses only the highest noise (HN) responses.

d. Details. Table I has a summary of the model and other simulation parameters. The measurement scale for the DV axis is fixed by choosing $\sigma_{LN}=1$. Therefore, the simulation model is described by the six parameters: μ, σ_{CS}, σ_{CN}, ρ_{SN}, σ_{LS}, and T. For $T=1$ the model reverts to the binormal ROC model.[27] The different intra-image correlation structures (IICS) investigated and the descriptive terms for them (NN: high-negative, N: medium-negative, Z: zero, P: medium-positive and PP: high-positive) are shown in Table II.

B. Jackknife analysis of FROC data

There are two basic steps to the analysis of FROC observer data—a scoring step and a statistical analysis step. The intent of the scoring step is to reduce the FROC data for a given observer and set of cases to a single number, a figure-of-merit θ (which is the analog of the area under the ROC curve) that rewards the observer for good decisions (true positives and true negatives) and penalizes the observer for bad decisions (false negatives and false positives). The intent of the statistical analysis step is to estimate a confidence interval for the θ—this is where one needs to be concerned with possibly underestimating the confidence interval by assuming that the same-case responses are uncorrelated.

TABLE II. The intra-image correlation structures (IICS) that were used in the simulations and the associated notation are summarized. One can think of NN as "highly negative," N as "Negative," Z as "Zero," P as "Positive," and PP as "highly positive." Note that we set $\sigma_{LS}=\sigma_{LN}=1$ in all cases. All symbols are defined in Table I.

Notation	σ_{CS}, σ_{CN}	ρ_{SN}
NN	10	−0.9
N	5	−0.5
Z	0	0
P	5	+0.5
PP	10	+0.9

1. Scoring step

The first step to scoring the marks is to classify them as true positives (TPs) or false positives (FPs). This is done by selecting an acceptance radius i.e., a clinically relevant distance with the understanding that a mark that is closer than this distance from an actual signal center will be scored as a TP. Marks that are further than this characteristic distance will be scored as FPs. The selection of the acceptance radius has an obvious impact on the analysis (a small distance will lead to small apparent FROC performance) and it is very important to state this parameter and to keep it consistant for the modalities being compared.

In traditional AFROC scoring each signal contributes a rating to the analysis. As noted earlier, when variable numbers of signals are present, this scoring gives unequal weighting to the abnormal cases. This concern can be met by a modified lesion scoring scheme in which one assigns *weights* to each lesion, where the weights add up to unity, and defines a figure-of-merit θ (which is the analog of the area under the ROC curve) that involves a weighted combination of the ratings. Defining N_T the total number of cases, indexed by i, N_A the total number of abnormal cases, indexed by j, n_j is the total number of lesions in the jth abnormal case, one calculates the figure-of-merit θ as follows:

$$\theta = \frac{1}{N_T N_A} \sum_{i=1}^{N_T} \sum_{j=1}^{N_A} \sum_{k=1}^{n_j} W_{jk} \psi(X_i, Y_{jk}),$$

$$\psi(X,Y) = \begin{cases} 1.0 & \text{if } Y > X \\ 0.5 & \text{if } Y = X, \\ 0.0 & \text{if } Y < X \end{cases} \tag{1}$$

$$\sum_{k=1}^{n_j} W_{jk} = 1.$$

Here X_i is the highest noise (HN) rating for normal case i, Y_{jk} is the signal rating for the kth target on case j, and W_{jk} is the relative importance of detecting the kth signal on abnormal case j. The statistic θ is the weighted-probability that a signal rating exceeds a noise rating. It is bounded between 0 (worst possible performance) and 1 (perfect performance). The lower limit of zero applies when the number of sites T is very large. This can be seen by considering the case of essentially undetectable lesions, where no lesion sites are marked, even by chance and therefore each lesion receives the default ''0'' rating. Only FP sites are marked and explicitly rated (i.e., receive ratings >0), so the HN-rating always exceeds the lesion rating.

The lesion weights W_{jk} can be determined by asking the clinician using the lesion information (e.g., the surgeon) the clinical significance of each lesion. For example, the surgeon could be asked to quantify the significance of missing each lesion from the point of view of its ultimate impact on the patient's outcome. The weights are normalized, so that they add up to unity for each case, and should be averaged over several clinicians. The weights become part of the truth information for each case. Since the weights add up to unity, the statistic θ is unaffected by a case that has a large number of lesions. If each case has a single lesion, then the weights are all unity and θ reduces to the Wilcoxon statistic.[28]

2. Analysis step

Jackknife analysis of FROC data follows the same basic rationale as jackknife analysis of ROC data. In the DBM method one deletes each case, one at a time, with replacement, and recalculates the area under the ROC curve. A transformation is used to convert the area to a pseudovalue (PV). The underlying idea is that jackknifing preserves the effect on the ROC area of the individual cases (prior to jackknifing the traditional methods collapsed the information of all cases into a single ROC area). Moreover, the pseudovalues can be regarded as observed data and can be modeled in a physically meaningful manner. The pseudovalues are analyzed by an analysis of variance (ANOVA) method[2] that accounts for the underlying physical model.

Note that in ROC analysis each case provides one rating so when this case is removed from the analysis a *single* rating is dropped. The extension of this concept to FROC analysis is straightforward. Now each case can provide multiple ratings, so that when a case is removed from the analysis *all* the ratings for that case are dropped. The pseudovalues are calculated in the normal manner [see Eq. (3)]. In the subsequent ANOVA each case is represented by a *single* pseudovalue. Note that no assumptions regarding the correlation structure of the multiple responses on a single case are needed to calculate the pseudovalue. In this manner the jackknife method circumvents the issue of how to account for the intra-image correlations. A similar idea is behind a recent bootstrapping extension of the original ROI method.[8]

For reasons that will become apparent, two jackknife schemes were investigated, that differed only in how the figure-of-merit quantity θ was calculated.

a. JAFROC method 1. In this analysis (henceforth termed JAFROC-1) one keeps in the analysis the highest noise (HN) rating on *all* images (normal and abnormal), and in particular the first summation in Eq. (1) runs, as shown, from 0 to N_T, the total number if images. In other words, and as in traditional AFROC scoring, each image contributes a HN event. The θ in Eq. (1) can be defined as the probability of the signal rating exceeding the HN rating, where the HN rating is measured on *any* image (normal or abnormal).

b. JAFROC method 2. A second method (termed JAFROC-2) was also investigated. This method consisted of ignoring the HN responses on abnormal images. Note that with this method the definition of θ changes to

$$\theta = \frac{1}{N_N N_A} \sum_{i=1}^{N_N} \sum_{j=1}^{N_A} \sum_{k=1}^{n_j} W_{jk} \psi(X_i, Y_{jk}). \tag{2}$$

Here the first summation runs from 1 to N_N, the total number of normal cases. All other definitions are unchanged. The revised θ in Eq. (2) can be defined as the probability of the signal rating exceeding the HN rating, where the HN rating is measured on *normal* image only.

C. Simulation testing

Simulations conducted to test the methods were resticted to two modalities and to model parameters $\sigma_{CS}=\sigma_{CN}$ and $\sigma_{LS}=1$, see Table II.

1. Generation of ROC ratings from the detection-localization model

The equivalent ROC rating for an image was defined as the highest of all the ratings for that image. In other words, it was assumed that if the observer was constrained to give a single rating for the image, he or she would give the most suspicious rating. This follows the same scheme described in our prior publication[3] and that used by others.[5]

2. MRMC algorithms

The MRMC algorithm based software used in this work, whose source code and executable were downloaded around 2001 from the University of Chicago web-site, produced output containing the identifiers ''LABMRMC (Macintosh PPC version 1.0 b3)'' and ''MRMC16 1.55''. This software was recompiled (with the assistance of Ben Herman) under the Microsoft Visual C++ compiler (version 6.0), tested and modified to allow us to examine its working and to monitor internal variables and to allow us to call it from the main IDL control code. The LABMRMC software generally uses the maximum likelihood (ML) estimate of the ROC area, and sometimes the ML algorithm can fail to converge. We monitored the nonconvergence condition and whenever LABMRMC failed to converge we incremented a counter and restarted the simulations with a different seed, i.e., we drew fresh cases and fresh readers and repeated the two-modality MRMC study. This was repeated until the method finally converged. Of course, in practice one does not have the luxury of repeating the entire ROC experiment with new readers and cases until the analysis program converges. In reporting simulation studies using the original LABMRMC code it is necessary to also report the fraction of times when the method fails altogether—if this is unacceptably high, the experimenter might, at the design stage, reconsider use of the method. We also evaluated a modified LABMRMC algorithm, which used the trapezoidal rule to calculate the ROC area, instead of the ML method, and consequently it did not have any convergence problems.

3. Variance structures

In a multi-reader multi-case experiment the same readers interpret the same set of cases in the different modalities. Therefore, in addition to including intra-image correlations as described earlier, one has to include inter-image correlations, describing the correlations between the ratings when different observers read the same sets of cases, and the correlations between the ratings when the observers read the same images in the two modalities. This is done by specifying additional Gaussian distributions describing reader and case variability, and various interaction terms. We specified these correlations by selecting from one of the eight variance structures originally proposed by Roe and Metz.[29,19] One of these, termed HL (for high data correlation, low reader variance), was extensively used in our work. In this paper the Roe and Metz correlation structures are referred to as case reader correlation structures (CRCS), to distinguish them from the intra-image correlation structures (IICS). The manner in which both inter- and intra-image correlations were included in the sampling model is described in Appendix B.

4. Binning

The above-described simulation model generated continuous DV values. To accommodate observer experiments in which a finite number of discrete ratings are employed, we binned this variable into BINS values. This was done by assuming that all the bin-widths were equal. Specifically, for each modality and each reader we determined the maxima and minima of the ($T\times$total number of cases) ratings, and divided this interval into BINS+1 equal segments.

5. Calculation of the pseudovalues

Each time a case is deleted one needs to calculate a figure of merit for the remaining cases. For the figure of merit we used the modified Wilcoxon statistic θ defined earlier. The calculation of the pseudo values PV_{ijk} followed the procedure in the original DBM paper, namely,

$$\mathrm{PV}_{ijk}=N_T\theta_{ij}-(N_T-1)\,\theta_{ij(k)}, \tag{3}$$

where θ_{ij} is the figure of merit for the ith modality and the jth reader, when all cases are used in the calculation, and $\theta_{ij(k)}$ is the figure of merit for the ith modality and the jth reader when case k is deleted, and N_T is the total number of cases.

6. Measure of statistical power

To measure power we have previously introduced a detectability parameter-like quantity,[3] which was defined as the separation between the NH and AH distributions, modeled as unit-normal distributions. We denote this parameter henceforth by the symbol d. A high value for d means that it is relatively easy to distinguish between the NH and AH, in other words, to detect the difference between the two modalities. The quantity d was estimated from the probability of the F-statistic for the AH exceeding that for the NH [the calculation of this probability is identical to the formula for ψ appearing in Eq. (1)]. This is the same as the trapezoidal area, A, under the power curve, defined as a plot of 1-β vs α (see Sec. I for the definitions of α and β). The advantage of using the trapezoidal area as a measure of power (rather than simply counting the number of times the AH is rejected) is that all the simulations, including those that did not reject the AH, but where the AH statistic exceeded the NH statistic, contribute to the trapezoidal area, thus resulting in a more stable estimate. By analogy to ROC analysis where the area under the ROC curve can be formally transformed to an equivalent detection index, the trapezoidal area (which ranges between 0.5 and 1) was converted to an equivalent separation of two unit normal distributions. The transformation is given by[30]

$$d = 2\ \text{erfinv}(2A-1). \tag{4}$$

Here erfinv is the inverse of the error function erf.[31] The transformation has the advantage that the resulting d behaves more linearly than A and, moreover, d is monotonically related to A. For example, with increasing power the area values cannot be readily differentiated (e.g., area values of 0.99 and 0.999 versus d values of 3.29 and 4.37). Note that this is a matter of convenience and that we are not assuming that the F-distribution is unit-normal (it is not). [A similar technique is used in the ROC context to calculate d_a—Refs. 30 and 32—the detection index, from the ROC area by using Eq. (4). The ROC area is calculated using a binormal model with generally unequal variances.[27] While it is true that d_a can be defined without making any assumption about equal variances, the physical meaning of this parameter is that of the separation of two unit normal distributions.] The relation between d, α and power ($P=1-B$) for a two-sided z-score test is as follows:

$$\alpha = \frac{2}{\sqrt{2\pi}} \int_z^\infty \exp\left(\frac{-t^2}{2}\right) dt, \tag{5}$$

$$P(d,\alpha) = 1 - \Phi(z-d) + \Phi(-z-d), \tag{6}$$

where $\Phi(x)$ is the Gaussian distribution function, i.e., the probability that a random variable sampled from a unit-normal distribution is less than x, and z is defined so that the probability that a random variable sampled from a unit-normal distribution exceeds z is the desired type-I error rate α. In the IDL programming language the z-function is implemented as GAUSS_CVF and Φ is implemented as GAUSS_PDF. It can be confirmed that $P(0,0.05)=0.05$ and $P(1.96,0.05)=0.5$. As noted above, Eq. (6) is approximate since the NH and AH distributions are not strictly normal.

7. Null hypothesis test

While d is important in judging the power of a method of analysis it is also important to have a test with the correct NH behavior. The observed probability of incorrectly rejecting the NH is denoted by P_{NH}. If the targeted level of the test is $\alpha=5\%$ and if $P_{\text{NH}}<0.05$ then one has a ''conservative test,'' meaning that the cutpoint internal to the ANOVA analysis, that was used to judge if the F-statistic was in the rejection region, is too high. This results in fewer than expected rejections of the NH. This is not desirable since the method will then also reject the AH more often than it should, and the expected power, based upon Eq. (6), will not be realized.

Assuming a binomial distribution the standard deviation of P_{NH}, denoted by $\sigma(P_{\text{NH}})$, is given by

$$\sigma(P_{\text{NH}}) = \sqrt{\frac{P_{\text{NH}}(1-P_{\text{NH}})}{N_{\text{TRIALS}}}}, \tag{7}$$

where N_{TRIALS} is the number of NH trials. The 95% confidence limits are given by ± 1.96 times $\sigma(P_{\text{NH}})$. If this included the nominal 5% level, then the method is regarded as having passed the NH test. Note that for 2000 NH simulations the confidence limits are approximately ± 0.01 (i.e., $1.96\sqrt{[0.05(0.95)/2000]}$).

8. Software implementation

The software was implemented in IDL (Research Systems Inc., Boulder, CO) with computationally intensive parts of the code implemented in C and FORTRAN. For example, the MRMC software source code is available in FORTRAN on the University of Chicago website, and we adopted that code with minimal changes and compiled it as a Dynamic Link Library (DLL) that was callable from IDL. The ANOVA implementation of the DBM method was also adapted from the MRMC code. The random number generator used is the one implemented in IDL, which is based on the ''ran1'' routine whose source code is available in Numerical Recipes. An important aspect of the simulation was the control over the random number sequences. This was accomplished by initializing the random number generator (this is known as ''seeding'') with known integer values (''seeds''). Independent random number sequences are generated using different seeds. Conversely, identical random number sequences can be generated using the same seeds. This control was necessary as we wished to test the methods on the same reader and same case samples, to satisfy the matched conditions under which the clinical experiments are commonly conducted.

III. RESULTS

A. Consistency check of simulation method

Note that if we set the parameter $T=1$ the simulation model becomes identical to a standard binormal ROC simulation model. To check our simulation methods against independent work, we applied it to two-modality NH simulations, which have been extensively studied previously by Roe and Metz.[19]

For the HL variance structure, with 25 normal and 25 abnormal cases, $\mu=2.50$, 5 readers, and 2000 trials of the NH condition, Roe and Metz observed $p\sim 0.018$ for the NH failure rate [our estimate from their Fig. 1A], which was significantly below the nominal 5% level. Under these same conditions and with the parameter BINS=20 (simulating quasicontinuous ratings), we observed $p=0.019$, which agrees very well with the Roe and Metz determination. A $\mu=1.50$ simulation with 100+100 cases, and 3 readers, confirmed the Roe and Metz result ($p\sim 0.07$), since we observed 0.059, both of which are significantly above the 5% level. In general Roe and Metz noticed that departures from the NH generally occurred when μ was large, the number of positive and negative cases was small, and the number of readers was small. In all the work detailed in the following we focused on the HL variance structure, 100 normal and 100 abnormal cases, 1 lesion per abnormal case, and 5 readers, both to keep the analysis manageable and since these conditions presented the least NH problems to ROC analysis. For NH simulations we set $\Delta\mu=0$ and for AH conditions (power estimates) we set $\Delta\mu=0.25$. NH behavior was as-

TABLE III. The results of simulation testing of the ROC and JAFROC-2 methods are presented. In all cases we used 2000 NH simulations, 2000 AH simulations, 100 normal and 100 abnormal cases, 5 readers, CRCS structure=HL and BINS=4. CRCS describes the case reader correlation structure, see Table I, and HL is defined in Ref. 19. This is for IICS=NN, i.e., for strong negative intra-image correlations. Cells with asterisks represent failed NH-tests.

		ROC			ROC	JAFROC-2	ROC	JAFROC-2
T	μ	ILF-NH	θ (ROC)	θ (JAFROC-2)	P_{NH}	P_{NH}	Power-parameter (d)	
3	0.75	0.142	0.661	0.599	0.049	0.043	0.575	1.187
	1.5	0.071	0.768	0.748	0.047	*0.042**	0.763	1.058
	2.5	0.012	0.882	0.880	0.059	0.057	0.656	0.716
5	0.75	0.143	0.645	0.567	0.050	0.051	0.497	1.208
	1.5	0.075	0.748	0.722	*0.041**	*0.042**	0.727	1.084
	2.5	0.014	0.868	0.864	0.048	0.054	0.691	0.788
20	0.75	0.274	0.608	0.494	0.048	0.053	0.324	1.189
	1.5	0.102	0.704	0.658	0.050	0.048	0.671	1.183
	2.5	0.020	0.833	0.825	0.048	0.047	0.611	0.773

sessed with 2000 simulations and 2000 AH simulations were conducted when power estimates were desired. The latter were only conducted when a method passed the NH testing. Unless otherwise noted we set the binning parameter BINS=4, corresponding to a 5-rating ROC study or a 4-rating FROC study. In the ROC case all ratings are explicit (i.e., observer assigned). In the FROC case in addition to the explicit ratings one has the ''0'' rating, which is an implicit rating (i.e., investigator assigned), corresponding to regions with DV below the lowest observer-cutoff. This is the reason why a 4-rating FROC study corresponds to a 5-rating ROC study.

B. Check of modified MRMC method

As noted previously, we used the trapezoidal area instead of the ML estimate for the area under the ROC curve. Because of the difference of our ROC analysis from the accepted MRMC method, we performed a comparison of the original and the modified MRMC algorithms. These simulations were conducted for IICS=NN, N, Z, P, and PP; T=3, 5, 20; μ=0.75, 1.5, and 2.5; and 1 lesion per abnormal case. A paired t-test showed significant differences (p =0.0433) between the two sets of d values, with the modified method yielding a higher d value on the average than the original method (0.6923 vs 0.6685). On an average of 3.2% of the trials (maximum 9.4%) the MRMC algorithm failed to converge, with convergence problems occurring predominantly for μ=2.5. We observed that in 19 out of 45 runs the original MRMC method failed the NH test, vs 4 out of 45 failures with the modified method. Based on these results we can conclude that for the CRCS=HL the modified MRMC method, which uses the trapezoidal area rather than the maximum likelihood area, outperformed the traditional MRMC method in terms of NH behavior, power and convergence properties. Note that these results, conducted with a simulation model which allowed for multiple sites per image (T =3, 5, or 20) do not contradict the earlier study,[19] which was conducted (effectively) with T=1.

C. NH behavior of AFROC

In this analysis the AFROC scored pseudo-ROC data are submitted to the MRMC software, and jackknifing occurs internal to the MRMC software. Since each abnormal image contributed two responses (a HN and a signal response) the

TABLE IV. The results of simulation testing of the ROC and JAFROC-2 methods are presented. In all cases we used 2000 NH simulations, 2000 AH simulations, 100 normal and 100 abnormal cases, 5 readers, CRCS structure=HL and BINS=4. CRCS describes the case reader correlation structure, see Table I, and HL is defined in Ref. 19. This is for IICS=N, i.e., for negative intra-image correlations. Cells with asterisks represent failed NH-tests.

		ROC			ROC	JAFROC-2	ROC	JAFROC-2
T	μ	ILF-NH	θ (ROC)	θ (JAFROC-2)	P_{NH}	P_{NH}	Power-parameter (d)	
3	0.75	0.129	0.651	0.597	0.042	0.049	0.612	1.182
	1.5	0.061	0.762	0.747	0.047	0.043	0.820	1.060
	2.5	0.011	0.881	0.879	*0.060**	0.057	0.660	0.698
5	0.75	0.185	0.634	0.565	*0.040**	*0.042**	0.501	1.168
	1.5	0.060	0.740	0.720	0.051	0.049	0.717	1.031
	2.5	0.005	0.865	0.863	0.049	0.051	0.709	0.768
20	0.75	0.254	0.596	0.490	0.045	*0.042**	0.371	1.254
	1.5	0.091	0.693	0.655	0.049	0.048	0.688	1.163
	2.5	0.017	0.829	0.823	0.051	0.048	0.754	0.870

TABLE V. Results of simulation testing of the ROC and JAFROC-2 methods are presented. In all cases we used 2000 NH simulations, 2000 AH simulations, 100 normal and 100 abnormal cases, 5 readers, CRCS structure=HL and BINS=4. CRCS describes the case reader correlation structure, see Table I, and HL is defined in Ref. 19. This is for IICS=Z, i.e., for zero intra-image correlations.

		ROC			ROC	JAFROC-2	ROC	JAFROC-2
T	μ	ILF-NH	θ (ROC)	θ (JAFROC-2)	P_{NH}	P_{NH}	Power-parameter (d)	
3	0.75	0.138	0.600	0.552	0.048	0.050	0.673	1.221
	1.5	0.025	0.723	0.712	0.049	0.049	0.889	1.095
	2.5	0.001	0.860	0.859	0.055	0.055	0.726	0.745
5	0.75	0.233	0.573	0.499	0.053	0.052	0.379	1.174
	1.5	0.070	0.687	0.668	0.053	0.047	0.809	1.107
	2.5	0.001	0.834	0.833	0.053	0.055	0.756	0.797
20	0.75	0.370	0.529	0.381	0.050	0.048	0.150	1.189
	1.5	0.106	0.604	0.554	0.047	0.046	0.605	1.214
	2.5	0.015	0.763	0.758	0.052	0.054	0.895	0.999

pseudo-ROC ratings submitted to MRMC consisted of 200 ''normal'' ratings (100 of which came from actually normal images and the remaining from actually abnormal images) and 100 signal ratings. The MRMC program treated the data set as originating from 200 normal cases and 100 abnormal cases, and jackknifed each of these cases to construct the pseudovalue matrix, which had 300 entries for each reader and modality. We found that the AFROC method generally failed the NH test—it was too conservative, typically rejecting the NH only about 1%–2% of the time, instead of the expected 5%.

D. NH behavior of JAFROC-1

Recall that in the JAFROC-1 method one uses the HN responses from all images, including the abnormal cases. The analysis program jackknifed 100 normal cases and 100 abnormal cases, removing two values (signal and HN) for each abnormal case, and the pseudovalue matrix had 200 entries for each reader and modality. While the p values were closer to 5% than with the AFROC method, the JAFROC-1 method also failed the NH test.

E. NH behavior of JAFROC-2 versus ROC

The JAFROC-2 method differs from JAFROC-1 essentially in the definition of θ. The jackknifing scheme is identical to that in JAFROC-1. This change resulted in a significant improvement in the NH behavior over that observed for JAFROC-1, and in view of this we performed a detailed comparison of the NH and power characteristics of this method and the ROC (i.e., modified MRMC) method. Tables III–VII present the results of this testing for one lesion per abnormal case. The conditions varied to generate these tables are listed in the first two-columns. They are T, the total number of sites per case evaluated by the reader, and μ, the separation of the signal and noise distributions shown in Fig. 3. Each row of the body of the tables was generated using a unique seed variable (not listed) which guaranteed that the same random ratings data sequences were used to assess the ROC and JAFROC-2 analysis methods. In columns 4 and 5 we list the mean figure-of-merit θ for the two methods. This was calculated without jackknifing any cases (this was referred to as θ_{ij} in Eq. 3) and we averaged this value over all readers, both modalities, and NH and AH conditions. The mean figure-of-merit θ is a measure of the task difficulty. Columns 6 and 7, labeled P_{NH}, shows the observed NH failure rates for the ROC and JAFROC-2 methods respectively, and values in italics indicate when this rate fell outside the 95% confidence limit of the expected value. On 8/60 occasions JAFROC-2 failed the NH test and on 5/60 occasions ROC failed the NH test. Note that even with an ideal algorithm one expects the NH test to fail in 5% of the trials. For 60 trials one expects 3 failures on average, with a 95% confidence range extending from 0 to 6, so the observed rate for JAFROC-2 is higher than expected. The number of NH failures was highest (three) for the large negative intra-image correlations (IICS=NN). If this correlation structure is excluded, the observed NH performance is close to optimal.

F. Power comparison of JAFROC-2 and ROC

The last two columns in Tables III–VII, labeled d, list the observed powers of the two methods, as measured by the effective separation d of the AH and NH distributions. Larger values of d imply greater power of the method at discriminating between the NH and AH conditions. We found that JAFROC-2 has greater power than the ROC method, and the difference was most pronounced for the negative correlation structures (IICS=NN or N) and difficult case samples (e.g., μ=0.75 or 1.5). The power advantage, expressed as a ratio of the d values, is typically a factor or 2 to 3, but can be as large as 8 in some situations. In column 3 of Tables III–VII, labeled ILF-NH, we list the incorrect localization fraction (ILF) under the NH condition for the ROC method (for the JAFROC-2 method this quantity is zero). We found that the power ratio was related to ILF-NH. This quantity is the fraction of abnormal cases where the HN rating exceeds the signal rating on the abnormal images. In other words, if ILF =0.10, then in 10 out of 100 abnormal cases a noise location was rated higher than a signal location, and would have led to incorrect localizations. The term ''incorrect localization'' owes its origin to the fact that the observer has rated a noise

TABLE VI. Results of simulation testing of the ROC and JAFROC-2 methods are presented. In all cases we used 2000 NH simulations, 2000 AH simulations, 100 normal and 100 abnormal cases, 5 readers, CRCS structure=HL and BINS=4. CRCS describes the case reader correlation structure, see Table I, and HL is defined in Ref. 19. This is for IICS=P, i.e., for positive intra-image correlations.

		ROC			ROC	JAFROC-2	ROC	JAFROC-2
T	μ	ILF-NH	θ (ROC)	θ (JAFROC-2)	P_{NH}	P_{NH}	Power-parameter (d)	
3	0.75	0.084	0.627	0.599	0.044	0.047	0.683	1.107
	1.5	0.008	0.751	0.747	0.053	0.049	0.892	0.993
	2.5	0.000	0.880	0.880	0.053	0.053	0.689	0.695
5	0.75	0.122	0.606	0.565	0.057	0.054	0.598	1.203
	1.5	0.018	0.726	0.719	0.053	0.054	0.904	1.046
	2.5	0.001	0.864	0.864	0.055	0.057	0.821	0.830
20	0.75	0.197	0.564	0.489	0.052	0.052	0.370	1.200
	1.5	0.048	0.670	0.654	0.052	0.055	0.839	1.114
	2.5	0.001	0.823	0.823	0.046	0.044	0.839	0.861

location higher than a signal location on the same case. This leads to the classic scoring ambiguity of the ROC-paradigm,[9] which rewards a reader who makes two canceling errors (a false positive and a false negative) on the same case. In contrast FROC analysis penalizes the observer for both errors, and it is perhaps not surprising that cases where incorrect localization occurs tend to accentuate the difference between the two methods. Figure 4 shows that the power advantage correlates well with ILF-NH.

Based on Fig. 4, the conclusion that JAFROC-2 has higher power than ROC (i.e., power advantage ratio >1) appears to be valid at an extremely high significance level (i.e., low p value). To quantify this statistically a paired t-test of d (JAFROC-2) vs d (ROC) was performed for all the data shown in Tables III–VII. The observed p value was 2.7×10^{-9}. We also performed 40 simulations of ROC and JAFROC-2 with varying seeds for $T=5$, IICS=N, $\mu=0.75$ and BINS=4 (this resulted in an ILF of 0.185, roughly in the middle of the range shown in Fig. 4) and tested the resulting power advantage ratio values with a two-tailed z test against the value of 1.0, obtaining a p value of 0 (mean=2.25, standard deviation=0.0979, and z statistic=12.8).

G. Reasons for differences in NH behavior of the methods

To further investigate the reasons for the failure of the NH with some methods, we computed the average histogram (normalized to unit count) of the pseudovalues over 20 trials and both treatment conditions, under the NH condition. Each trial returned $2\times5\times N$ pseudovalues, where N was 200, 300, 200, and 200 for ROC, AFROC, JAFROC-1, and JAFROC-2, respectively. The factor of 2 is from the two modalities and 5 from the number of readers. The binning interval for the histogram was 0.2 and separate histograms were averaged for normal and abnormal cases. With 20 trials the observed histogram had negligible sampling error. Figure 5 shows the observed NH histogram for $T=3$, IICS=N, one signal per abnormal case, 100 normal+100 abnormal cases, 5 readers, $\mu=0.75$, for the ROC analysis methods. The jagged structure visible in these plots is related to the finite number of bins employed (BINS=4, corresponding to a 5-rating ROC study or a 4-rating FROC study), which causes peaks in the observed pseudovalue histogram, and makes it somewhat difficult to see trends. Therefore we also calcu-

TABLE VII. Results of simulation testing of the ROC and JAFROC-2 methods are presented. In all cases we used 2000 NH simulations, 2000 AH simulations, 100 normal and 100 abnormal cases, 5 readers, CRCS structure=HL and BINS=4. CRCS describes the case reader correlation structure, see Table I, and HL is defined in Ref. 19. This is for IICS=PP, i.e., for strong positive intra-image correlations. Cells with asterisks represent failed NH-tests.

		ROC			ROC	JAFROC-2	ROC	JAFROC-2
T	μ	ILF-NH	θ (ROC)	θ (JAFROC-2)	P_{NH}	P_{NH}	Power-parameter (d)	
3	0.75	0.047	0.616	0.600	0.056	0.053	0.856	1.182
	1.5	0.005	0.750	0.749	0.048	0.048	1.014	1.045
	2.5	0.000	0.880	0.880	0.051	0.051	0.720	0.720
5	0.75	0.073	0.593	0.567	0.048	0.051	0.682	1.155
	1.5	0.009	0.724	0.722	*0.041**	*0.041**	1.048	1.112
	2.5	0.000	0.865	0.865	0.050	0.050	0.778	0.778
20	0.75	0.152	0.549	0.493	0.050	0.044	0.450	1.203
	1.5	0.021	0.665	0.659	0.046	0.048	0.910	1.071
	2.5	0.000	0.825	0.825	0.050	0.050	0.833	0.835

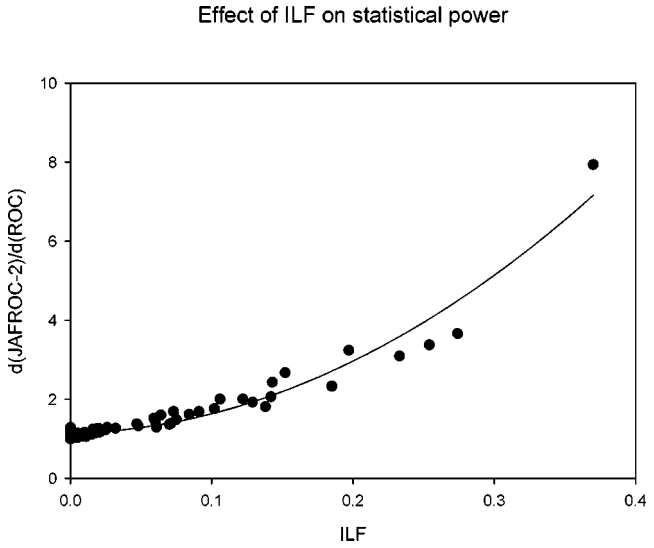

FIG. 4. Data from Tables III to VII plotted to show the dependence of the ratio d (JAFROC-2)/d (ROC) on the incorrect localization fraction (ILF), where the latter is defined as the fraction of times a noise site was rated higher than a signal site on an abnormal image. The quantity d is the separation of the NH and AH distributions, see Fig. 2. It is a measure of the statistical power of the analysis method. Note that d (JAFROC-2) is always greater than d (ROC), and as ILF increases, the ratio approaches 9. This shows the power advantage of the JAFROC-2 method over ROC analysis.

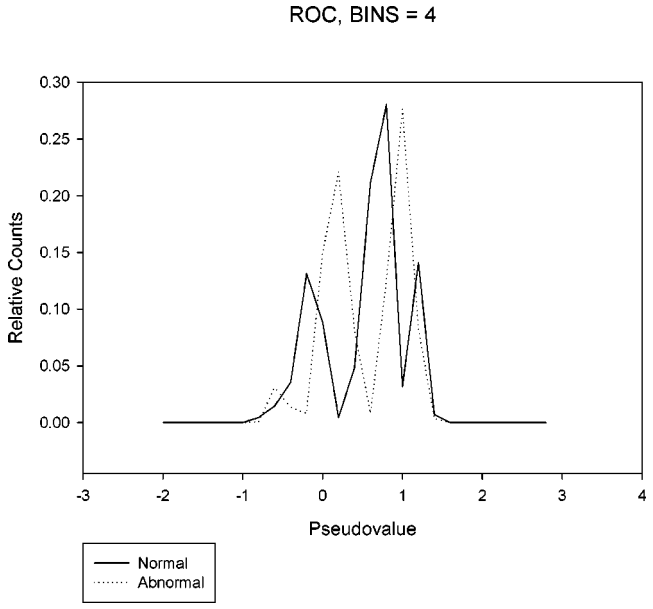

FIG. 5. A histogram of the ROC pseudovalues for BINS=4 computed under the NH of no modality difference. Each pseudovalue measures the effect on the ROC area of removing a single case [see Eq. (3)]. See Sec. III G for more details. The two plots are for normal and abnormal cases. The jagged structure is due to the small number of ratings categories (4) employed, which causes certain pseudovalues to be more likely, and which makes trends difficult to see. Therefore in succeeding plots we set BINS=20, to remove this structure.

lated the histograms for BINS=20, and the results are shown in Figs. 6(a)–6(d), for the following analyses: (a)=ROC, (b)=AFROC, (c)=JAFROC-1, and (d)=JAFROC-2. It should be noted that the ANOVA linear model for the pseudovalues has no dependence on the truth, and to the extent that the normal and abnormal case histograms have different shapes (e.g., have different variances) the pseudovalues are violating the ANOVA model. It is seen that the normal and abnormal case histograms are negligibly different for ROC analysis, that they are grossly different for AFROC analysis, and that they are closer but still significantly different for JAFROC-1 analysis, and that JAFROC-2 yields histograms that are similar. The pseudovalues are also violating the normality assumption of the ANOVA but, based on past experience, this is expected to be less important than the violation of equality of variances (Dr. Steve Hillis, private communication).

H. Explanations

Examination of the results shown in Tables III–VII can yield additional insights. These results can be understood as follows.

1. θ (ROC)$\geqslant\theta$ (JAFROC-2)

The FOM quantity θ shown in these tables is a measure of task difficulty. For the ROC method it ranges from 0.5 to 1.0, and for the JAFROC-2 method it ranges from 0 to 1.0. Note that θ (ROC) always exceeds the corresponding θ (JAFROC-2) and that the differences are larger for smaller values of μ. Recall that θ (ROC) was defined as the probability of an abnormal case ROC-rating exceeding a normal case ROC-rating. The abnormal ROC-case rating is the higher of the signal or the HN rating on the same abnormal image, and the normal case ROC-rating is the HN rating on the normal case. By contrast θ (JAFROC-2) was defined as the probability of the signal rating exceeding the HN rating on normal images. Careful contemplation of these definitions should convince one that θ (ROC)$\geqslant\theta$ (JAFROC-2). When μ is small the HN event on abnormal cases can more often exceed the signal rating (the frequency of this occurring is measured by the incorrect localization fraction quantity, ILF). Therefore it is expected that the difference θ (ROC)-θ (JAFROC-2) will be larger for smaller μ as observed in Tables III–VII.

2. θ increases with μ

It is observed that both θ (ROC) and θ (JAFROC-2) increase as μ increases and that ILF decreases as μ increases. Since μ is the separation of the case noise and signal distributions (Fig. 3) increasing μ is always expected to increase the probability of the signal rating exceeding the noise rating. One also expects that the ILF will decrease as μ increases as fewer HN events will exceed the signal rating.

3. Dependence of θ on T

It is observed that as T increases both (ROC) and θ (JAFROC-2) decrease and ILF increases. Recall that T measures the total number of sites that the observer must evaluate, including the signal site on each abnormal image. In other words, with $T=3$, the observer makes decisions on 3

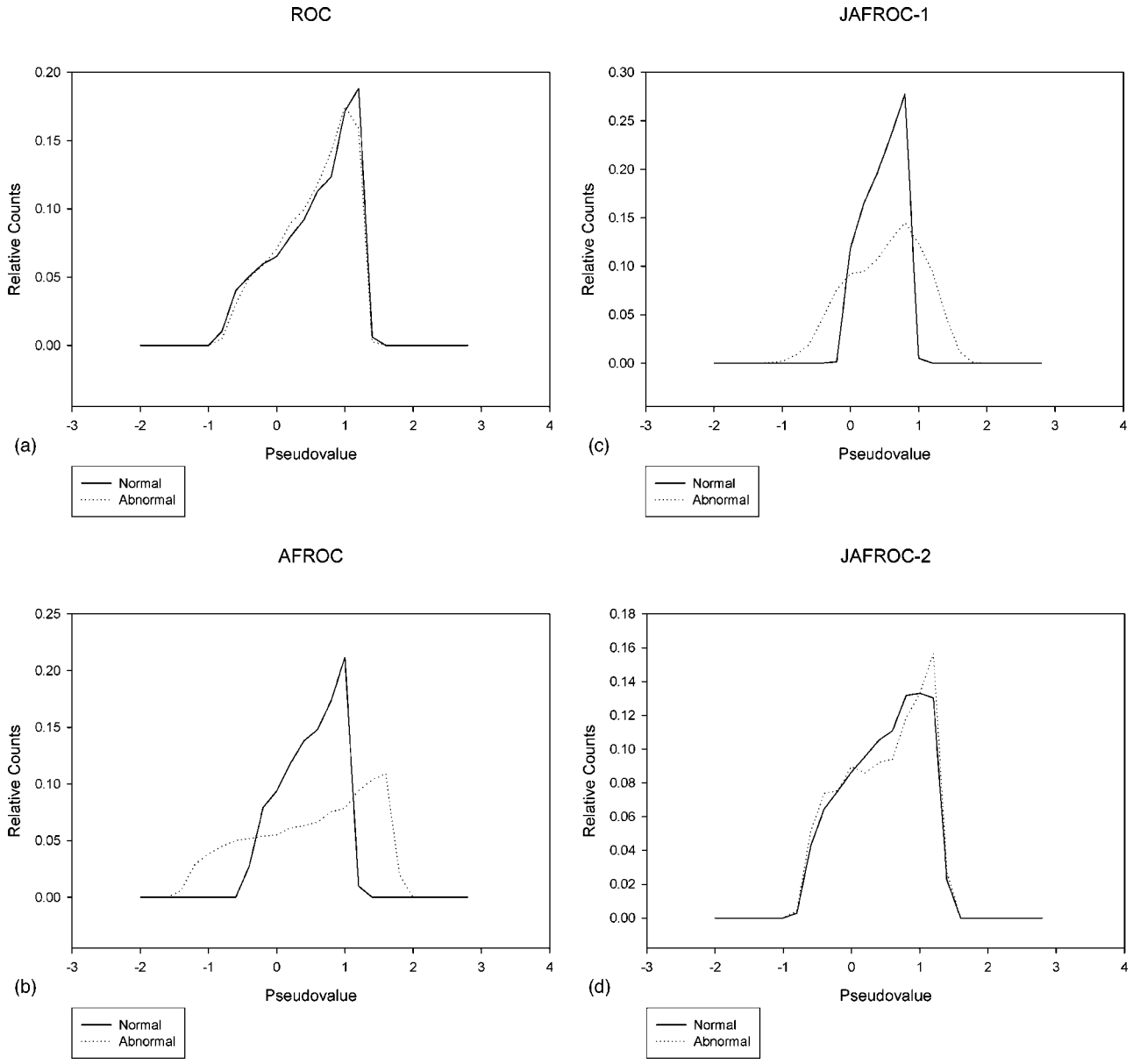

FIG. 6. (a) Histogram of the ROC pseudovalues for BINS=20. The jagged structure is no longer seen due to large number of ratings categories employed. While the plots do not support normally distributed pseudovalues, the normal and abnormal distributions are seen to be indistinguishable. (b) Histogram of the AFROC pseudovalues for BINS=20. In addition to problems with normality the normal and abnormal distributions are seen to be quite different, which violates a central assumption of the MRMC algorithm. This is the reason AFROC analysis is no longer recommended. (c) Histogram of the JAFROC-1 pseudovalues for BINS=20. The normal and abnormal distributions are still quite different, which is why JAFROC-1 is not recommended either. (d) Histogram of the JAFROC-2 pseudovalues for BINS=20. The normal and abnormal distributions are now almost indistinguishable. This is the likely explanation for the observed satisfactory statistical behavior of JAFROC-2.

noise sites on normal images and 2 noise sites and 1 signal site on abnormal images. This observer is more experienced than the observer modeled by $T=5$ who makes decisions on 5 noise sites on abnormal images and 4 noise sites and 1 signal site on abnormal images. This difference in expertise is reflected in both θ (ROC) and θ (JAFROC-2), but the difference in expertise is seen to have a larger effect on θ (JAFROC-2), since the FROC scoring credits the observer only for true signal detections, as opposed to signal or HN detections with ROC scoring. These results confirm what had been alluded to before, that all other conditions being equal, the novice who looks at more locations than the expert will generate more FPs and smaller θ values.

4. Dependence of θ (JAFROC-2) on intra-image correlations

It is observed from Tables III–VII that θ (JAFROC-2) increases as the magnitude of the intra-image correlations (IICS) increases, and that it is independent of the sign of the

correlation. Because in JAFROC-2 analysis one ignores the HN rating on abnormal cases, the HN and signal samples are from different cases. This means that the parameter ρ_{SN} in Fig. 3, which describes the correlation of noise and signal ratings within the same image, should have no effect on θ (JAFROC-2). From its definition θ (JAFROC-2) is the probability that a sample from the LS distribution (Fig. 3) exceeds all samples from the LN distribution. With zero intra-image correlations the T noise samples on a normal case become independent, so P(lesion rating exceeds all noise rating)$\sim$[P(lesion rating exceeds one noise rating)]^T. With highly correlated samples P(lesion rating exceeds all noise ratings)$\sim$P(lesion rating exceeds one noise rating), which is larger than the corresponding value for zero correlations. This explains why θ (JAFROC-2) increases with the magnitude fo ρ_{SN}.

5. Dependence of ILF on intra-image correlations

It is observed from Tables III–VII that all other factors being the same, the Incorrect Localization Fraction (ILF) increases monotonically for IICS=PP,P,Z in that order, and is approximately constant thereafter. The ILF is the probability that a sample from the LS distribution in Fig. 3 is smaller than the HN sample from the LN distribution, where all samples are from the same case. A detailed explantation of this result is given in Appendix C.

IV. DISCUSSION

We have described a simulation model for generating ratings data such as are observed in FROC experiments. An important parameter in the model is the total number of sites per case, T. This parameter, which can be measured using eye-position recording, reflects the expertise of the reader, the case difficulty, and the task. The model parameter μ, defined as the separations of the two case distributions in Fig. 3, will also depend on the expertise of the reader (expert readers have a better idea of what the lesion is supposed to look like), the case difficulty (e.g., lesion contrast, edge sharpness, etc.), and the task (e.g., larger lesions will yield a larger μ). The μ parameter is analogous to the corresponding parameter that enters the conventional ROC binormal model. Note that the new model allows reader expertise at two distinct levels, which could be independent. For example, an experienced reader with failing eye-sight will know where not to look (i.e., T will be small) but μ will also be small (inability to optimally process the information in the image). According to this thinking observer performance measured using the well-known signal-known-exactly location-known-exactly (SKE/LKE) two-alternative forced choice (2AFC) paradigm would reflect μ, but not T.

The simulation model accommodates correlations between ratings on the same image. In the special case of $T=1$ the model reverts to the conventional ROC binormal model. Unlike the standard ROC binormal model this simulation model is capable of encompassing a wider range of observer performance phenomena. Some of these phenomena, e.g., multiple responses per case, have been the subject of the present work, but the model can also describe phenomena such as satisfaction-of-search.[33–36] In fact, the assumption we made in this study that T (the number of noise and signal sites) was independent of the truth status has a built-in SOS-effect, as it implies that the number of noise sites on lesion-containing images is smaller.

We have applied the simulation model to test the ROC, AFROC, and two jackknife methods for analyzing detection-localization data. Tested were null hypothesis (NH) and alternative hypothesis (AH) properties. In all cases we used simple definitions of the figure-of-merit θ that quantified reader performance for a single modality that did not involve any curve fitting. In the ROC case the θ measure was equivalent to the trapezoidal area under the ROC curve. The analysis methods differed in how they handled multiple responses per case and localization information. The ROC method ignored the location information and forced the reader to reduce multiple responses to a single response. All three free-response methods used the highest noise rating on each case; lower rated noise responses were ignored. The figure of merit quantity θ was identical for all free-response methods. The jackknife-FROC approaches differed from the AFROC method in how they estimated the variability of θ. The AFROC method regarded all lesion and highest noise responses as independent. In the first jackknife-FROC method (JAFROC-1) all highest noise responses were used in the analysis. In the second jackknife-FROC method (JAFROC-2) only the highest noise responses from normal cases were used in the analysis.

We found that both the ROC and JAFROC-2 methods passed the NH test, and that the AFROC and JAFROC-1 methods failed the NH test. The JAFROC-2 method consistently outperformed the ROC (modified MRMC) method in statistical power. The power advantage was especially pronounced in situations where the number of incorrect localizations (noise sites rated higher than lesion sites) was large. Due to its poor NH properties, use of the AFROC method is no longer recommended because it is too conservative in rejecting the NH. Past investigators who have used the AFROC method and rejected the NH at the 5% level (i.e., they found a modality difference) were probably rejecting it at a smaller p value. While this is a conservative error, it risks undue loss of power and these investigators may have missed actual differences (type II errors).

Regarding the validation of JAFROC-2, each simulation conducted (5 readers reading 200 images in both modalities) represents a controlled study that is dependent only on the validity of the model and its software implementation. The results shown in Tables III–VII represent close to 100 000 simulations under varying conditions of model parameters. In the vast majority of these simulations the JAFROC-2 method yielded higher power. The model is based on current thinking about how observers scan images for lesions. A consideration that argues against software errors is that the simulation results agree with what is expected from our understanding of the model—see in particular Sec. III H and Appendix C. To further preclude software errors we are making available (on the website http://jafroc.radiology.pitt.edu)

the entire simulation software that was used to generate these results. This should enable any one who wishes to test the model to do so with a minimum need to recode the algorithms. We know of no comparable testing of other methods—indeed, as has been noted in Sec. I such testing was not feasible prior to the introduction of the FROC simulation model in 2002.[3]

Based on this work we believe several avenues remain to be explored. A localization model needs to be developed that captures the number, spatial spread, and average proximity of noise sites to signal sites. This will allow other proposed methods of analyzing detection-localization data, e.g., the LROC and ROI methods to be evaluated. Methods that can handle multiple indications per case are needed for evaluations of CAD algorithms for CT lung-screening studies because such algorithms provide multiple indications per case. ROI and free response methods require different types of data, so before evaluation studies are conducted, there needs to be some consensus on the evaluation method to be used. Recent publications[37,8,38] recommend the ROI method for such studies. Our problem with the ROI paradigm is that the task is fundamentally different from how experts interpret images. Studies need to be performed on the ROI method using a detection and localization simulation model that accounts for the number of sites evaluated by radiologists. That radiologists and CAD algorithms may make multiple responses within an investigator-specified region and that some investigator-specified sites may not be as likely to draw the expert's attention as others has not been included in the limited evaluation[37] of the ROI method that has been conducted.

A similar cautionary note applies to the usage of the LROC paradigm. This method has not been evaluated using a detection and localization model. The limited evaluation that has appeared has consisted of applications to a few clinical data sets. For the reasons noted in Sec. I, this is not an acceptable way of validating a method—since there is no way of knowing the error rates α and β. It is not widely appreciated that LROC analysis also makes an independence assumption. Specifically, it assumes that the probability distribution function for the HN decision variable is independent of the number of signals present [see the development leading to Eq. (2) in Swensson's work[5]]. The LROC experimental paradigm also involves a clinically unrealistic data acquisition paradigm—a clinical report is *not* equivalent to a rating and a forced localization of the most suspicious region. An additional unrealistic aspect is that the reader is forced to indicate a location even when they are highly confident that the image is normal. While the LROC analysis program has recently been revised to not require a location response when the rating is smaller than an investigator-specified value,[39,40] we are not aware of any validation of this program.

Our rather extensive simulation studies clearly suggest that as noise sites become more highly rated than signal sites, the JAFROC-2 method outdistances the ROC method in statistical power. In experiments in which noise sites are not often rated higher that signal sites, the JAFROC-2 method offers the same statistical power as the ROC method. Of course, the experimenter may not know in advance which of these conditions may obtain in an experiment. If not, the safer choice of method may be the JAFROC-2. This conclusion must be tempered by the consideration that it rests on the findings of a single paper. We consider the conclusion tentative because ours is the first and only simulation study to demonstrate a statistical advantage for JAFROC-2. At this time, only the individual experimenter can determine whether using the more conventional method is worth the added risk of failing to detect a real difference.

ACKNOWLEDGMENTS

This work was supported in part by a grant from the Department of Health and Human Services, National Institutes of Health, R01-CA75145 and 8 RO1-EB002120. D.P.C. is grateful to Dr. Claudia Mello-Thoms and Dr. Harold Kundel for stimulating discussions and for providing Fig. 1, and to Ben Herman, MS, for help with compilation of the MRMC code, and to Dr. Steve Hillis for pointing out a simpler derivation of the results given in Appendix A. The simulation and analysis software is available on the following website: http://jafroc.radiology.pitt.edu. The first author is grateful to Hong-Jan Yoon, MS, for creating and maintaining this website.

APPENDIX A

Let $N(\mu,\sigma)$ denote the normal distribution with mean μ and standard deviation σ and let $N_2(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$ denote the bivariate normal distribution with mean-parameters μ_1 and μ_2, standard deviations σ_1 and σ_2, and correlation ρ. Let us denote a random variable ξ sampled from the distribution Δ by $\xi\sim\Delta$. According to the sampling model, depicted schematically in Fig. 3, the net decision variable (x,y) is obtained by pairwise addition of the location samples (γ,δ) and the case samples (ξ,Ψ). One has

$$\gamma\sim N(0,\sigma_{\mathrm{LN}}),\quad \delta\sim N(0,\sigma_{\mathrm{LS}}),$$
$$(\xi,\Psi)\sim N_2(0,\mu,\sigma_{\mathrm{CN}},\sigma_{\mathrm{CS}},\rho_{\mathrm{SN}}). \tag{A1}$$

One can regard γ and δ as samples from a bivariate distribution with zero correlation

$$(\gamma,\delta)\sim N_2(0,0,\sigma_{\mathrm{LN}},\sigma_{\mathrm{LS}},0). \tag{A2}$$

The net decision variable (x,y) is the sum of corresponding samples from the two bivariate distributions. It is a standard result in statistics that the sum of two bivariate distributed random variables is itself bivariate distributed, with covariance matrix equal to the sum of the individual covariance matrices, and mean equal to the sum of the individual means. Therefore,

$$(x,y)\sim N_2(0,\mu,\sigma_x,\sigma_y,\rho_{xy}). \tag{A3}$$

The covariance matrices are given by

$$\sum_{\xi,\psi}=\begin{bmatrix}\sigma_{CN}^2 & \rho_{SN}\sigma_{CN}\sigma_{CS}\\ \rho_{SN}\sigma_{CN}\sigma_{CS} & \sigma_{CS}^2\end{bmatrix},$$

$$\sum_{\gamma,\delta}=\begin{bmatrix}\sigma_{LN}^2 & 0\\ 0 & \sigma_{LS}^2\end{bmatrix}, \tag{A4}$$

$$\sum_{x,y}=\begin{bmatrix}\sigma_x^2 & \rho_{xy}\sigma_x\sigma_y\\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2\end{bmatrix}.$$

Using the addition property of the covariance matrices one sees that

$$\sigma_x^2=\sigma_{CN}^2+\sigma_{LN}^2,\quad \sigma_y^2=\sigma_{CS}^2+\sigma_{LS}^2,$$
$$\rho_{xy}=\rho_{SN}\sigma_{CN}\sigma_{CS}/(\sigma_x\sigma_y). \tag{A5}$$

We note that Eq. (A3) and Fig. 3 apply to *x*, *y* samples from the *same* case. For samples from *different* cases $(k,q,k\neq q)$, one has

$$(x_k,y_q)\sim N_2(0,\mu,\sigma_x,\sigma_y,0)\quad \text{for } (k\neq q). \tag{A6}$$

APPENDIX B

The variance structures introduced by Roe and Metz specify values for the variance-components entering the ROC-decision variable model. In particular, each variance structure specifies the value of the pure case-variance, referred to as σ_C^2 in the Roe and Metz publication. This appendix describes how we generated samples according to the sampling model described in Sec. II A 3 and subject to the constraint that the net variance of noise (and signal) events are equal to the specified value σ_C^2.

In our simulations we set $\sigma_{CN}=\sigma_{CS}=\sigma_G$ and $\sigma_{LN}=1$. We defined F_0 and F_1 by

$$F_0=\sqrt{\frac{\sigma_C^2}{1+\sigma_G^2}},\quad F_1=\sqrt{\frac{\sigma_C^2}{\sigma_{LS}^2+\sigma_G^2}}. \tag{B1}$$

The case samples were produced as shown in the following:

$$(\xi,\psi)=N_2(0,\mu,\sigma_GF_0,\sigma_GF_1,\rho_{SN}). \tag{B2}$$

The local samples were produced as shown in the following:

$$(\gamma,\delta)=N_2(0,\mu,F_0,\sigma_{LS}F_1,0). \tag{B3}$$

Since the net noise and signal samples *x*, *y* are given by the pairwise sums of the case and location samples, then using the results in Appendix A it can be easily confirmed that the variances of *x*, *y* are each given by σ_C^2.

APPENDIX C

We had noted earlier that the Incorrect Localization Fraction (ILF) was least for highly positively correlated ratings and increased as the correlations become more negative. The ILF is a measure of how many times the noise random variable *x* exceeds the signal random variable *y* on the same abnormal case. Consider the random variable *x*-*y* which is distributed as $N(-\mu,\ \sigma_{x\text{-}y})$. The fraction of this distribution that exceeds zero, which is a measure of ILF, decreases as the ratio $\mu/\sigma_{x\text{-}y}$ becomes larger, since the distribution of *x*-*y* becomes narrower and further below zero, so that *x*-*y* is less likely to exceed zero. For the *Z* intra-image correlation structure one has $\sigma_x=1$, $\sigma_y=1$, $\rho_{xy}=0$, and using the above formulae $\sigma_{x\text{-}y}\sim\sqrt{2}$ and the ratio of $\mu/\sigma_{x\text{-}y}\sim\mu/\sqrt{2}$. If the correlation was perfect $(\rho_{xy}=1)$ the *x* and *y* samples would vary in perfect unison and their difference *x*-*y* would be constant $(\sigma_{x\text{-}y}=0)$ and the ratio of $\mu/\sigma_{x\text{-}y}$ would be infinite. This explains why ILF is least for the PP correlation structure. For perfect negative correlations $(\rho_{xy}=-1)$ the *x* and *y* samples will vary opposite to each other in perfect unison, and the difference *x*-*y* would be more variable (i.e., $\sigma_{x\text{-}y}$ is large) and $\mu/\sigma_{x\text{-}y}$ would be small. This explains why ILF is greatest for the NN correlation structure. The simplistic arguments above do not take into account the multiple *x*-samples occurring on the same abnormal case and the fact that one needs to consider the highest of them to determine the overlap with the signal sample. These factors presumably lead to the *T*-dependence of the ILF-results in Table III–VII.

[a]This work was partially presented at the 2002 Meeting of the Radiological Society of North America and at the 2003 Meeting of the Medical Image Perception Society.

[b]Electronic mail: dpc10@pitt.edu

[1] R. Wagner, S. V. Beiden, G. Campbell, C. E. Metz, and W. M. Sacks, "Assessment of medical imaging and computer-assist systems: Lessons from recent experience," Acad. Radiol. **9**, 1264–1277 (2002).

[2] D. D. Dorfman, K. S. Berbaum, and C. E. Metz, "ROC characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method," Invest. Radiol. **27**, 723–731 (1992).

[3] D. P. Chakraborty, "Statistical power in observer performance studies: A comparison of the ROC and free-response methods in tasks involving localization," Acad. Radiol. **9**, 147–156 (2002).

[4] S. J. Starr, C. E. Metz, L. B. Lusted, and D. J. Goodenough, "Visual detection and localization of radiographic images," Radiology **116**, 533–538 (1975).

[5] R. G. Swensson, "Unified measurement of observer performance in detecting and localizing target objects on images," Med. Phys. **23**, 1709–1725 (1996).

[6] N. A. Obuchowski, M. L. Lieber, and K. A. Powell, "Data analysis for detection and localization of multiple abnormalities with application to mammography," Acad. Radiol. **7**, 516–525 (2000).

[7] K. O. Hajian-Tilaki, J. A. Hanley, L. Joseph, and J. P. Collet, "Extension of receiver operating characteristic analysis to data concerning multiple signal detection tasks," Acad. Radiol. **4**, 222–229 (1997).

[8] C. M. Rutter, "Bootstrap estimation of diagnostic accuracy with patient-clustered data," Acad. Radiol. **7**, 413–419 (2000).

[9] P. C. Bunch, J. F. Hamilton, G. K. Sanderson, and A. H. Simmons, "A free-response approach to the measurement and characterization of radiographic-observer performance," J. Appl. Photogr. Eng. **4**, 166–171 (1978).

[10] D. P. Chakraborty, "Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data," Med. Phys. **16**, 561–568 (1989).

[11] D. P. Chakraborty and L. Winter, "Free-response methodology: Alternate analysis and a new observer-performance experiment," Radiology **174**, 873–881 (1990).

[12] C. E. Metz, "Evaluation of digital mammography by ROC analysis," in *Digital Mammography '96*, edited by K. Doi *et al.* (Elsevier Science, Amsterdam, 1996).

[13] D. P. Chakraborty, "Data analysis for detection and localization of multiple abnormalities with application to mammography. [letter; comment]," Acad. Radiol. **7**, 553–554 (2000); discussion 554–556.

[14] E. A. Krupinski, "Visual scanning patterns of radiologists searching mammograms," Acad. Radiol. **3**, 137–144 (1996).

[15] H. L. Kundel and P. S. J. La Follette, "Visual search patterns and experience with radiological images," Radiology **103**, 523–528 (1972).

[16] A. Hillstrom, ''Repetition effects in visual search,'' Percept. Psychophys. **2**, 800–817 (2000).
[17] C. E. Metz and H. Kronman, ''Statistical significance tests for binormal ROC curves,'' J. Math. Psychol. **22**, 218–242 (1980).
[18] J. A. Hanley, ''The robustness of the 'binormal' assumptions used in fitting ROC curves,'' Med. Decis. Making **8**, 197–203 (1988).
[19] C. Roe and C. E. Metz, ''Dorfman–Berbaum–Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: Validation with computer simulation,'' Acad. Radiol. **4**, 298–303 (1997).
[20] D. D. Dorfman, K. S. Berbaum, R. V. Lenth, Y.-F. Chen, and B. A. Donaghy, ''Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: Factorial experimental design,'' Acad. Radiol. **5**, 591–602 (1998).
[21] D. D. Dorfman, K. S. Berbaum, R. V. Lenth, and Y.-F. Chen, ''Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: Split plot experimental design,'' in *Proceedings of SPIE, Medical Imaging, Image Perception and Performance*, 1999.
[22] C. H. Hu, H. L. Kundel, C. F. Nodine, E. A. Krupinski, and L. C. Toto, ''Searching for bone fractures: A comparison with pulmonary nodule search,'' Acad. Radiol. **1**, 25–32 (1994).
[23] H. L. Kundel, C. F. Nodine, and D. Carmody, ''Visual scanning, pattern recognition and decision-making in pulmonary nodule detection,'' Invest. Radiol. **13**, 175–181 (1978).
[24] C. F. Nodine and H. L. Kundel, ''Using eye movements to study visual search and to improve tumor detection,'' Radiographics **7**, 1241–1250 (1987).
[25] R. E. Hendrick *et al.*, *Mammography Quality Control Manual*, 4th ed. (American College of Radiology, Committee on Quality Assurance in Mammography, 1999).
[26] D. P. Chakraborty, ''Proposed solution to the FROC problem and an invitation to collaborate,'' Proc. SPIE **5034**, 204–212 (2003).
[27] D. Dorfman and E. J. Alf, ''Maximum likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—rating method data,'' J. Math. Psychol. **6**, 487–496 (1969).
[28] J. A. Hanley and B. J. McNeil, ''A method of comparing the areas under receiver operating characteristic curves derived from the same cases,'' Radiology **148**, 839–843 (1983).
[29] C. Roe and C. E. Metz, ''Variance-component modeling in the analysis of receiver operating characteristic index estimates,'' Acad. Radiol. **4**, 587–600 (1997).
[30] A. E. Burgess, ''Comparison of receiver operating characteristic and forced choice observer performance measurement methods,'' Med. Phys. **22**, 643–655 (1995).
[31] W. H. Press, B. P. Flannery, S. A. Teulosky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge University Press, Cambridge, 1988), p. 735.
[32] ICRU, *Medical Imaging: The Assessment of Image Quality* (International Commission on Radiation Units and Measurements, Bethesda, MD, 1996), Vol. 54.
[33] K. S. Berbaum, E. A. Franken, D. D. Dorfman, S. A. Rooholamini, M. H. Kathol, T. J. Barloon, F. M. Behlke, Y. Sato, C. H. Lu, G. Y. El-Khoury, F. W. Flickinger, and W. J. Montgomery, ''Satisfaction of search in diagnostic radiology,'' Invest. Radiol. **25**, 133–140 (1990).
[34] K. S. Berbaum, G. Y. El-Khoury, E. A. Franken, D. M. Kuehn, D. M. Meis, D. D. Dorfman, N. G. Warnock, B. H. Thompson, S. Kao, and M. H. Kathol, ''Missed fractures resulting from satisfaction of search effect,'' Emerg. Radiol. **1**, 242–249 (1994).
[35] K. S. Berbaum, E. A. Franken, D. D. Dorfman, E. M. Miller, R. T. Caldwell, D. M. Kuehn, and M. L. Berbaum, ''Role of faulty visual search in the satisfaction of search effect in chest radiography,'' Acad. Radiol. **5**, 9–19 (1998).
[36] S. Samuel, H. L. Kundel, C. F. Nodine, and L. C. Toto, ''Mechanism of satisfaction of search: Eye position recordings in the reading of chest radiographs,'' Radiology **194**, 895–902 (1995).
[37] N. A. Obuchowski, M. L. Lieber, and K. A. Powell, ''Data analysis for detection and localization of multiple abnormalities with application to mammography [see comments],'' Acad. Radiol. **7**, 516–525 (2000).
[38] X.-H. Zhou, D. K. McClish, and N. A. Obuchowski, *Statistical Methods in Diagnostic Medicine* (Wiley-Interscience, New York, 2002).
[39] R. G. Swensson, G. Maitz, J. L. King, and D. Gur, ''Using incomplete and imprecise localization data on images to improve estimates of detection accuracy,'' Proc. SPIE **3663**, 74–81 (1999).
[40] R. G. Swensson, J. L. King, and D. Gur, ''A constrained formulation for the receiver operating characteristic (ROC) curve based on probability summation,'' Med. Phys. **28**, 1597–1609 (2001).