

Maximum Likelihood Fitting of FROC Curves Under an Initial-Detection-and-Candidate-Analysis Model

Darrin C. Edwards

*Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology,
The University of Chicago, Chicago, IL 60637*

Matthew A. Kupinski

*Optical Sciences Center, Department of Radiology,
The University of Arizona, Tucson, AZ 85721*

Charles E. Metz and Robert M. Nishikawa

*Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology,
The University of Chicago, Chicago, IL 60637*

(Received 11 June 2002; accepted for publication 6 October 2002; published 27 November 2002)

We have developed a model for FROC curve fitting that relates the observer's FROC performance not to the ROC performance that would be obtained if the observer's responses were scored on a per image basis, but rather to a hypothesized ROC performance that the observer would obtain in the task of classifying a set of "candidate detections" as positive or negative. We adopt the assumptions of the Bunch FROC model, namely that the observer's detections are all mutually independent, as well as assumptions qualitatively similar to, but different in nature from, those made by Chakraborty in his AFROC scoring methodology. Under the assumptions of our model, we show that the observer's FROC performance is a linearly scaled version of the candidate analysis ROC curve, where the scaling factors are just given by the FROC operating point coordinates for detecting initial candidates. Further, we show that the likelihood function of the model parameters given observational data takes on a simple form, and we develop a maximum likelihood method for fitting a FROC curve to this data. FROC and AFROC curves are produced for computer vision observer datasets and compared with the results of the AFROC scoring method. Although developed primarily with computer vision schemes in mind, we hope that the methodology presented here will prove worthy of further study in other applications as well. © 2002 American Association of Physicists in Medicine. [DOI: 10.1118/1.1524631]

Keywords: ROC, FROC, AFROC, curve fitting, maximum likelihood

I. INTRODUCTION

Well known and widely accepted methods exist for obtaining maximum likelihood (ML) estimates of curve parameters for fitting data acquired in receiver operating characteristic (ROC) experiments.^{1,2} Similar methods for relating free-response operating characteristic (FROC) curves to the ROC curves that would be obtained by disregarding location information in a given experiment,^{3,4} as well as for fitting FROC curves directly,⁵ are also well known, if perhaps less widely accepted.

Some of the reluctance in accepting such FROC methods may be due to the additional assumptions that must be imposed beyond those involved in ROC analysis — namely, the assumptions that the detections in an image are statistically independent, that the occurrences of signals in images are independent events, and that the false-positive detections follow a particular model (usually Poisson). For example, work by Swensson⁶ on a "location-response" operating characteristic (LROC) formulation related to the FROC model suggested that, at least for human observers, independence of detections held only when the observer failed to find a signal present in an image; that is, a normal location was less likely to be reported if a signal were detected than if it were not. On the other hand, recent work by Chakraborty⁷ demonstrated that greater statistical power to detect differences between modalities was available from free-response analysis than from ROC analysis.

We have developed a model for fitting a restricted type of FROC curve. This model does not relax any of the assumptions stated above, but it may be more flexible than previous methods, because it does not relate the observer's FROC performance to the ROC performance that would be obtained if the observer's responses were scored on a per image basis. Instead, it relates the observer's performance to the ROC performance that the observer would obtain in the task of classifying a set of "candidate detections" as positive or negative. (This phrase will be explained in Sec. II.) This model was inspired primarily by our experience in working with computer vision schemes for computer-aided diagnosis; in particular, we believe that the assumption that the observer's detections be statistically independent may be more tenable for computerized observers than for human observers.

In Sec. II A, we introduce a model for a particular type of observer who produces response data in an FROC experiment. In Sec. II B, we show how error bars on this response data can be calculated easily in a manner analogous to the calculation of error bars on standard ROC data. The method to calculate the FROC curve parameters is developed in Sec. II C. In Secs. III and IV, we describe the application of this method to actual datasets and present the results obtained. Finally, we discuss some implications of the methodology and our experimental results in Sec. V, followed by our conclusions in Sec. VI.

II. THEORY

A. The IDCA FROC model

Receiver operating characteristic (ROC) analysis is the most widely accepted methodology for characterizing the performance of an observer in a two-class classification task. The observer’s decision as to whether a given case is in the “actually-positive” class or the “actually-negative” class is presumed to depend on the critical value adopted for a latent continuous decision variable. The true-positive fraction (TPF), defined as the probability that an actually-positive case is decided positive, is then a function of this critical value, as is the false-positive fraction (FPF), the probability that an actually-negative case is decided positive. The observer’s performance — the parametric curve (FPF(t), TPF(t)) where t is the critical value of the latent decision variable — is dependent on only two probability density functions (PDFs). Furthermore, flexible and empirically valid models for these PDFs can be constructed. Thus, it is possible to apply general results from signal detection theory to problems in ROC analysis; *e. g.*, the performance of the ideal observer given the data PDFs may be determined.⁸ If it is not practical or not possible to obtain reliable continuously distributed decision variables from the observer, as may be the case with a human observer, one may seek to estimate the observer’s performance from a partition of the t -axis, referred to as a set of categories. Since methods have been developed for obtaining ML estimates of an ROC curve from either categorical¹ or continuously-distributed data,⁹ this distinction is of limited practical significance.

It is well known^{1,2} that an ML fit of a set of category data obtained in an ROC experiment can be achieved by

1. assuming a model (*e. g.*, the conventional binormal model, or a “proper” likelihood ratio model derived from an underlying pair of normal distributions) for the PDFs of the latent decision variables which generate the category data;
2. assigning probabilities to the categories (*e. g.*, by cutpoints on the latent decision variable PDFs); and then
3. finding values for the model parameters and the category cutpoints which maximize the likelihood function

$$\mathbf{L}_{\text{ROC}} = M! N! \prod_i \frac{p_i^{\mathbf{j}_i} q_i^{\mathbf{k}_i}}{\mathbf{j}_i! \mathbf{k}_i!}. \quad (1)$$

In Eq. (1), M and N are the numbers of actually negative and actually positive cases, respectively, used in the experiment; p_i represents the probability of an actually negative case, and q_i the probability of an actually positive case, being assigned to category i ; and \mathbf{j}_i and \mathbf{k}_i are the number of actually negative and actually positive cases, respectively, assigned to category i . We use boldface type to indicate statistically variable quantities. Metz, Herman, and Shen have shown that a likelihood function with the form shown in Eq. (1) can be used to fit an ROC curve to continuously-distributed data as well.⁹

The situation becomes more complicated when one considers free-response operating characteristic (FROC) experiments. Here the observer is shown a set of I images containing a total of N true signals, and must produce for each image a set of locations where the observer believes that signals may be present. We can consider each image to be divided into a large number of subregions, where each subregion is small enough that at most one detection can be made within it.¹⁰ (That is, the number of subregions is much larger than the expected number of signals or detections in an image.) The image-based detection problem is then reduced to a classification task within each subregion. Bunch *et al.* showed^{3,4} that if the observer uses a constant critical value of a latent decision variable t to make decisions for all detections, the observer’s performance can be characterized by two quantities: the signal detection fraction (SDF), defined similarly to the TPF above as the probability that an actually positive signal is detected; and the expectation value of the number of false-positive detections per image (FPpI). (The effect of a statistically variable critical value has been shown¹¹ to be equivalent, in the context of ROC analysis, to an increase in the widths of the latent decision variable’s conditional PDFs; we have not included this effect in our model.) Although the observer’s performance is still completely characterizable by a single parametric curve, namely (FPpI(t), SDF(t)), it is no longer clear that

this curve depends only on two PDFs; it is conceivable that if the image properties are sufficiently different in each subregion (as in a typical medical image), we will need as many PDFs as subregions to completely model t .

Bunch^{3,4} proposed an FROC model which makes the assumption that the observer's detections are all mutually independent. This implies that the number of true-positive (TP) detections in each image follows a binomial distribution, while the number of false-positive (FP) detections approaches a Poisson distribution (the validity of the Poisson approximation, given that detections are assumed independent of one another, increases with the number of subregions into which the image is divided). Under these assumptions it is possible to relate the observer's FROC curve, given by $(\text{FPpI}(t), \text{SDF}(t))$, to the ROC curve obtained by considering an image with one or more detections to be a "true-positive" if it contains any actual signals, and to be a "false-positive" if it contains no actual signals but one or more false-positive detections. Furthermore, Bunch showed that the observer's FROC curve could also be related to a different formulation, which Chakraborty⁵ later called the alternative free-response operating characteristic (AFROC). In this formulation, the observer's performance is characterized by the SDF as defined in the FROC formulation, and by the probability of obtaining one or more FP detections in an image (also referred to as $P(\text{FP} \geq 1)$, the "probability of a false-positive image"). The AFROC formulation is useful because it provides a "hybrid" characterization of performance with respect to both signals and images. Furthermore, under the assumptions of the Bunch model, the relationship between the FROC curve and the AFROC curve is particularly simple.

Let us now consider a somewhat simplified (and therefore admittedly nongeneral) experiment in which initial candidate detections are first located in the image by the observer. The observer then, in a second stage of the detection process, considers each individual candidate in turn and produces a decision variable value for that candidate detection. The initial candidate selection can be thought of as corresponding to a particular operating point in a postulated "candidate" FROC space. That is, the candidate selection process involves some parameter or parameters which could in principle be varied to produce different such operating points in this space. In the model here, however, we assume that those parameters have been chosen in some way and are fixed for the following analysis. We shall refer to this model as the "initial detection and candidate analysis" (IDCA) model.

This model was motivated primarily by the schemes for computer-aided diagnosis developed in our laboratory. Typically in such schemes, a filter is applied to the image to locate a large number of candidate detections. More sophisticated feature analysis techniques are then applied to each candidate to reduce the number of false-positive detections. In general, the choice of operating parameters, and indeed the methodology, of these stages will effectively be independent. Furthermore, many more free parameters are available in a filter than in a typical feature analysis technique. Therefore, the candidate detection filter is usually considered fixed for our purposes, and optimization efforts are focused on the feature analysis stage of the scheme.

Let the operating point of the candidate detection stage be $(\text{FPpI}_c, \text{SDF}_c)$, and let the performance of the analysis stage for a corresponding population of selected candidates be given by an ROC curve $(\text{FPF}(t), \text{TPF}(t))$. Notice that for a given decision threshold t , the observer's SDF is just the conditional probability that a TP decision is made for a given candidate detection, multiplied by the probability that the candidate was detected in the first place:

$$\begin{aligned} \text{SDF}(t) &\equiv P(\text{signal detected as candidate and decided "+"}) \\ &= P(\text{"+" cand. decided "+"} | \text{sig. detected as cand.}) \times P(\text{sig. detected as cand.}) \\ &= \text{TPF}(t) \times \text{SDF}_c. \end{aligned} \tag{2}$$

It is important to note here that the parameters involved in the candidate selection process, *i. e.*, the choice of FPpI_c and SDF_c , will strongly influence the "spectrum" of cases presented to the observer, and thus the values $\text{FPF}(t)$ and $\text{TPF}(t)$ will depend on the candidate selection parameters. Although we will not indicate this explicitly, since in our simplified model the candidate selection parameters are regarded as "fixed" for the duration of the experiment, this point should be kept in mind. Also note that, under this model, the observer's decision criterion t is assumed *not* to influence the detection rates of the set of initial candidates. Again, the parameters governing these initial candidate detection rates are assumed fixed.

Similarly, the false-positive rate $\text{FPpI}(t)$ can be evaluated in terms of $\text{FPF}(t)$ and FPpI_c :

$$\begin{aligned} \text{FPpI}(t) &\equiv \langle \text{No. of "-" detections in 1 image} \rangle \\ &= \sum_f f P(f \text{ "-" candidates detected and decided "+"}) \\ &= \sum_f f \sum_{g \geq f} P(f \text{ out of } g \text{ "-" cands. decided "+"} | g \text{ "-" cands.}) \times P(g \text{ "-" cands.}) \\ &= \sum_g P(g \text{ "-" cands.}) \sum_{f \leq g} f P(f \text{ out of } g \text{ "-" cands. decided "+"} | g \text{ "-" cands.}) \end{aligned}$$

$$\begin{aligned}
&= \sum_g P(g \text{ “-” cand.}) g \text{ FPF}(t) \\
&= \text{FPF}(t) \times \text{FPpI}_c.
\end{aligned} \tag{3}$$

(Here f and g are particular numbers of candidate detections, and the angular brackets indicate that the expectation value of the enclosed expression is being taken.) Thus the IDCA model implies that the overall FROC performance is a linearly scaled version of the analysis stage ROC curve, where the scaling factors are given by the candidate detection FROC operating point coordinates.

We claim that with the IDCA model, a curve can be fit to the observer’s category data in an ML sense by

1. assuming a model (here binormal) for the latent decision variable PDFs of the candidate population (the population of initial candidates which are detectable given the fixed parameters of the candidate selection process);
2. assigning probabilities to the categories generated by the latent decision variables;
3. assuming a model (here Poisson) for the distribution of the number of actually-negative candidate detections, and a model (here binomial) for the distribution of the number of actually-positive candidate detections; and then
4. finding values for the latent decision variable model parameters, the category cutpoints, and the initial candidate detection rates $\widehat{\text{SDF}}_c$ and $\widehat{\text{FPpI}}_c$ which maximize a likelihood function \mathbf{L}_{FROC} .

Note that $\widehat{\text{SDF}}_c$ and $\widehat{\text{FPpI}}_c$ must be considered unknown *a priori* in the sense that their values SDF_c and FPpI_c for the data population must in practice be estimated from finite samples. The actual number N of true signals in the dataset, and the number I of images in the dataset, are taken as known, however.

In the remainder of this section, we show that, under the IDCA model, error bars on the original points in the FROC plot can be defined in a way similar to the ROC case. We will also show that the likelihood function \mathbf{L}_{FROC} has a fairly simple form, and is maximized by maximizing the equivalent \mathbf{L}_{ROC} for the ROC decision task implied by points (1) and (2) above.

B. Operating Point Estimation

In order to make our discussion of FROC operating point statistics clearer, we will first briefly review the equivalent results for ROC operating points. In the typical model for an ROC experiment, the observer produces a set of category data \mathbf{j} for M actually negative cases such that \mathbf{j}_i is the number of responses in category i for the actually negative cases, and a second set \mathbf{k} for N actually positive cases such that \mathbf{k}_i is the number of responses in category i for the actually positive cases. The categories i are a discrete ordinal set of possible responses indicating the observer’s “confidence” that a case is positive; *e. g.*, a number from 1 through 5, where 1 denotes “most likely negative” and 5 denotes “most likely positive”.

The response vectors are statistically variable in the sense that we can repeat the experiment with another sample of M actually-negative cases and N actually-positive cases, obtaining in general different category responses (but drawn from the same PDFs). To estimate an operating point on the observer’s ROC curve, we select a category number i_0 above which responses are interpreted as “positive” (allowing a “null” category such that all other category responses are interpreted as positive when this null category is chosen as i_0), and count the number \mathbf{J} of actually negative cases and \mathbf{K} of actually positive cases interpreted as positive:

$$\mathbf{J} \equiv \sum_{i>i_0} \mathbf{j}_i, \tag{4}$$

$$\mathbf{K} \equiv \sum_{i>i_0} \mathbf{k}_i. \tag{5}$$

Since the cases and category decisions for each case are statistically independent by hypothesis, it follows that the probabilities of obtaining particular values of \mathbf{J} and \mathbf{K} are binomial:

$$P(\mathbf{J} = J) = \frac{M!}{J!(M-J)!} \text{FPF}(i_0)^J [1 - \text{FPF}(i_0)]^{M-J}, \tag{6}$$

$$P(\mathbf{K} = K) = \frac{N!}{K!(N-K)!} \text{TPF}(i_0)^K [1 - \text{TPF}(i_0)]^{N-K}, \tag{7}$$

in which $\text{FPF}(i_0)$ and $\text{TPF}(i_0)$ are the expected values of FPF and TPF that are obtained when categorical responses above i_0 are interpreted as “positive” — *i. e.*, $\text{FPF}(t)$ and $\text{TPF}(t)$ when t is set equal to the upper boundary of category i_0 on the latent decision-variable axis.

Equations (6) and (7) allow us to write expressions for the means and variances of the true- and false-positive counts:

$$\langle \mathbf{J} \rangle = M \text{FPF}(i_0), \quad (8)$$

$$\langle (\mathbf{J} - \langle \mathbf{J} \rangle)^2 \rangle = M \text{FPF}(i_0) [1 - \text{FPF}(i_0)], \quad (9)$$

$$\langle \mathbf{K} \rangle = N \text{TPF}(i_0), \quad (10)$$

$$\langle (\mathbf{K} - \langle \mathbf{K} \rangle)^2 \rangle = N \text{TPF}(i_0) [1 - \text{TPF}(i_0)]. \quad (11)$$

From the above it follows immediately that $\widehat{\mathbf{FPF}}(i_0) \equiv \mathbf{J}/M$ is an unbiased estimator of $\text{FPF}(i_0)$ with variance $\text{FPF}(i_0)[1 - \text{FPF}(i_0)]/M$, and $\widehat{\mathbf{TPF}}(i_0) \equiv \mathbf{K}/N$ is an unbiased estimator of $\text{TPF}(i_0)$ with variance $\text{TPF}(i_0)[1 - \text{TPF}(i_0)]/N$. Furthermore,

$$\hat{\sigma}_{\widehat{\mathbf{FPF}}(i_0)}^2 \equiv \frac{1}{M-1} \left(\frac{\mathbf{J}}{M} \right) \left(1 - \frac{\mathbf{J}}{M} \right) \quad (12)$$

is an unbiased estimator of the variance of $\widehat{\mathbf{FPF}}(i_0)$, and

$$\hat{\sigma}_{\widehat{\mathbf{TPF}}(i_0)}^2 \equiv \frac{1}{N-1} \left(\frac{\mathbf{K}}{N} \right) \left(1 - \frac{\mathbf{K}}{N} \right) \quad (13)$$

is an unbiased estimator of the variance of $\widehat{\mathbf{TPF}}(i_0)$. (The factors $M-1$ and $N-1$ are needed to make the estimates unbiased.) The expressions in Eqs. (12) and (13) can be used to plot error bars (plus or minus two standard errors) for the points in an ROC plot, as was done in Fig. 1 for a set of simulated category data.

It is straightforward to show that the above results can be applied directly to estimation of operating points on an AFROC curve if all responses are statistically independent. To achieve this, we must replace M with I , the total number of images; and $\text{FPF}(i_0)$ with $P(\text{FP} \geq 1)[i_0]$, the probability of a false-positive image when the confidence threshold is i_0 . Furthermore, N must now be interpreted as the number of actually positive signals present across all the images; $\vec{\mathbf{k}}$ is the vector of category responses for the TP detected signals; and $\vec{\mathbf{j}}$ is a vector of category responses such that \mathbf{j}_i is the number of *images* in which i is the highest category response among the FP detections in that image.

The situation is slightly more complicated in the IDCA FROC case. Consider a set of I images randomly sampled from a population, subject to the criterion that there are a total of N signals in the set. We take as our experimental outcome the category response vector $\vec{\mathbf{r}}$ for the set of \mathbf{B} actually negative candidate detections, and the category response vector $\vec{\mathbf{s}}$ for the set of \mathbf{C} actually positive candidate detections. That is, the element \mathbf{r}_i represents the number of detections in category i due to actually negative candidates, whereas the element \mathbf{s}_i represents the number of detections in category i due to actually positive candidates. Note that the category response vectors, as well as the total numbers of candidate detections \mathbf{B} and \mathbf{C} , are statistically variable; we can repeat the experiment with another sample of I images containing N signals, and obtain in general different candidate detection counts and category responses for the detected candidates (but drawn from the same PDFs).

Let us select a category number i_0 above which responses are interpreted as “positive” (allowing a “null” category such that all other category responses are interpreted as positive when this null category is chosen as i_0), and count the number \mathbf{R} of detections scored as FPs, and the number \mathbf{S} of detections scored as TPs:

$$\mathbf{R} \equiv \sum_{i>i_0} \mathbf{r}_i, \quad (14)$$

$$\mathbf{S} \equiv \sum_{i>i_0} \mathbf{s}_i. \quad (15)$$

If we consider a given set of B actually negative and C actually positive candidate detections, then reasoning as in the ROC situation above it should be clear that

$$P(\mathbf{R} = R | \mathbf{B} = B) = \frac{B!}{R! (B-R)!} \text{FPF}(i_0)^R [1 - \text{FPF}(i_0)]^{B-R}, \quad (16)$$

$$P(\mathbf{S} = S | \mathbf{C} = C) = \frac{C!}{S! (C-S)!} \text{TPF}(i_0)^S [1 - \text{TPF}(i_0)]^{C-S} \quad (17)$$

if all responses are statistically independent. Here $\text{TPF}(i_0)$ and $\text{FPF}(i_0)$ are the respective probabilities of a TP or FP candidate detection having a response category above i_0 ; intuitively, this is the ROC performance of the observer in the restricted task of distinguishing between TP and FP signals for the population of candidate detections.

Figure 2 shows a schematic representation of a hypothetical image in which three signals are present (indicated by plus signs) and five candidate detections have been located by the observer (indicated by circles). Three of the detections are actually negative ($B = 3$) and two are actually positive ($C = 2$); next to each detection is indicated the category i reported by the observer for that detection. At a threshold of $i_0 = 2$, we have $R = 2$ and $S = 2$ from Eqs. (14) and (15); thus the SDF of the observer at this threshold is 0.67 and the FPPi is 2 (for this single image).

In order to obtain expressions for the unconditional PDFs of \mathbf{R} and \mathbf{S} , we must sum over the conditional values

$$P(R) = \sum_{B=R}^{\infty} \frac{B!}{R!(B-R)!} \text{FPF}(i_0)^R [1 - \text{FPF}(i_0)]^{B-R} P(B), \quad (18)$$

$$P(S) = \sum_{C=S}^N \frac{C!}{S!(C-S)!} \text{TPF}(i_0)^S [1 - \text{TPF}(i_0)]^{C-S} P(C). \quad (19)$$

In order to evaluate these sums explicitly, we require models for $P(B)$ and $P(C)$. We will adopt a Poisson model for the distribution of \mathbf{B} , with mean and variance $I \cdot \text{FPpI}_c$. In practice such a model may be flawed; for example there is likely to be an upper limit to the possible number of candidate detections that can be made in a given image (it is unlikely to be greater than the number of pixels in the image, for example). Nevertheless, if we wish to adhere to the assumption that different candidate detections are statistically independent, then the Poisson model (which can be thought of as a binomial model in the limit of very small probability of an event occurring and very large number of independent trials in which the event may occur) is the most tractable. By similar reasoning we will take as our model for $P(C)$ a binomial distribution in which each of a set of N signals has a probability SDF_c of being detected. The mean of \mathbf{C} is $N \cdot \text{SDF}_c$, and the variance is $N \cdot \text{SDF}_c(1 - \text{SDF}_c)$.

Using the well known relation that $\langle g(\mathbf{x}) \rangle = \langle \langle g(\mathbf{x}) | \mathbf{y} \rangle_x \rangle_y$,¹² we can easily evaluate the mean and variance of \mathbf{R} :

$$\begin{aligned} \langle \mathbf{R} \rangle &= \langle \langle \mathbf{R} | \mathbf{B} \rangle_R \rangle_B \\ &= \langle \mathbf{B} \cdot \text{FPF}(i_0) \rangle_B \\ &= I \cdot \text{FPpI}_c \cdot \text{FPF}(i_0); \\ \langle (\mathbf{R} - \langle \mathbf{R} \rangle)^2 \rangle &= \langle \mathbf{R}^2 \rangle - \langle \mathbf{R} \rangle^2 \\ &= \langle \langle \mathbf{R}^2 | \mathbf{B} \rangle_R \rangle_B - \langle \mathbf{R} \rangle^2 \\ &= \langle \mathbf{B} \cdot \text{FPF}(i_0) [1 - \text{FPF}(i_0)] + \mathbf{B}^2 \cdot \text{FPF}(i_0)^2 \rangle - [I \cdot \text{FPpI}_c \cdot \text{FPF}(i_0)]^2 \\ &= I \cdot \text{FPpI}_c \cdot \text{FPF}(i_0) [1 - \text{FPF}(i_0)] + (I \cdot \text{FPpI}_c + I^2 \cdot \text{FPpI}_c^2) \text{FPF}(i_0)^2 \\ &\quad - [I \cdot \text{FPpI}_c \cdot \text{FPF}(i_0)]^2 \\ &= I \cdot \text{FPpI}_c \cdot \text{FPF}(i_0) [1 - \text{FPF}(i_0)] + I \cdot \text{FPpI}_c \cdot \text{FPF}(i_0)^2 \\ &= I \cdot \text{FPpI}_c \cdot \text{FPF}(i_0). \end{aligned} \quad (20)$$

The calculation of the mean and variance of \mathbf{S} is similar:

$$\begin{aligned} \langle \mathbf{S} \rangle &= \langle \langle \mathbf{S} | \mathbf{C} \rangle_S \rangle_C \\ &= \langle \mathbf{C} \cdot \text{TPF}(i_0) \rangle_C \\ &= N \cdot \text{SDF}_c \cdot \text{TPF}(i_0); \\ \langle (\mathbf{S} - \langle \mathbf{S} \rangle)^2 \rangle &= \langle \mathbf{S}^2 \rangle - \langle \mathbf{S} \rangle^2 \\ &= \langle \langle \mathbf{S}^2 | \mathbf{C} \rangle_S \rangle_C - \langle \mathbf{S} \rangle^2 \\ &= \langle \mathbf{C} \cdot \text{TPF}(i_0) [1 - \text{TPF}(i_0)] + \mathbf{C}^2 \text{TPF}(i_0)^2 \rangle - [N \cdot \text{SDF}_c \cdot \text{TPF}(i_0)]^2 \\ &= N \cdot \text{SDF}_c \cdot \text{TPF}(i_0) [1 - \text{TPF}(i_0)] \\ &\quad + (N \cdot \text{SDF}_c [1 - \text{SDF}_c] + N^2 \cdot \text{SDF}_c^2) \text{TPF}(i_0)^2 - [N \cdot \text{SDF}_c \cdot \text{TPF}(i_0)]^2 \\ &= N \cdot \text{SDF}_c \cdot \text{TPF}(i_0) [1 - \text{TPF}(i_0)] + N \cdot \text{SDF}_c [1 - \text{SDF}_c] \text{TPF}(i_0)^2 \\ &= N \cdot \text{SDF}_c \cdot \text{TPF}(i_0) [1 - \text{SDF}_c \cdot \text{TPF}(i_0)]. \end{aligned} \quad (22)$$

From Eqs. (20) and (22), it is clear that

$$\widehat{\text{FPpI}}(i_0) \equiv \mathbf{R}/I \quad (24)$$

is an unbiased estimator of $\text{FPpI}_c \cdot \text{FPF}(i_0)$, the overall FP detection rate per image of the observer at the confidence threshold i_0 ; and similarly,

$$\widehat{\text{SDF}}(i_0) \equiv \mathbf{S}/N \quad (25)$$

is an unbiased estimator of $\text{SDF}_c \cdot \text{TPF}(i_0)$, the overall SDF rate of the observer at the confidence threshold i_0 . Furthermore, note that

$$\hat{\sigma}_{\text{FPpI}}^2(i_0) \equiv \frac{\mathbf{R}}{I^2} \quad (26)$$

$$\begin{aligned} \langle \hat{\sigma}_{\text{FPpI}}^2(i_0) \rangle &= \left\langle \frac{\mathbf{R}}{I^2} \right\rangle \\ &= \frac{\text{FPpI}_c \cdot \text{FPF}(i_0)}{I}, \end{aligned} \quad (27)$$

implying that $\hat{\sigma}_{\text{FPpI}}^2(i_0)$ is an unbiased estimator of the variance of $\widehat{\text{FPpI}}(i_0)$. Similarly,

$$\hat{\sigma}_{\text{SDF}}^2(i_0) \equiv \frac{1}{N-1} \left(\frac{\mathbf{S}}{N} \right) \left(1 - \frac{\mathbf{S}}{N} \right) \quad (28)$$

$$\begin{aligned} \langle \hat{\sigma}_{\text{SDF}}^2(i_0) \rangle &\equiv \left\langle \frac{1}{N-1} \left(\frac{\mathbf{S}}{N} \right) \left(1 - \frac{\mathbf{S}}{N} \right) \right\rangle \\ &= \frac{1}{N-1} \left\langle \frac{\mathbf{S}}{N} - \frac{\mathbf{S}^2}{N^2} \right\rangle \\ &= \frac{1}{N-1} \left(\text{SDF}_c \cdot \text{TPF}(i_0) - \frac{N \cdot \text{SDF}_c \cdot \text{TPF}(i_0) + N(N-1)\text{SDF}_c^2 \cdot \text{TPF}(i_0)^2}{N^2} \right) \\ &= \frac{1}{N-1} \left[\frac{N-1}{N} (\text{SDF}_c \cdot \text{TPF}(i_0) - \text{SDF}_c^2 \cdot \text{TPF}(i_0)^2) \right] \\ &= \frac{\text{SDF}_c \cdot \text{TPF}(i_0) [1 - \text{SDF}_c \cdot \text{TPF}(i_0)]}{N}, \end{aligned} \quad (29)$$

implying that $\hat{\sigma}_{\text{SDF}}^2(i_0)$ is an unbiased estimator of the variance of $\widehat{\text{SDF}}(i_0)$. The expressions in Eqs. (26) and (28) were used to produce the error bars on the FROC operating points shown in Fig. 3, which was made using the same simulated data as in Fig. 1.

Note that if one makes only the Bunch assumptions in an FROC model, one can derive the same expressions for the means and variances of estimators of the operating points as those given in Eqs. (24), (25), (26), and (28), provided we recall Eqs. (2) and (3); *i. e.*, $\text{SDF}(i_0) = \text{TPF}(i_0) \cdot \text{SDF}_c$, and $\text{FPpI}(i_0) = \text{FPF}(i_0) \cdot \text{FPpI}_c$. In this sense the IDCA FROC model can be viewed as a special case of the Bunch model.

C. ML Estimation of IDCA FROC Curve Parameters

In Sec. II A, the standard methodology for fitting an ROC curve to measured observer category data was reviewed briefly. Given a model for the distributions of the observer's decision variables, a relationship between the decision variable distributions and the measured category data, and the assumption of case-to-case independence of observer decisions, one obtains the likelihood function given in Eq. (1). For a given set of measured category data, this likelihood function is then maximized to yield ML estimates of the ROC curve parameters.^{1,2,9} Figure 1 shows the category data fitted using both the LABROC4 method⁹ and the PROPROC method.²

Chakraborty⁵ proposed a method for fitting FROC curves under the Bunch model by making two additional assumptions: first, that the observer's latent decision variable for the set of TP detections follows a normal distribution (or a monotonic transformation thereof); and second, that the observer's latent decision variable for the set of "most salient" FP detections per image (that is, the FP detection in each image with the highest value of the decision variable) also follows a normal distribution (or the same monotonic transformation thereof). This results in a model which is formally equivalent to the familiar binormal ROC model,^{1,9} but applicable to AFROC space rather than ROC space. The proposed method, referred to as AFROC scoring, originally involved transforming the resulting AFROC curve to FROC space using the transformation appropriate to the Bunch model,^{3,4} namely $P(\text{FP} \geq 1) = 1 - e^{-\text{FPpI}}$.

More recently, it has been recommended that AFROC scoring results be reported directly in terms of the curve fitted in AFROC space, rather than the curve obtained in FROC space *via* the above transformation.¹³

In our model, in order to obtain the ML estimate of the IDCA FROC curve parameters given a set of measured observer category data, we require an explicit expression for the likelihood of the data given the curve parameters. As in the preceding section, we will find it easier to construct this total probability from simpler conditional probability expressions.

Let

$$\mathbf{L}_{\text{FROC}} \equiv P(\vec{\mathbf{r}}, \vec{\mathbf{s}}) \quad (30)$$

represent the likelihood of obtaining a set of category responses $\vec{\mathbf{r}}$ for the actually negative candidate detections and, jointly, a set of category responses $\vec{\mathbf{s}}$ for the actually positive candidate detections, as defined in the preceding section. \mathbf{L}_{FROC} is implicitly a function of the model parameters to be estimated. Note that this expression can be expanded using Bayes's theorem¹² to obtain

$$\mathbf{L}_{\text{FROC}} = \sum_{B,C} P(\vec{\mathbf{r}}, \vec{\mathbf{s}} | B, C) P(B, C), \quad (31)$$

where B is a particular number of actually negative candidate detections and C is a particular number of actually positive candidate detections. That is, B and C are the counts of the candidate detections initially detected in the set of I images, while $\vec{\mathbf{r}}$ and $\vec{\mathbf{s}}$ contain the category responses for the candidates (each element corresponds to a different category, and the sums of the category counts are the number of actually negative and actually positive candidates, respectively).

Because we are considering the location and category information to be separable in this fashion by hypothesis, the quantity $P(\vec{\mathbf{r}}, \vec{\mathbf{s}} | B, C)$ is formally equivalent to the expression for \mathbf{L}_{ROC} given in Eq. (1). To see this, consider each of the B actually negative candidate detections, and each of the C actually positive candidate detections, to be a single "case"; $\vec{\mathbf{r}}$ and $\vec{\mathbf{s}}$ then represent the category data for these cases. Thus, after altering variable names appropriately in Eq. (1), we can write

$$\mathbf{L}_{\text{FROC}} = \sum_{B,C} B! C! \prod_i \left\{ \frac{p_i^{\mathbf{r}_i} q_i^{\mathbf{s}_i}}{\mathbf{r}_i! \mathbf{s}_i!} \right\} \times P(B, C). \quad (32)$$

Notice, however, that there can be only one term in the sum, namely the term such that $B = \sum_i r_i$ and $C = \sum_i s_i$ (the total actually negative and actually positive candidate counts equal the sums of the respective category counts). Since these variables are no longer being summed over, and since they are sums of the elements of the random vectors $\vec{\mathbf{r}}$ and $\vec{\mathbf{s}}$, we conclude that the candidate counts themselves should be treated as random variables:

$$\mathbf{L}_{\text{FROC}} = \mathbf{B}! \mathbf{C}! \prod_i \left\{ \frac{p_i^{\mathbf{r}_i} q_i^{\mathbf{s}_i}}{\mathbf{r}_i! \mathbf{s}_i!} \right\} P(\mathbf{B}) P(\mathbf{C}). \quad (33)$$

By employing the Poisson model for $P(B)$ and the binomial model for $P(C)$ as stated in Sec. II A, and collecting terms in particular random variables, we arrive at the expression

$$\mathbf{L}_{\text{FROC}} = \prod_i \left\{ \frac{p_i^{\mathbf{r}_i} q_i^{\mathbf{s}_i}}{\mathbf{r}_i! \mathbf{s}_i!} \right\} \times \mathbf{B}! \frac{(I \cdot \text{FPpI}_c)^{\mathbf{B}}}{\mathbf{B}!} e^{-I \cdot \text{FPpI}_c} \times \mathbf{C}! \frac{N!}{\mathbf{C}! (N - \mathbf{C})!} \text{SDF}_c^{\mathbf{C}} (1 - \text{SDF}_c)^{N - \mathbf{C}}. \quad (34)$$

Taking the logarithm of this expression, and deleting terms that do not depend on the quantities to be estimated (the ROC curve parameters implicit in p_i and q_i , and the candidate FROC operating point values SDF_c and FPpI_c), we are left with

$$\begin{aligned} \mathbf{L}\mathbf{L}_{\text{FROC}} \equiv & \sum_i \{ \mathbf{r}_i \ln p_i + \mathbf{s}_i \ln q_i \} + [\mathbf{B} \ln(\text{FPpI}_c) - I \cdot \text{FPpI}_c] \\ & + [\mathbf{C} \ln(\text{SDF}_c) + (N - \mathbf{C}) \ln(1 - \text{SDF}_c)]. \end{aligned} \quad (35)$$

Notice that this sum has three terms: a term which depends only on the category data $\vec{\mathbf{r}}, \vec{\mathbf{s}}$ and the analysis stage ROC model parameters; a term which depends only on the actually negative count data \mathbf{B} and the candidate FPpI_c rate; and a third term which depends only on the actually positive count data \mathbf{C} and the candidate SDF_c rate. Clearly, solving this maximization problem is completely equivalent to solving three separate maximization problems. The first is equivalent to the ROC curve-fitting problem which has been solved.^{1,2} The second is simply

$$\frac{\partial \mathbf{LL}_{\text{FROC}}}{\partial \text{FPpI}_c} = \frac{\mathbf{B}}{\text{FPpI}_c} - I \quad (36)$$

$$\widehat{\text{FPpI}}_c = \frac{\mathbf{B}}{I}, \quad (37)$$

where the estimate is taken as that value of the parameter which maximizes the likelihood function. The third maximization problem is similarly solved by taking

$$\frac{\partial \mathbf{LL}_{\text{FROC}}}{\partial \text{SDF}_c} = \frac{\mathbf{C}}{\text{SDF}_c} - \frac{N - \mathbf{C}}{1 - \text{SDF}_c} \quad (38)$$

$$\widehat{\text{SDF}}_c = \frac{\mathbf{C}}{N}. \quad (39)$$

It is interesting to note that under the particular models we have chosen for \mathbf{B} and \mathbf{C} , the estimators in Eqs. (37) and (39) are unbiased, although the curve parameter estimates themselves may be biased in general.

All that remains is to scale the fitted ROC curve by the estimates $\widehat{\text{SDF}}_c$ and $\widehat{\text{FPpI}}_c$ as required to satisfy Eqs. (2) and (3). Figure 3 shows the application of the above method to the simulated category data used to generate the operating points in that and Fig. 1.

III. MATERIALS AND METHOD

We applied the results developed in the preceding section to two datasets examined previously in our laboratory for the development of computer vision schemes for computer-aided diagnosis. The first dataset, which we will refer to as dataset A, consisted of 43 computed tomography (CT) volume sets containing a total of 171 biopsy-confirmed lung nodules. Computer vision results, in the form of the output of a linear discriminant analysis (LDA) algorithm, were also available for each of 132 actually positive candidates and 7,165 actually negative candidates.¹⁴ A detection was scored as a TP if and only if it contained either the centroid pixel or the maximum-gray-level pixel of an actual nodule; we note this here and for the other datasets because it is impossible to evaluate FROC results without an explicit description of the scoring method.^{15,16}

The second dataset, which we will refer to as dataset B, consisted of 50 mammograms containing 41 biopsy-confirmed clusters of microcalcifications. Computer vision results, in the form of BANN output, were available for the 36 actually positive and 112 actually negative candidates in these mammograms;¹⁷ a detection was scored as a TP if and only if the detected cluster's centroid was within 6mm of the centroid of an actual cluster, and additionally the cluster contained at least 2 TP microcalcifications (defined as detected signals within 0.4mm of a true microcalcification).

For both of these datasets, we computed the estimated operating points, plus and minus two standard errors, in both FROC and AFROC spaces, from the effective category data implied by the computer vision output. We then fitted FROC curves to the data with the IDCA method, using both LABROC4⁹ and PROPROC² for the second (analysis) stage. For comparison, we also used LABROC4 to fit the data directly in AFROC space using the AFROC scoring method.¹³ The IDCA FROC curves were transformed to AFROC space, and the fitted AFROC curves were transformed to FROC space, using the Bunch model transformation ($P(\text{FP} \geq 1) = 1 - e^{-\text{FPpI}}$), to facilitate this comparison.

IV. RESULTS

Figure 4 shows the estimated operating points, plus and minus two standard errors, in AFROC space for dataset A, along with the curve fitted to these points using the AFROC scoring method. Figure 5 shows the estimated operating points, plus and minus two standard errors, in FROC space for the same dataset, along with the IDCA FROC curves fitted to these points using LABROC4 and PROPROC for the analysis stage ROC fit. For comparison, the AFROC scoring curve in Fig. 4 was transformed to FROC space using the Bunch model relation, and is shown in Fig. 5; similarly, the IDCA FROC curves in Fig. 5 were transformed to AFROC space and are shown in Fig. 4.

Similarly, Fig. 6 shows the estimated operating points, plus and minus two standard errors, in AFROC space for dataset B, along with the curve fitted to these points using the AFROC scoring method. Figure 7 shows the estimated operating points, plus and minus two standard errors, in FROC space for the same dataset, along with the IDCA FROC curves fitted to these points using LABROC4 and PROPROC for the analysis stage ROC fit. The AFROC scoring curve in Fig. 6 was transformed to FROC space and is shown in Fig. 7; similarly, the IDCA FROC curves in Fig. 7 were transformed to AFROC space and are shown in Fig. 6.

V. DISCUSSION

Figures 4 and 6 show that the AFROC scoring method fits the operating point data reasonably well in AFROC space for both datasets. However, in Fig. 6, it appears that the IDCA FROC curves, when transformed to AFROC space, actually lie below the majority of the estimated operating points. Conversely, Figs. 5 and 7 seem to show the opposite effect; the IDCA FROC curves appear to fit the estimated operating points in FROC space fairly well, while the transformed AFROC scoring curves are above many of the estimated operating points, especially at high FP rates. This observation can be understood in light of the large numbers of FPs present in these datasets at low observer confidence thresholds, especially in dataset A (Figs. 4 and 5). Note that FP rates of 5 per image and higher correspond to points in AFROC space to the right of the line $P(\text{FP} \geq 1) = 0.99$; thus the majority of points visible in Fig. 4 to the left of this line, which are in a sense “controlling” the AFROC scoring method fit, actually correspond to points in the lower left corner of the corresponding FROC curve (Fig. 5). In an intuitive sense, then, we may regard this “overshooting” on the part of the AFROC scoring method in FROC space as a form of extrapolation error. Similar remarks apply to the curves for dataset B, although the effect is not quite so pronounced.

In Fig. 6, it is evident that the transformed IDCA FROC curves terminate at a point interior to the AFROC unit square, unlike the AFROC scoring curve which passes smoothly to the point (1,1) in AFROC space. This is in fact a direct consequence of the IDCA FROC method, by design, not extrapolating beyond the upper-rightmost estimated operating point in FROC space — that is, the point corresponding to the initial candidate detection operating point. Recall from Sec. II A that the fundamental assumption of the IDCA method is that this initial candidate detection operating point is held fixed by the observer. Although the point itself must be estimated from the observer’s performance on a dataset of finite size, resulting in nonzero error bars on this operating point, the IDCA method cannot extrapolate beyond this point. That is, no matter how lax a criterion the observer adopts in the second (analysis) stage, the most detections (both TP and FP) that can be made are those already present in the set of initial candidates.

We wish to emphasize strongly at this point that the AFROC scoring method is currently recommended to be performed in AFROC space,¹³ and *not* transformed back to FROC space as in the originally proposed FROCFIT method.⁵ Thus the transformed AFROC scoring curves in Figs. 5 and 7 should not be interpreted as any criticism of the AFROC scoring method, but merely an attempt to provide some context for the new method we are developing.

Finally, we would like to reiterate that the IDCA FROC curve fitting method was inspired primarily by our observations of, and conjectures regarding, computer vision methods. In particular, we suspected that the Bunch model assumption regarding the independence of observer detections might be less radical for computer vision methods than for human observers. We currently have no reason to believe that the results of applying the IDCA FROC method to human observer data will be as encouraging, and further work is necessary in order to assess the generality of this model.

VI. CONCLUSION

We have developed an IDCA model for fitting FROC curves to observer data which gives the observer’s FROC curve as a scaled version of a “per-signal” ROC curve, where the scale factors are given by the coordinates of the initial candidate detection operating point in FROC space. Fitting FROC curves to measured observer category data is difficult in general, due primarily to the difficulties in modeling the latent decision variable process in terms of both location and confidence threshold. By explicitly separating these stages in our IDCA model, and by introducing fairly reasonable assumptions concerning the distributions of TP and FP detections in image sets of a given size containing a fixed number of signals, we have been able to make use of previously developed methods for ML fitting of ROC curves. Although this model may prove to be of limited use in directly measuring the performance of human observers, we hope that for a variety of applications, including for example computer vision schemes used in computer-aided diagnosis, the methodology presented here is worthwhile of further study.

ACKNOWLEDGMENTS

Special thanks to Dr. Sam Armato for providing one of the datasets. Thanks to Dr. Phillip Bunch for providing an out-of-print paper, and to Dr. Dev Chakraborty for helpful comments regarding his papers on FROCFIT and AFROC scoring. The first author thanks Ben Herman for invaluable assistance on very short notice with the LABROC4 program. Preliminary results of this work were presented at the Medical Image Perception Conference IX in September, 2001; the first author’s attendance at that conference was supported in part by a grant from the National Cancer Institute. This

work was supported in parts by grant R01-CA60187 from the National Cancer Institute (R.M. Nishikawa, principal investigator) and grant R01-GM57622 from the National Institutes of Health (C.E. Metz, principal investigator). R.M. Nishikawa and C.E. Metz are shareholders in R2 Technology, Inc. (Los Altos, CA). It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interest which would reasonably appear to be directly and significantly affected by the research activities.

- ¹ D. D. Dorfman and E. 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).
- ² C. E. Metz and X. Pan, "'Proper' binormal ROC curves: Theory and maximum-likelihood estimation," *J. Math. Psychol.* **43**, 1–33 (1999).
- ³ 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," in *Application of Optical Instrumentation in Medicine VI*, edited by J. E. Gray and W. R. Hendee, Proc. SPIE **127**, pp. 124–135 (1977).
- ⁴ 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**, 166–172 (1978).
- ⁵ D. P. Chakraborty, "Maximum likelihood analysis of free-response operating characteristic (FROC) data," *Med. Phys.* **16**, 561–568 (1989).
- ⁶ R. G. Swenson, "Unified measurement of observer performance in detecting and localizing target objects on images," *Med. Phys.* **23**, 1709–1725 (1996).
- ⁷ D. Chakraborty, "Statistical power in observer-performance studies: Comparison of the receiver operating characteristic and free-response methods in tasks involving localization," *Acad. Radiol.* **9**, 147–156 (2002).
- ⁸ J. P. Egan, *Signal Detection Theory and ROC Analysis* (Academic Press, New York, 1975).
- ⁹ C. E. Metz, B. A. Herman, and J.-H. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," *Stat. Med.* **17**, 1033–1053 (1998).
- ¹⁰ C. E. Metz, S. J. Starr, and L. B. Lusted, "Observer performance in detecting multiple radiographic signals," *Radiology* **121**, 337–347 (1976).
- ¹¹ D. J. Goodenough and C. E. Metz, "Implications of a 'noisy' observer to data processing techniques," in *Information Processing in Scintigraphy*, edited by C. Raynaud and A. Todd-Pokropek (Commisariat a l'Energie Atomique, Departement de Biologie, Service Hospitalier Frederic Joliot, Orsay, France, 1975), pp. 400–419.
- ¹² A. Papoulis, *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, Inc., New York, 1991).
- ¹³ D. P. Chakraborty, "Chapter 16: The FROC, AFROC, and DROC variants of the ROC analysis," in *Handbook of Medical Imaging, Vol. 1: Physics and Psychophysics*, edited by J. Beutel, H. L. Kundel, and R. L. Van Metter (SPIE, Bellingham, WA, 2000), pp. 771–796.
- ¹⁴ S. G. Armato III, M. L. Giger, and H. MacMahon, "Automated detection of lung nodules in CT scans: Preliminary results," *Med. Phys.* **28**, 1552–1561 (2001).
- ¹⁵ R. M. Nishikawa and L. M. Yarusso, "Variations in measured performance of CAD schemes due to database composition and scoring protocol," in *Medical Imaging 1998: Image Processing*, edited by K. M. Hanson, Proc. SPIE **3338**, pp. 840–844 (1998).
- ¹⁶ M. Kallergi, G. M. Carney, and J. Gaviria, "Evaluating the performance of detection algorithms in digital mammography," *Med. Phys.* **26**, 267–275 (1999).
- ¹⁷ D. C. Edwards, J. Papaioannou, Y. Jiang, M. A. Kupinski, and R. M. Nishikawa, "Eliminating false-positive microcalcification clusters in a mammography CAD scheme using a Bayesian neural network," in *Medical Imaging 2001: Image Processing*, edited by M. Sonka and K. Hanson, Proc. SPIE **4322**, pp. 1954–1960 (2001).

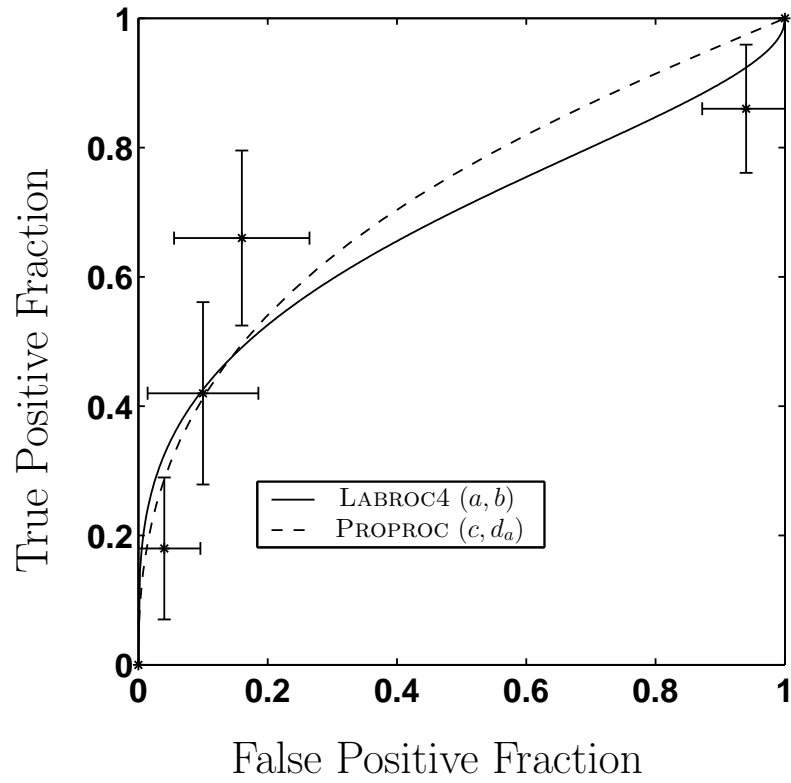


FIG. 1. Estimates of TPF and FPF obtained from simulated category data for 50 actually positive and 50 actually negative cases, plus and minus two standard errors in the means. Also shown are curves fitted to these data using LABROC4 to find ML estimates of the conventional (a, b) parameters, and PROPROC to find ML estimates of the “proper” (c, d_a) parameters.

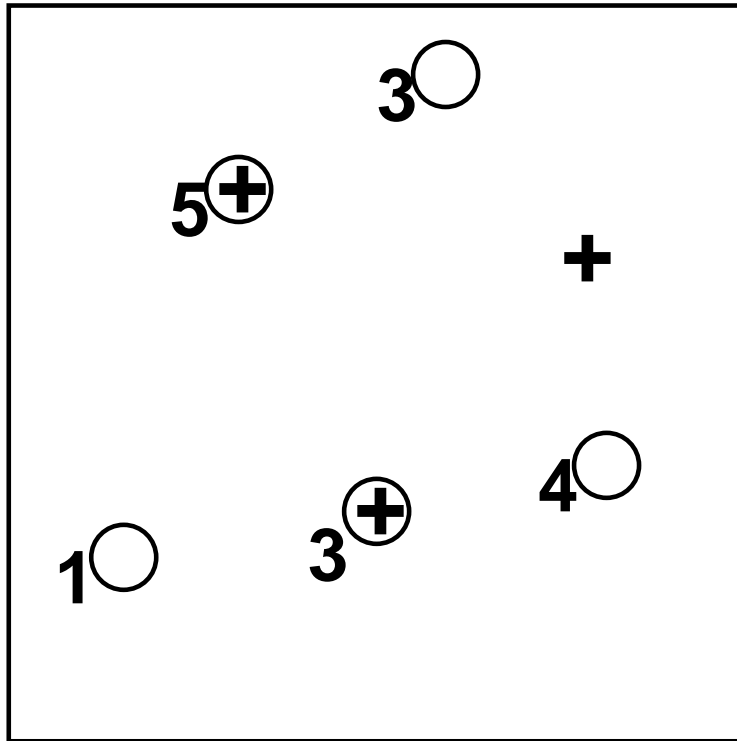


FIG. 2. A schematic image containing three signals, indicated by plus signs, and five candidate detections, indicated by circles; two of these detections are actually positive and three are actually negative [$B = 3$ and $C = 2$ in Eqs. (16) and (17)]. Next to each detection is indicated the category i reported by the observer for that detection; at a threshold of $i_0 = 2$, the observer's SDF is 0.67 and its FPP is 2.

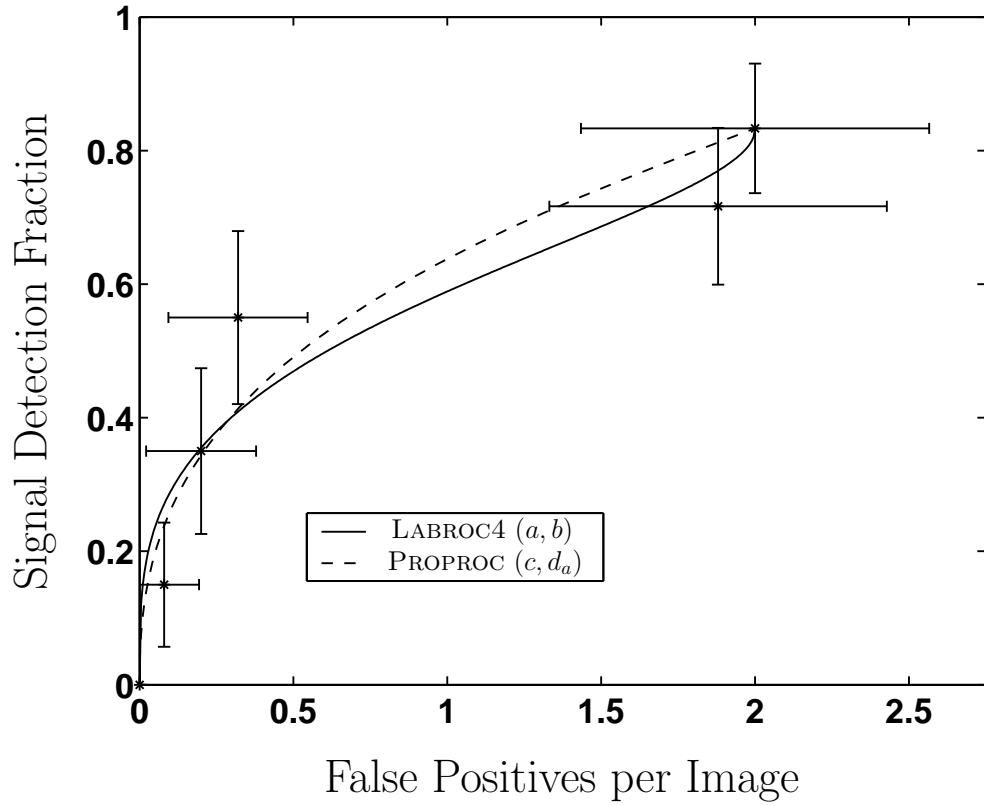


FIG. 3. Estimates of SDF and FPpI, plus and minus two standard errors, obtained from simulated category data for 50 true-positive and 50 false-positive candidate detections, out of 60 true signals and 25 images. Also shown are curves fitted by ML estimation of the analysis stage ROC curve parameters [using LABROC4 to find (a, b) , and PROPROC to find (c, d_a)], $\widehat{\text{FPpI}}_c$, and $\widehat{\text{SDF}}_c$.

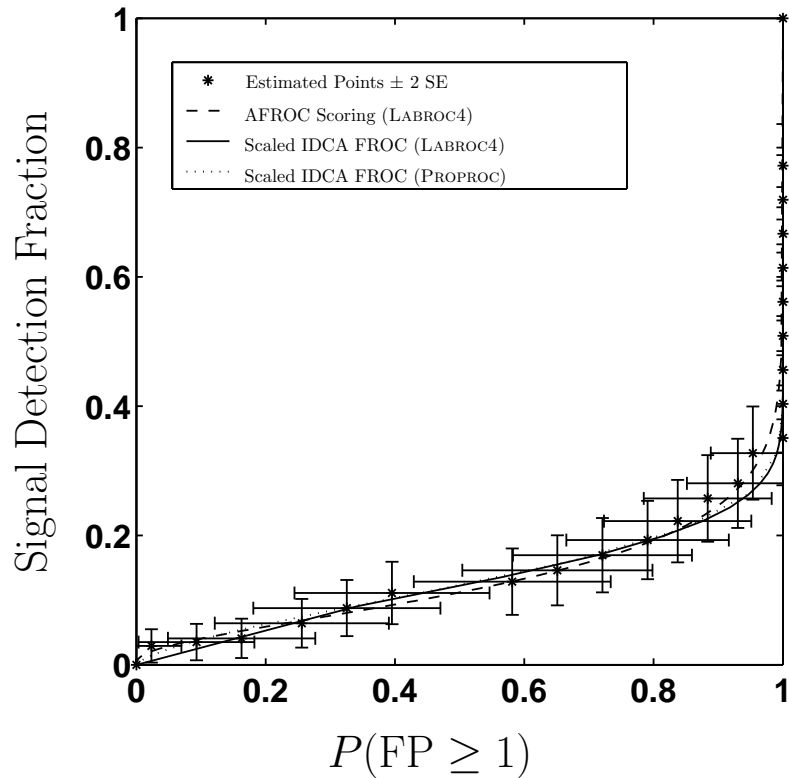


FIG. 4. Estimated AFROC operating points, plus and minus two standard errors, for dataset A, consisting of an LDA algorithm applied to detections in 47 CT volumes with 171 lung nodules. Also shown is the AFROC scoring fit to these data, as well as transformed versions of the IDCA FROC curves fitted using LABROC4 and PROPROC. Note that the LABROC4 and PROPROC curves are effectively indistinguishable here. (For clarity, operating points which were spaced too closely together to be distinguishable have been omitted.)

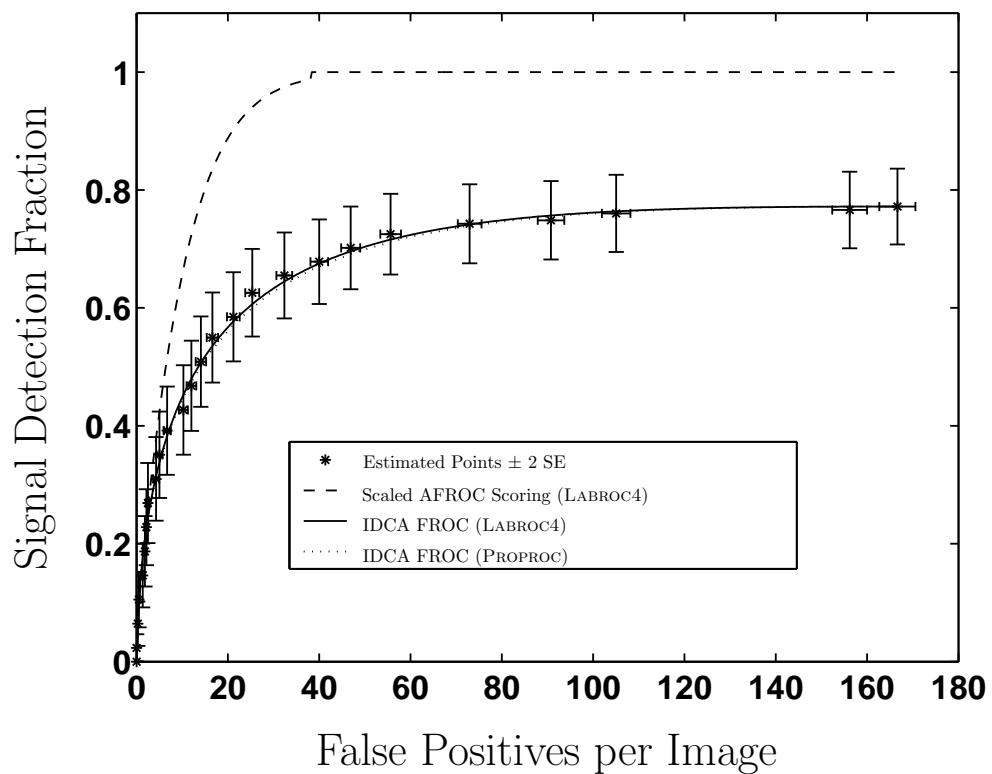


FIG. 5. Estimated FROC operating points, plus and minus two standard errors, for dataset A, consisting of an LDA algorithm applied to detections in 47 CT volumes with 171 lung nodules. Also shown are the IDCA FROC curves fitted using LABROC4 and PROPROC, as well as a transformed version of the AFROC scoring fit to these data. Note that the LABROC4 and PROPROC curves are effectively indistinguishable here. (For clarity, operating points which were spaced too closely together to be distinguishable have been omitted.)

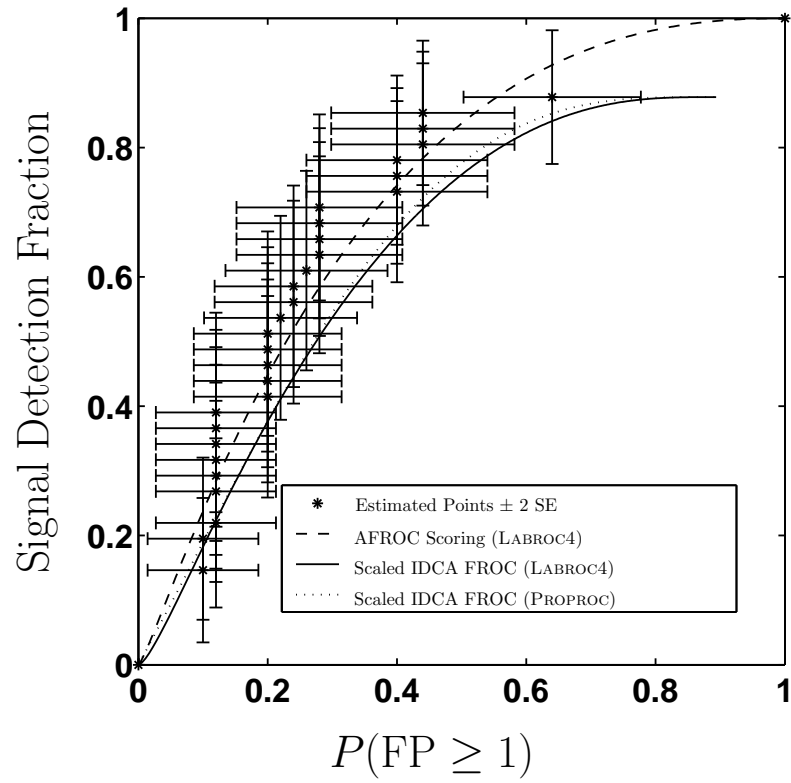


FIG. 6. Estimated AFROC operating points, plus and minus two standard errors, for dataset B, consisting of a BANN applied to detections in 50 mammograms with 41 clusters of microcalcifications. Also shown is the AFROC scoring fit to these data, as well as transformed versions of the IDCA FROC curves fitted using LABROC4 and PROPROC.

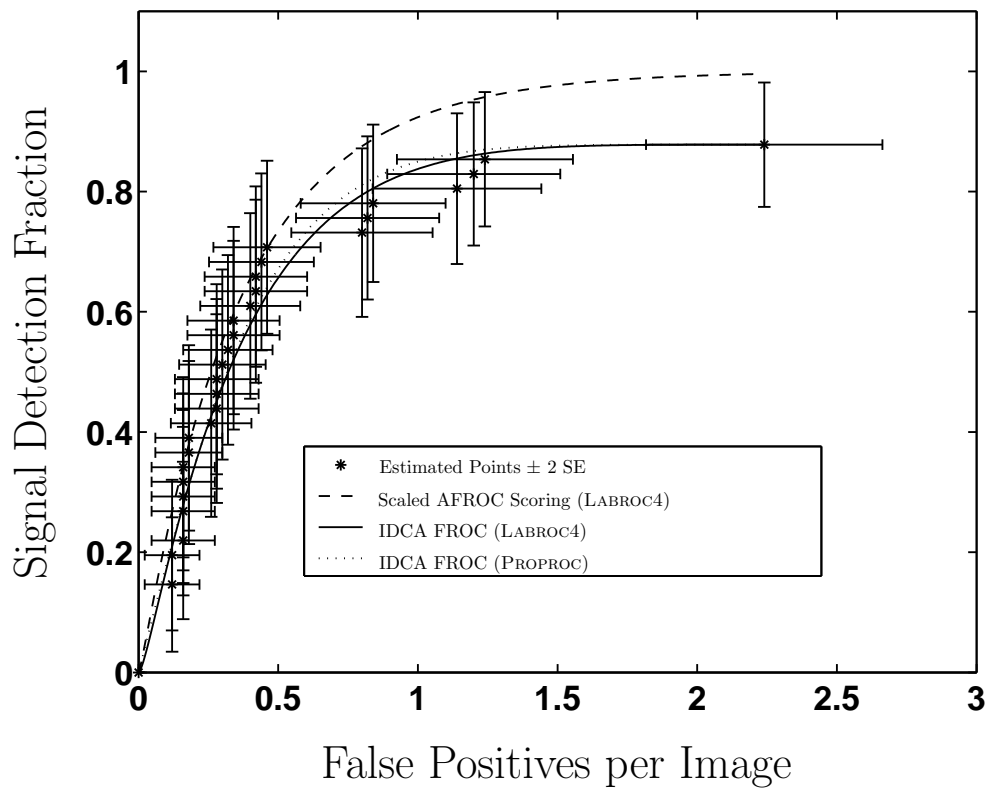


FIG. 7. Estimated FROC operating points, plus and minus two standard errors, for dataset B, consisting of a BANN applied to detections in 50 mammograms with 41 clusters of microcalcifications. Also shown are the IDCA FROC curves fitted using LABROC4 and PROPROC, as well as a transformed version of the AFROC scoring fit to these data.