On comparing methods for discriminating between actually negative and actually positive subjects with FROC type data

Tao Song, Andriy I. Bandos, and Howard E. Rockette

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania 15261

David Gur^{a)}

Department of Radiology, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15213

(Received 8 November 2007; revised 14 January 2008; accepted for publication 9 February 2008; published 24 March 2008)

The task of searching and detecting multiple abnormalities depicted on an image, or a series of images, is a common problem in different areas such as military target detection or diagnostic medical imaging. A free response receiver operating characteristic (FROC) approach for assessing performance in many of these scenarios entails marking the locations of suspected abnormalities and indicating a level of suspicion at each of the marked locations. One of the important characteristics of a system being evaluated under the FROC paradigm is its performance in the conventional ROC domain, namely classifying a subject (or a unit of interest) as “negative” or “positive” in regard to the presence of the abnormality (or any of the abnormalities) of interest. With FROC data we can compare subjects by specifying a function of multiple scores within a subject. This approach allows formulating subject-based ROC type indices that can be estimated using existing ROC concepts. In this article we focus on indices that reflect the ability of the system to discriminate between actually negative and actually positive subjects. We consider a previously proposed index that is based on the comparison of the highest scores on subjects and two new indices that are based on potentially more stable comparison functions, namely comparison of average scores and stochastic dominance. Based on these indices we develop nonparametric procedures for comparing subject-based discriminative ability of diagnostic systems being evaluated under the FROC paradigm. We also investigate the properties of the statistical procedures in a simulation study. © 2008 American Association of Physicists in Medicine. [DOI: [10.1118/1.2890410](https://doi.org/10.1118/1.2890410)]

Key words: FROC, ROC, nonparametric statistics, subject-based, discriminative ability

I. INTRODUCTION

Assessing performance under a free response approach requires searching, detecting, and classifying abnormalities within a unit of interest (i.e., a subject) when there is a possibility for none, one, or several abnormalities to be present in one subject. The importance of the free-response approach has been recognized in different fields such as acoustics,¹ radiographic signal detection,² and imagery of military targets.³ Perhaps the area in which it is most frequently used is diagnostic medical imaging.^{4–7} Analyses of data collected under the free-response paradigm are frequently termed a free-response receiver operating characteristic (FROC) analysis. A free-response approach entails detecting and marking the locations of all suspected abnormalities as well as indicating a level of suspicion (rating) regarding the specific abnormality at each marked location. Under the FROC paradigm the locations that are rated are neither fixed nor predetermined and in the FROC analysis the number of rated locations within one subject is an important characteristic of the system being evaluated.

The FROC paradigm enables assessment of the ability of a system to correctly locate the abnormalities within subjects. One method that has been used to analyze FROC data is to construct an FROC curve defined as a plot of the pro-

portion of true positive (*TP*) marks versus the average number of false positive (*FP*) marks per subject. A FROC curve is obtained by varying the “rating threshold” which is used to classify marked locations into “positive” and “negative” categories. Investigators have used both parametric^{4,5} and nonparametric⁶ approaches to fit a FROC curve and to estimate different summary indices based on the fitted FROC curve.

The consequences of missing an abnormal location and the consequences of missing an abnormal subject are fundamentally different. Sometimes it is more important to correctly identify a subject as positive or negative rather than to correctly locate the abnormal regions within a subject. It is also important to note that a superior system with respect to discrimination between actually negative and actually positive findings on a “per location” performance level is not necessarily superior with respect to discrimination between actually negative and actually positive subjects. For example if a system forms the opinion about a subject according to the maximum of the within subject ratings then its discriminative ability on a subject level will depend upon the number of marks on actually positive and actually negative subjects. If the number of marks on the actually negative subjects is sufficiently larger than the number of marks on the actually positive subjects, then the subject-level discriminative ability

FIG. 1. Histograms of 10 000 simulated *FP* and *TP* ratings generated from skewed distributions. The ratings below the 40th percentile of the distribution of the *FP* ratings were assigned to the 40th percentile and the ratings above the 60th percentile of the distribution of the *TP* ratings were assigned the value of the 60th percentile. The skewness for *FP* ratings is 0.832, and the skewness for *TP* ratings is -0.832 .

of the system may be negligible or even unreasonable (e.g., <0.5) even though the location level discriminative stability is reasonable (e.g., >0.5).

Given a set of FROC data a subject-based assessment can be achieved by combining the information on suspicious locations within subjects. However, such reduction of the FROC data is not unique and can be achieved by using a wide range of combination functions. Swenson⁸ assumed the highest-rated score on each individual subject as an observer's "first-choice" report. A figure of merit θ can then be defined as the probability that the highest score on an actually positive subject exceeds that on an actually negative subject, and this approach was considered under both nonparametric⁹ and parametric¹⁰ frameworks. Although the highest (maximum) score on the subject has long been assumed to represent the opinion of the observer regarding this subject's abnormality, no comprehensive investigation of this assumption has been performed. It seems reasonable to expect that an effective decision scheme could "form an opinion" about the entire subject based not only on the single most suspicious location but also considering all suspicious locations and their associated ratings. It is possible that depending on the pattern of location-based ratings in actually positive and actually negative subjects, a given combination function can lead to a different conclusion about the subject-based discriminative ability of a system. At the same time, the use of the "maximum" rating within a subject as a combination function may have inferior statistical properties as compared to the mean, which is the most commonly used combination function to compare two sets of multiple scores.

In this article we focus on a general framework that encompasses the maximum score approach (θ) as well as other indices. As alternatives to θ we consider two natural indices,

develop simple nonparametric procedures for statistical inferences, and compare the three indices in a simulation study. All three indices belong to a general family that includes indices for estimating the probability of a correct discrimination in a corresponding subject-based two-alternative-forced-choice (2-AFC) experiment. The considered indices quantify the ability of the system to discriminate between actually positive and actually negative subjects. Each index in the considered family is determined by a specific function for comparing the collection of scores on two different subjects (dominance function). The maximum score index, θ , corresponds to the comparison of maximum scores on two subjects, while other two indices correspond to the comparisons based on the relative order of averages, A_1 , and the stochastic dominance of the scores, A_2 .

II. METHODS

II.A. Indices of subject-based discriminative ability under the FROC paradigm

Under the FROC paradigm, the marks placed on a subject can be classified as a True Positive (*TP*) or False Positive (*FP*) at a certain threshold or "cut point" (being used to classify locations as "positive") by determining first whether the mark is located in an area where there is an actual abnormality ("correct location") and then determining whether the score ("suspicion level") is above the threshold.

The data from a FROC experiment for S_0 actually negative and S_t actually positive subjects with a fixed number of abnormalities t can be summarized as follows:

$$\{x_{s'c}^0\}_{c=1}^{n_{s'}}, \quad s' = 1, \dots, S_0 \leftrightarrow \text{"actually negative"}$$

TABLE I. Estimated expectations and standard errors of summary indices under a simulation model based on normal distributions for $\lambda=1$. The symbols in the table heading are the following simulation parameters: AUC - is the degree of separation between the distribution of FP and TP ratings; b - is the ratio of the standard deviation of FP ratings to that of TP ratings; λ - is the average number of FP marks per subject; ν - is the average proportion of the marked lesions; t - is the number of lesions on an actually positive subject. Two-sample jackknife standard error(2s-jackk.) is the average of standard errors calculated by using Eq. (10) for each dataset. Indices A_i and their expectations are defined in Eqs. (4)–(9). The estimates of the expectations and standard errors are based on 10 000 simulations. All empirical estimates of the standard error fall within the inter-quartile interval of the empirical distribution of the corresponding two-sample jackknife standard errors. For example, for the parameters $b=1$, $t=1$, $\nu=0.5$, $AUC=0.5$, the inter-quartile interval of the two-sample jackknife estimate of the standard error for $\hat{\theta}$ is (0.0875, 0.0934) and the empirical estimate of 0.0894 is well within this interval. For the above table the inter-quartile ranges of the estimates of the 2s-jackk. standard error change from 0.0051 to 0.0243.

			AUC=0.5									
			$\widehat{E(\hat{\theta})}$	Empirical $se(\hat{\theta})$	2s-jackk. $se(\hat{\theta})$	$\widehat{E(\hat{A}_1)}$	Empirical $se(\hat{A}_1)$	2s-jackk. $se(\hat{A}_1)$	$\widehat{E(\hat{A}_2)}$	Empirical $se(\hat{A}_2)$	2s-jackk. $se(\hat{A}_2)$	
1/b	t	v										
1	1	0.5	0.609	0.0894	0.0900	0.593	0.0909	0.0910	0.593	0.0881	0.0880	
		0.9	0.695	0.0842	0.0850	0.666	0.0870	0.0880	0.666	0.0832	0.0840	
	3	0.5	0.728	0.0794	0.0810	0.660	0.0874	0.0890	0.661	0.0837	0.0850	
		0.9	0.810	0.0707	0.0700	0.683	0.0901	0.0900	0.683	0.0864	0.0860	
	2	1	0.5	0.621	0.0873	0.0890	0.592	0.0888	0.0900	0.592	0.0863	0.0880
			0.9	0.718	0.0794	0.0810	0.666	0.0854	0.0870	0.666	0.0819	0.0830
3		0.5	0.764	0.0746	0.0750	0.660	0.0864	0.0870	0.660	0.0827	0.0830	
		0.9	0.871	0.0541	0.0540	0.684	0.0853	0.0860	0.684	0.0825	0.0820	
AUC=0.9												
			$\widehat{E(\hat{\theta})}$	Empirical $se(\hat{\theta})$	2s-jackk. $se(\hat{\theta})$	$\widehat{E(\hat{A}_1)}$	Empirical $se(\hat{A}_1)$	2s-jackk. $se(\hat{A}_1)$	$\widehat{E(\hat{A}_2)}$	Empirical $se(\hat{A}_2)$	2s-jackk. $se(\hat{A}_2)$	
1/b	t	v										
1	1	0.5	0.711	0.0811	0.0820	0.691	0.0824	0.0840	0.678	0.0804	0.0820	
		0.9	0.883	0.0545	0.0530	0.845	0.0628	0.0620	0.822	0.0613	0.0610	
	3	0.5	0.896	0.0524	0.0500	0.859	0.0606	0.0590	0.844	0.0597	0.0580	
		0.9	0.981	0.0188	0.0150	0.940	0.0383	0.0350	0.932	0.0376	0.0360	
	2	1	0.5	0.719	0.0810	0.0810	0.702	0.0826	0.0830	0.678	0.0817	0.0810
			0.9	0.894	0.0519	0.0500	0.864	0.0578	0.0570	0.821	0.0608	0.0600
3		0.5	0.908	0.0492	0.0470	0.880	0.0548	0.0540	0.845	0.0578	0.0570	
		0.9	0.992	0.0110	0.0080	0.964	0.0262	0.0240	0.936	0.0336	0.0320	

$$(\{x_{sc}^t\}_{c=1}^{n_s^t}, \{y_{sc}\}_{c=1}^{m_s}), \quad s = 1, \dots, S_t \leftrightarrow \text{"actually positive"}, \quad (1)$$

where s' indexes an actually negative subject and s indexes an actually positive subject. We use $n_{s'}^0$ and n_s^t to represent the number of FP marks on an actually negative and an actually positive subject respectively; m_s to represent the number of TP marks on an actually positive subject and $\vec{x}^0$ and $\vec{x}^t$ to represent the collection of ratings for the FP marks for an actually negative subject and an actually positive subject respectively. The term $\vec{y}$ is the collection of ratings for TP marks and c is used to index the collection of ratings for each individual subject. We use $x_{s'c}^0$ to represent the c th FP rating on the s' th actually negative subject, x_{sc}^t to represent the c th FP rating on the s th actually positive subject, and y_{sc} to represent the value for the c th TP rating on the actually

positive subject s . We treat the observed data as a realization of a collection of random variables and distinguish between the random quantities from their realizations with capital letters. Thus, $\vec{X}^0$, $\vec{X}^t$, and $\vec{Y}$ are the random vectors of the scores on a subject with length N^0 , N^t , and M , respectively, namely:

$$\vec{X}^0 = \begin{pmatrix} X_1^0 \\ \vdots \\ X_{N^0}^0 \end{pmatrix} \quad \vec{X}^t = \begin{pmatrix} X_1^t \\ \vdots \\ X_{N^t}^t \end{pmatrix} \quad \vec{Y}_s = \begin{pmatrix} Y_1 \\ \vdots \\ Y_M \end{pmatrix}.$$

A natural and commonly used index for summarizing the discriminative ability of a diagnostic system is the percent of correct discriminations in pairs of actually positive- actually negative subjects. The task of discrimination in such a pair of subjects is called a 2-alternative forced choice (2AFC) task. If the decision in 2AFC is guided by the comparison of a

TABLE II. Expected frequencies of subjects without any rated marks under the simulation model based on normal distributions. Expected frequencies are calculated based on binomial and Poisson distributions. The frequency of no mark on an actually positive subject is $P(\text{no TP marks}) \times P(\text{no FP marks}) = P(M=0|M \sim \text{Bin}(t, v)) \times P(N'=0|N' \sim \text{Poisson}(\lambda'))$. The frequency of no mark on an actually negative subject is $P(\text{no FP marks}) = P(N^0=0|N^0 \sim \text{Poisson}(\lambda^0))$. When the distribution of the FP and TP marks are identically distributed ($AUC=0.5, b=1$), $E(\hat{A}_1)$ and $E(\hat{A}_2)$ can be calculated based on the parametric assumptions using Eqs. (5) and (6). For a pair of an actually negative and an actually positive subject that both are marked: $P(\text{mean}(\{Y, X'\}) > \text{mean}(X^0))=0.5$ and $P(w(X^0, \{Y, X'\}) > 0.5)=0.5$.

		$\lambda=0.5$		$\lambda=1.0$			$\lambda=2.0$	
		Actually positive subject	Actually negative subject	Actually positive subject	Actually negative subject	$E(\hat{A}_1)$ and $E(\hat{A}_2)$ under $AUC=0.5, b=1$	Actually positive subject	Actually negative subject
1	0.5	0.30	0.61	0.18	0.37	0.595	0.07	0.14
	0.9	0.06	0.61	0.04	0.37	0.665	0.01	0.14
3	0.5	0.08	0.61	0.05	0.37	0.660	0.02	0.14
	0.9	0.00	0.61	0.00	0.37	0.685	0.00	0.14

certain subject-specific scalar (ordinal rating) then the percent correct in 2AFC is equivalent to the area under the ROC curve constructed on these ratings.^{11,12} However, not every 2AFC task can be described by an underlying ordinal rating and hence the percent correct in 2AFC is a more general index than the area under the ROC curve. In order to compute the percent correct in a 2AFC experiment we need to know the value of a corresponding comparison function ψ for every pair of actually positive-actually negative subjects. For the 2AFC task that is guided by comparing subject-specific ratings the ψ function can be defined as a result of comparison of two numbers, namely:

$$\psi(b, c) = \begin{cases} 1 & b < c \\ 1/2 & b = c \\ 0 & b > c \end{cases}.$$

With FROC data the 2-AFC task is complicated by the need to compare vectors of observations $\vec{b}$ and $\vec{c}$. For this purpose we define a generalization of the ψ function, i.e.:

$$\tilde{\psi}(\vec{b}, \vec{c}) = \begin{cases} 1 & \vec{b} \triangleleft \vec{c} \\ 1/2 & \vec{b} \triangleleft \triangleright \vec{c}, \\ 0 & \vec{b} \triangleright \vec{c} \end{cases}, \quad (2)$$

where $\vec{b} \triangleleft \vec{c}$ indicates that the collection of scores in $\vec{c}$ dominates the collection of scores in $\vec{b}$ based on a pre-selected rule ($\triangleleft$); and $\vec{b} \triangleleft \triangleright \vec{c}$ corresponds to the subject where the scores in $\vec{b}$ and $\vec{c}$ are equivalent according to a pre-selected rule ($\triangleleft \triangleright$).

For the problem considered in this paper we use $\vec{b} = \vec{x}^0$ and $\vec{c} = \{\vec{x}^t, \vec{y}\}$, i.e., we compare scores on an actually negative subject to the scores on an actually positive subject. In a FROC experiment there might be subjects with no marks, hence subjects without any scores. This leads to the possibility of having empty vectors $\vec{b}$ and $\vec{c}$. Therefore we augment the definition of $\tilde{\psi}$ in Eq. (2) by adopting the approach pro-

posed by Chakraborty.¹⁰ For a pair of an actually positive and an actually negative subject, if only the actually negative ($\vec{x}^0 = \emptyset$ and $\{\vec{x}^t, \vec{y}\} \neq \emptyset$) or only the actually positive subject ($\vec{x}^0 \neq \emptyset$ and $\{\vec{x}^t, \vec{y}\} = \emptyset$) is not marked, we assign $\tilde{\psi}$ to be 1 or 0, respectively; if both subjects are not marked ($\vec{x}^0 = \emptyset$ and $\{\vec{x}^t, \vec{y}\} = \emptyset$), we assign $\tilde{\psi}$ to be 0.5. Specifically, we define $\tilde{\psi}$ as follows:

$$\tilde{\psi}(\vec{b}, \vec{c}) = \begin{cases} 1 & \vec{b} \triangleleft \vec{c}, & \text{or } \vec{b} = \emptyset \text{ and } \vec{c} \neq \emptyset \\ 1/2 & \vec{b} \triangleleft \triangleright \vec{c}, & \text{or } \vec{b} = \vec{c} = \emptyset \\ 0 & \vec{b} \triangleright \vec{c}, & \text{or } \vec{b} \neq \emptyset \text{ and } \vec{c} = \emptyset \end{cases}.$$

In this article we consider three specific indices from a family that quantifies the probability of correct discrimination in a 2-AFC task. Each of the indices in this family can be written as follows:

$$A = E[\tilde{\psi}(\{\vec{X}^0\}, \{\vec{X}^t, \vec{Y}\})]. \quad (3)$$

Let us first focus on a pair of an actually negative and an actually positive subject on which at least one subject has no marks. If only the actually negative subject is not marked, it contributes a value of $1 \times P(\vec{X}^0 = \emptyset)[1 - P(\{\vec{X}^t, \vec{Y}\} = \emptyset)]$ to the expectation; if only the actually positive subject is not marked, it contributes a value of zero to the expectation; if neither of them is marked, it contributes a value of $1/2 \times P(\vec{X}^0 = \emptyset)P(\{\vec{X}^t, \vec{Y}\} = \emptyset)$ to the expectation. All pairs of subjects with at least one subject without any marks contribute $P(\vec{X}^0 = \emptyset) \times [1 - 1/2 P(\vec{X}^t = \emptyset, \vec{Y} = \emptyset)]$ to the expectation in Eq. (3) for each of the indices in the family.

Using the maximum of the components of the vector is a special case of Eq. (3), and we can represent the “figure of merit”¹⁰ θ in the following manner:

$$\theta = A_0 = E[\psi(\max\{\vec{X}^0\}, \max\{\vec{X}^t, \vec{Y}\}) \times I(\{\vec{X}^0\} \neq \emptyset, \{\vec{X}^t, \vec{Y}\} \neq \emptyset)] + P(\vec{X}^0 = \emptyset) \times [1 - \frac{1}{2} P(\vec{X}^t = \emptyset, \vec{Y} = \emptyset)]. \quad (4)$$

The maximum score is only one of multiple possible functions that can be used with this approach. One commonly used summary statistic is a sample mean and our first index (A_1) is based on the comparison of averages of the sets of scores in a pair of subjects, namely:

$$A_1 = E \left[\psi \left(\frac{\sum_{c_1=1}^{N^0} X_{c_1}^0}{N^0}, \frac{\sum_{c_2=1}^{N^t} X_{c_2}^t + \sum_{c_3=1}^M Y_{c_3}}{N^t + M} \right) \times I(\{\vec{X}^0\} \neq \emptyset, \{\vec{X}^t, \vec{Y}\} \neq \emptyset) + P(\vec{X}^0 = \emptyset) \times \left[1 - \frac{1}{2} P(\vec{X}^t = \emptyset, \vec{Y} = \emptyset) \right] \right]. \quad (5)$$

Another commonly using index is a Wilcoxon statistic and our second index (A_2) is based on the stochastic dominance of the set of scores in one of the subjects, namely:

$$A_2 = E \{ \psi[0.5, w(\vec{X}^0, \{\vec{X}^t, \vec{Y}\})] \times I(\{\vec{X}^0\} \neq \emptyset, \{\vec{X}^t, \vec{Y}\} \neq \emptyset) + P(\vec{X}^0 = \emptyset) \times \left[1 - \frac{1}{2} P(\vec{X}^t = \emptyset, \vec{Y} = \emptyset) \right] \}, \quad (6)$$

where

$$\hat{\theta} = \hat{A}_0 = \frac{\sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} \tilde{\psi}_{\max}(\{x_{s',c}^0\}_{c=1}^{n_{s'}^0}, \{\{x_{sc}^t\}_{c=1}^{n_s^t}, \{y_{sc}^m\}_{c=1}^{m_s}\})}{S_0 \times S_t}$$

or

$$\hat{\theta} = \hat{A}_0 = \frac{\sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} \psi(\max(\{x_{s',c}^0\}_{c=1}^{n_{s'}^0}, \max(\{x_{sc}^t\}_{c=1}^{n_s^t}, \{y_{sc}^m\}_{c=1}^{m_s})) \times I(n_{s'}^0, (n_s^t + m_s) \neq 0)}{S_0 \times S_t} + \left[\frac{\sum_{s'=1}^{S_0} I(n_{s'}^0 = 0)}{S_0} \right] \times \left[1 - \frac{1}{2} \times \frac{\sum_{s=1}^{S_t} I(n_s^t + m_s = 0)}{S_t} \right] \quad (7)$$

$$\hat{A}_1 = \frac{\sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} \tilde{\psi}_{\text{mean}}(\{x_{s',c}^0\}_{c=1}^{n_{s'}^0}, \{\{x_{sc}^t\}_{c=1}^{n_s^t}, \{y_{sc}^m\}_{c=1}^{m_s}\})}{S_0 \times S_t}$$

or

$$\hat{A}_1 = \frac{\sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} \psi \left(\frac{\sum_{c_1=1}^{n_{s'}^0} x_{c_1}^0}{n_{s'}^0}, \frac{\sum_{c_2=1}^{n_s^t} x_{c_2}^t + \sum_{c_3=1}^{m_s} y_{c_3}^m}{n_s^t + m_s} \right) \times I(n_{s'}^0, (n_s^t + m_s) \neq 0)}{S_0 \times S_t} + \left[\frac{\sum_{s'=1}^{S_0} I(n_{s'}^0 = 0)}{S_0} \right] \times \left[1 - \frac{1}{2} \times \frac{\sum_{s=1}^{S_t} I(n_s^t + m_s = 0)}{S_t} \right] \quad (8)$$

$$w(\vec{b}, \vec{c}) = \frac{\sum_{i=1}^m \sum_{j=1}^n \psi(b_i, c_j)}{m \times n}, \quad \text{if } \vec{b} = (b_1, \dots, b_m) \quad \text{and} \quad \vec{c} = (c_1, \dots, c_n).$$

The three indices formulated above have the interpretation of the percent of correct decisions in a 2AFC task where the decisions are determined according to a chosen comparison function (based on maximum, average or stochastic dominance). The two indices based on the maximum of ratings (A_0) and based on the average of ratings (A_1) are equivalent to the areas under the ROC curves constructed based on the maximum and average of the within-subject ratings correspondingly (with an artificial lowest rating assigned to the subjects with no marks). This, however, is not true in general for the index A_2 which is based on the “stochastic dominance” of the ratings on one of the subjects, because the stochastic dominance relation does not in general possess the transitivity property inherent to standard order relations.

II.B. Statistical inferences

The nonparametric estimators of the summary statistics considered in the previous section can be applied to data given in the format of Eq. (1):

$$\widehat{A}_2 = \frac{\sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} \tilde{\psi}_{wilcoxon}(\{x_{s',c}^0\}_{c=1}^{n_{s'}^0}, \{\{x_{sc}^t\}_{c=1}^{n_s^t}, \{y_{sc}\}_{c=1}^{m_s}\})}{S_0 \times S_t}$$

or

$$\begin{aligned} \widehat{A}_2 = & \frac{\sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} \psi \left\{ 0.5, \frac{\sum_{c_1=1}^{n_{s'}^0} \sum_{c_2=1}^{n_s^t} \sum_{c_3=1}^{m_s} [\psi(x_{s',c_1}^0, x_{s,c_2}^t) + \psi(x_{s',c_1}^0, y_{s,c_3})]}{n_{s'}^0 \times (n_s^t + m_s)} \right\}}{S_0 \times S_t} \times I(n_{s'}^0, (n_s^t + m_s) \neq 0) + \left[\frac{\sum_{s'=1}^{S_0} I(n_{s'}^0 = 0)}{S_0} \right] \\ & \times \left[1 - \frac{1}{2} \times \frac{\sum_{s=1}^{S_t} I(n_s^t + m_s = 0)}{S_t} \right]. \end{aligned} \quad (9)$$

The estimators in Eqs. (7)–(9) have a structure similar to a two-sample U statistic. This structure permits the development of a closed form expression for a jackknife or bootstrap estimate of the variance when these re-sampling techniques consider subject as a sampling unit. In this article we propose to use a two-sample jackknife variance.¹³ The two-sample jackknife variance when applied to the area under the empirical ROC curve is known to be equivalent to the variance proposed by DeLong *et al.*¹⁴ Hence, we present a computational algorithm in a manner similar to the conventional description of DeLong *et al.*'s variance estimator.

Specifically, we:

1. Find the simple averages of each row and column in the matrix $\{\tilde{\psi}_{s',s}\}$ where $\tilde{\psi}_{s',s} = \tilde{\psi}(\{\tilde{x}_{s'}^0\}, \{\tilde{x}_s^t, \tilde{y}_s\})$. Namely, for the s' th row (corresponding to the s' th actually negative subject): $\overline{\tilde{\psi}_{s',\bullet}} = \sum_{s=1}^{S_t} \tilde{\psi}_{s',s} / (S_t)$ and for the s th column (corresponding to the s th actually positive subject): $\overline{\tilde{\psi}_{\bullet,s}} = \sum_{s'=1}^{S_0} \tilde{\psi}_{s',s} / (S_0)$.

2. Compute the unbiased estimates of the variance of the averages of the rows and columns or the variance elements that result from actually negative and actually positive subjects correspondingly.

Namely, for the rows (due to actually negative subjects): $\widehat{V}_N = \sum_{s'=1}^{S_0} (\overline{\tilde{\psi}_{s',\bullet}} - \overline{\tilde{\psi}_{\bullet,\bullet}})^2 / (S_0 - 1)$ and for the columns (due to actually positive subjects): $\widehat{V}_A = \sum_{s=1}^{S_t} (\overline{\tilde{\psi}_{\bullet,s}} - \overline{\tilde{\psi}_{\bullet,\bullet}})^2 / (S_t - 1)$

3. Compute the variance using: $\widehat{V}(\widehat{A}) = \widehat{V}_N / S_0 + \widehat{V}_A / S_t$ or

$$\widehat{V}(\widehat{A}) = \frac{\sum_{s'=1}^{S_0} (\overline{\tilde{\psi}_{s',\bullet}} - \overline{\tilde{\psi}_{\bullet,\bullet}})^2}{S_0 \times (S_0 - 1)} + \frac{\sum_{s=1}^{S_t} (\overline{\tilde{\psi}_{\bullet,s}} - \overline{\tilde{\psi}_{\bullet,\bullet}})^2}{S_t \times (S_t - 1)}. \quad (10)$$

When comparing two diagnostic systems evaluated under the FROC paradigm with the proposed indices we use the difference between the modality-specific indices. When assessing statistical significance of the differences in the indi-

ces observed for the two modalities, we propose an asymptotic procedure with the following test statistic:

$$Z_i = \frac{\hat{A}_i^2 - \hat{A}_i^1}{\sqrt{\hat{V}(\hat{A}_i^2 - \hat{A}_i^1)}}, \quad i = 0, 1, 2.$$

To conduct a statistical test we compare the Z statistic described above with the pre-specified percentile of the standard normal distribution. In an unpaired design where different subjects are evaluated by different systems the estimator of the variance of the difference is simply the sum of the corresponding variance estimators. In a paired design, where the same set of subjects is evaluated under both modalities, the estimator of the variance of the difference can be computed using equation (10) where $\tilde{\psi}$ is replaced with $\tilde{\psi}^1 - \tilde{\psi}^2$.

III. RESULTS

III.A. Simulation description

We evaluated all three indices under different scenarios in each of which we generate 10 000 independent datasets based on a set of pre-determined parameters. Each dataset consisted of 20 actually positive and 20 actually negative subjects. The notations for the observations that are typically obtained in a FROC experiment are summarized in expression (1). In this simulation study we consider the scenario where the sample consists of a group of actually negative subjects with zero known abnormalities and a group of actually positive subjects with t abnormalities. Originally proposed by Bunch *et al.*² and further employed by other researchers,^{4,5,10,15} the number of FP marks N on a subject is usually treated as a Poisson variable with parameter λ . The parameter λ can be viewed as the mean number of FP marks on a subject and a smaller value is expected for an experienced observer.^{9,10,15} The number of TP marks on a subject is typically modeled by the binomial distribution. The trial size is equal to the total number of lesions t on a subject. The

success rate ν regulates the proportion of lesions that are actually detected namely, marked at the right locations.

For all subjects, regardless of their actual scores, the number of FP marks, n , was generated from a Poisson distribution with expectation λ of 0.5, 1.0, and 2.0. For every actually positive subject the number of TP marks, m , was generated from a binomial distribution with number of trials t of 1 and 3 and probability of success ν of 0.5, 0.7 and 0.9.¹⁵ The ratings for FP , $\tilde{x}$, and TP , $\tilde{y}$, marks were generated independently from normal distributions with variances and means chosen to achieve a pre-specified separation between FP and TP ratings corresponding to AUC of 0.5, 0.7 and 0.9 and we allow $1/b = \sigma_Y/\sigma_X$ to be either 1 or 2, the latter representing the case where the ratings for the abnormalities are more variable than the ratings for normal regions as is often the case in ROC studies.¹⁶ In the evaluation of all three indices we used a range of parameters that include values that we have observed in breast imaging studies.

We also evaluated the three indices when ratings follow a pair of skewed distributions with a non-zero mass at the extremes. These distributions were created by grouping normal distributions. The ratings below the 40th percentile of the distribution of the FP ratings were assigned to the 40th percentile and the ratings above the 60th percentile of the distribution of the TP ratings were assigned the value of the 60th percentile. As a result, the distribution of the FP ratings becomes right skewed (e.g., the skewness is 0.832 for $AUC=0.8$ and $b=1$) and the distribution of the TP ratings becomes left-skewed (e.g., the skewness is -0.832 for $AUC=0.8$ and $b=1$) (Fig. 1).

The simulation model described above is similar but slightly different from the search model presented by Chakraborty.¹⁰ In our model we allow for more flexibility by permitting the variance of the ratings to be different for FP and TP marks. We also evaluate the different methods under a non-normal distribution. Because of the large number of simulation scenarios we considered, only a fraction of these are included in the tables.

III.B. Estimates of the expectation and variance of the three indices

Table I summarizes the estimated expectations and standard errors of the three estimators when data are generated from the simulation model based on a binormal distribution and $\lambda=1.0$. The standard errors were estimated both with a sample variance divided by the number of simulated realizations and by using the two-sample jackknife variance described in Sec. II A. For each of the three indices and for all scenarios that we considered, the empirical standard error is covered by to the inter-quartile interval of the empirical distribution of the two-sample jackknife estimate of the standard error. The standard error of $\hat{\theta}$ ("maximum score") tends to be lower than the standard error of $\hat{A}_1$ or $\hat{A}_2$, with the difference increasing as the expected number of identified lesions increases or as AUC increases.

In the adopted simulation model, when the distributions of the ratings of the actually negative and actually positive

subjects (the latter have a mixture of distribution of ratings for FP and TP marks) have the same location (corresponding to $AUC=0.5$), the estimated expectations range from about 0.59 to 0.87, with A_1 and A_2 being below 0.7 for most of the simulated scenarios. The phenomenon that the expectation of the indices is substantially greater than 0.5 when $AUC=0.5$ can be partially attributed to an imbalance in frequencies of actually positive and actually negative subjects without any marks. Specifically, the substantially larger frequency of actually positive subjects without any marks as compared with the actually negative subjects (Table II) poses an imbalanced frequency of $\tilde{\psi}=1$ (high) and $\tilde{\psi}=0$ (low), and hence, shifts the expectation of the comparison function $\tilde{\psi}$ towards higher values. In fact, under the simulation model in which the ratings of the FP and TP marks follow the same normal distributions ($AUC=0.5, b=1$), the expectations for the indices A_1 and A_2 can be computed directly using formulas (5) and (6) and expected frequencies of subjects without any marks (Table II). The expectation of θ can also be computed using the approach provided by Chakraborty.¹⁰ Thus, unlike the index of the average performance in a conventional ROC experiment, the expectations of the considered indices are not necessarily 0.5 when the ratings of actually negative and actually positive subjects have the same location ($AUC=0.5$).

A comparison of the estimated expectations of the three indices suggests that $E(\hat{\theta})$ is always higher than $E(\hat{A}_1)$ and $E(\hat{A}_2)$. In fact, under our simulation scenario in which the FP and TP rating follow the same normal distribution ($AUC=0.5, b=1$) the expectation of $\hat{\theta}$, unlike that of $\hat{A}_1$ and $\hat{A}_2$, would be higher than 0.5 even in a subpopulation of subjects with at least one mark. The reason for this phenomenon is that the "maximum rating," hence, $\hat{\theta}$, is substantially affected not only by the actual value of ratings but also by the number of rated marks, and in our simulation model the actually positive subjects have on average a higher total number of marks (FP and TP) than the total number of marks on actually negative subjects (FP only). This property also partially contributes to the fact that the expectation of $\hat{\theta}$ is greater than that of $\hat{A}_1$ and $\hat{A}_2$ when distributions of the FP and TP ratings are separated ($AUC(FP, TP) > 0.5$).

Table II shows the frequencies of the subjects for which there are no marks. From the formulation of the indices in Eqs. (4)–(6) one can see that a pair of an actually positive and an actually negative subject in which at least one subject has no marks affects all three considered indices in the same manner. Hence, an increasing frequency of subjects without any marks can be expected to make the differences between the indices less profound and thus to attenuate the differences in the power of the corresponding statistical tests.

III.C. Type I error rate and statistical power for testing differences between the two systems

For both normal and skewed distributions with nonzero mass at the extremes (Table III), the estimated type I error rate is close to the nominal value. However, when both the

TABLE III. Estimated type I error rates for testing the equality of the indices under a simulation model based on normal distributions and skewed distribution (in parentheses). The symbols in the table heading are the following simulation parameters: AUC - is the degree of separation between the distribution of FP and TP ratings; b - is the ratio of the standard deviation of FP ratings to that of TP ratings; λ - is the average number of FP marks per subject; ν - is the average proportion of the marked lesions t - is the number of lesions on an actually positive subject. The estimates of the type I error rate are based on 10 000 simulations which results in a maximum possible estimation standard error of 0.005.

			AUC=0.5			AUC=0.7			AUC=0.9			
			$\lambda=1.0$			$\lambda=1.0$			$\lambda=1.0$			
			θ	A_1	A_2	θ	A_1	A_2	θ	A_1	A_2	
1/b	t	v										
1	1	0.5	0.055	0.057	0.055	0.056	0.052	0.053	0.051	0.050	0.051	
			(0.056)	(0.058)	(0.058)	(0.057)	(0.053)	(0.054)	(0.051)	(0.051)	(0.050)	
		0.9	0.048	0.052	0.053	0.051	0.053	0.052	0.041	0.046	0.050	
			(0.048)	(0.053)	(0.052)	(0.052)	(0.055)	(0.055)	(0.042)	(0.047)	(0.048)	
	3	0.5	0.055	0.056	0.054	0.053	0.053	0.051	0.042	0.046	0.049	
			(0.057)	(0.056)	(0.057)	(0.054)	(0.051)	(0.052)	(0.044)	(0.047)	(0.048)	
		0.9	0.046	0.054	0.055	0.037	0.051	0.050	0.002	0.025	0.031	
			(0.053)	(0.055)	(0.055)	(0.048)	(0.054)	(0.053)	(0.004)	(0.027)	(0.030)	
	2	1	0.5	0.054	0.055	0.058	0.054	0.053	0.054	0.053	0.054	0.053
(0.057)				(0.058)	(0.059)	(0.054)	(0.054)	(0.054)	(0.054)	(0.051)	(0.053)	
0.9			0.049	0.048	0.050	0.046	0.051	0.050	0.038	0.045	0.047	
			(0.050)	(0.047)	(0.050)	(0.047)	(0.049)	(0.051)	(0.039)	(0.042)	(0.046)	
3		0.5	0.049	0.054	0.053	0.046	0.050	0.049	0.039	0.044	0.047	
			(0.051)	(0.056)	(0.056)	(0.048)	(0.052)	(0.050)	(0.039)	(0.045)	(0.046)	
		0.9	0.035	0.045	0.046	0.019	0.049	0.050	0.000	0.011	0.027	
			(0.049)	(0.049)	(0.050)	(0.031)	(0.052)	(0.050)	(0.000)	(0.009)	(0.025)	
AUC=0.8												
			$\lambda=0.5$			$\lambda=1.0$			$\lambda=2.0$			
			θ	A_1	A_2	θ	A_1	A_2	θ	A_1	A_2	
1/b		t	v									
1	1	0.5	0.056	0.055	0.054	0.052	0.053	0.054	0.051	0.055	0.054	
			(0.056)	(0.055)	(0.056)	(0.052)	(0.054)	(0.055)	(0.052)	(0.055)	(0.056)	
		0.9	0.046	0.046	0.045	0.046	0.049	0.053	0.049	0.046	0.048	
			(0.046)	(0.047)	(0.047)	(0.048)	(0.052)	(0.050)	(0.050)	(0.052)	(0.049)	
	3	0.5	0.047	0.049	0.051	0.045	0.050	0.048	0.051	0.056	0.057	
			(0.049)	(0.052)	(0.051)	(0.048)	(0.051)	(0.048)	(0.054)	(0.056)	(0.056)	
		0.9	0.005	0.028	0.033	0.018	0.042	0.045	0.035	0.047	0.047	
			(0.012)	(0.033)	(0.032)	(0.036)	(0.045)	(0.048)	(0.051)	(0.048)	(0.049)	
	2	1	0.5	0.054	0.053	0.053	0.054	0.055	0.056	0.052	0.054	0.055
(0.061)				(0.061)	(0.061)	(0.051)	(0.051)	(0.051)	(0.052)	(0.055)	(0.054)	
0.9			0.041	0.043	0.046	0.041	0.047	0.048	0.046	0.051	0.052	
			(0.040)	(0.043)	(0.045)	(0.043)	(0.045)	(0.047)	(0.047)	(0.052)	(0.052)	
3		0.5	0.041	0.046	0.046	0.047	0.045	0.047	0.047	0.052	0.054	
			(0.039)	(0.041)	(0.040)	(0.046)	(0.048)	(0.048)	(0.049)	(0.052)	(0.050)	
		0.9	0.001	0.021	0.033	0.003	0.032	0.036	0.014	0.042	0.048	
			(0.001)	(0.016)	(0.027)	(0.006)	(0.036)	(0.043)	(0.017)	(0.046)	(0.044)	

number of abnormalities on the actually positive subjects and the separation between the FP and TP ratings (as measured by AUC) are extremely large the estimated type I error rate is low for all three indices, and the procedure based on the maximum score index, θ , demonstrates the greatest degree of conservativeness.

The complex structure of FROC data results in multiple ways in which two diagnostic systems may differ, and several of these differences affect the ability of a correct discrimination between actually positive and actually negative subjects. In this article we focus on two types of such differences. First, we consider two diagnostic systems which are

TABLE IV. Estimated power for detecting differences resulting from differing separation between *FP* and *TP* ratings (*AUC*) for $\lambda=1.0$. The symbols in the table heading are the following simulation parameters: *AUC* - is the degree of separation between the distribution of *FP* and *TP* ratings; *b* - is the ratio of the standard deviation of *FP* ratings to that of *TP* ratings; λ - is the average number of *FP* marks per subject; ν - is the average proportion of the marked lesions *t* - is the number of lesions on an actually positive subject. The estimates of the statistical power are based on 10 000 simulations which results in a maximum possible estimation standard error of 0.005.

			Under normal distributions								
			$AUC^*=0.7$ versus $AUC^*=0.5$			$AUC^*=0.9$ versus $AUC^*=0.5$			$AUC^*=0.9$ versus $AUC^*=0.7$		
			θ	A_1	A_2	θ	A_1	A_2	θ	A_1	A_2
$1/b$	t	ν									
1	1	0.5	0.071	0.072	0.071	0.140	0.133	0.120	0.083	0.075	0.069
		0.9	0.115	0.114	0.109	0.460	0.374	0.315	0.200	0.148	0.121
	3	0.5	0.124	0.139	0.132	0.416	0.450	0.420	0.162	0.168	0.156
		0.9	0.185	0.217	0.212	0.698	0.762	0.762	0.195	0.308	0.324
2	1	0.5	0.068	0.076	0.072	0.135	0.157	0.121	0.078	0.080	0.071
		0.9	0.113	0.132	0.109	0.456	0.485	0.334	0.183	0.184	0.118
	3	0.5	0.124	0.179	0.145	0.353	0.553	0.426	0.128	0.198	0.158
		0.9	0.188	0.320	0.249	0.631	0.922	0.843	0.088	0.433	0.370
Under skewed distributions											
			$AUC^*=0.7$ versus $AUC^*=0.5$			$AUC^*=0.9$ versus $AUC^*=0.5$			$AUC^*=0.9$ versus $AUC^*=0.7$		
			θ	A_1	A_2	θ	A_1	A_2	θ	A_1	A_2
$1/b$	t	ν									
1	1	0.5	0.071	0.072	0.072	0.150	0.143	0.122	0.091	0.081	0.074
		0.9	0.121	0.117	0.113	0.506	0.410	0.339	0.219	0.172	0.134
	3	0.5	0.136	0.138	0.130	0.490	0.464	0.420	0.191	0.184	0.160
		0.9	0.227	0.187	0.192	0.818	0.720	0.726	0.305	0.305	0.312
2	1	0.5	0.076	0.075	0.074	0.156	0.158	0.125	0.081	0.082	0.071
		0.9	0.145	0.131	0.113	0.538	0.496	0.343	0.194	0.192	0.123
	3	0.5	0.188	0.170	0.147	0.489	0.530	0.406	0.144	0.195	0.144
		0.9	0.398	0.268	0.230	0.890	0.858	0.771	0.140	0.356	0.310

equal with respect to all the parameters except in regard to the discrimination between *FP* and *TP* ratings (degree of separation between *FP* and *TP* ratings, parameter *AUC*). Second, we consider two diagnostic systems that differ only with respect to the average number of the *FP* marks on a subject (λ).

Table IV shows the estimates of the statistical power for the scenario where the two diagnostic systems differ in regard to separation between the *FP* and *TP* ratings. From this table one can observe that for samples of subjects in which all actually positive subjects have a single abnormality ($t=1$), and where *FP* and *TP* ratings follow normal distributions with equal variance ($b=1$), in most of the scenarios the statistical test based on θ has the greatest statistical power to detect the differences between the two systems. However, in all other instances where *FP* and *TP* ratings follow normal distributions, the mean score index, *A*₁, results in a more powerful statistical test. In the scenarios with skewed distributions

with non-zero mass at extremes the statistical power of the test based on θ is greater than that for the other two indices in most instances.

Tables V and VI demonstrate the estimates of the statistical power for the scenario when the two diagnostic systems differ only with respect to the average number of *FP* marks on a subject. In this scenario the index based on stochastic dominance of the scores, *A*₂, results in higher statistical power for all considered scenarios. For the considered combinations of parameters and a sample size of 20 actually positive and 20 actually negative subjects the estimates of statistical power are quite low. To verify whether the patterns remain the same for higher sample sizes we additionally considered a sample size of 100 actually positive and 100 actually negative subjects (Table VI).

IV. DISCUSSION

In this article we have focused on methods of analysis that evaluate the ability to discriminate between actually negative

TABLE V. Estimated power for detecting differences resulting from differing average numbers of *FP* marks (λ) for $AUC=0.8$. The symbols in the table heading are the following simulation parameters: AUC - is the degree of separation between the distribution of *FP* and *TP* ratings; b - is the ratio of the standard deviation of *FP* ratings to that of *TP* ratings; λ - is the average number of *FP* marks per subject; ν - is the average proportion of the marked lesions; t - is the number of lesions on an actually positive subject. The estimates of the statistical power are based on 10 000 simulations which results in a maximum possible estimation standard error of 0.005.

			Under normal distributions								
			$\lambda=1.0$ versus $\lambda=0.5$			$\lambda=2.0$ versus $\lambda=0.5$			$\lambda=2.0$ versus $\lambda=1.0$		
			θ	A_1	A_2	θ	A_1	A_2	θ	A_1	A_2
$1/b$	t	ν									
1	1	0.5	0.056	0.063	0.069	0.081	0.112	0.127	0.061	0.069	0.074
		0.9	0.081	0.109	0.130	0.222	0.334	0.415	0.095	0.129	0.146
	3	0.5	0.067	0.092	0.103	0.152	0.248	0.291	0.081	0.098	0.112
		0.9	0.028	0.093	0.113	0.147	0.264	0.309	0.065	0.104	0.112
2	1	0.5	0.055	0.064	0.069	0.071	0.097	0.132	0.056	0.068	0.077
		0.9	0.069	0.093	0.133	0.131	0.264	0.414	0.066	0.103	0.144
	3	0.5	0.055	0.079	0.107	0.085	0.188	0.291	0.054	0.084	0.112
		0.9	0.004	0.065	0.111	0.020	0.203	0.349	0.014	0.079	0.116
Under skewed distributions											
			$\lambda=1.0$ versus $\lambda=0.5$			$\lambda=2.0$ versus $\lambda=0.5$			$\lambda=2.0$ versus $\lambda=1.0$		
			θ	A_1	A_2	θ	A_1	A_2	θ	A_1	A_2
$1/b$	t	ν									
1	1	0.5	0.059	0.062	0.067	0.084	0.106	0.118	0.062	0.068	0.073
		0.9	0.086	0.103	0.116	0.237	0.308	0.372	0.098	0.118	0.136
	3	0.5	0.073	0.088	0.098	0.177	0.228	0.268	0.093	0.098	0.107
		0.9	0.053	0.093	0.107	0.237	0.236	0.291	0.104	0.093	0.109
2	1	0.5	0.057	0.063	0.068	0.073	0.093	0.123	0.057	0.064	0.073
		0.9	0.069	0.092	0.122	0.135	0.240	0.365	0.069	0.100	0.132
	3	0.5	0.056	0.077	0.096	0.090	0.172	0.259	0.056	0.081	0.103
		0.9	0.005	0.062	0.102	0.027	0.182	0.322	0.021	0.077	0.107

and actually positive subjects. We acknowledge that consideration of the ability to discriminate between subjects is but one aspect of analyzing FROC data. Indices for summarizing such discriminative ability attempt to mimic a process of combining information which might occur when the system is forced only to rate an entire subject. Hence, these indices ignore information on correct localization within subjects. For assessing an FROC system with respect to performance related factors that may be directly affected by correct localization, a different type of a summary index may be more appropriate.

The indices we proposed were developed in part by applying concepts commonly used in ROC analysis. Specifically, all indices are defined in the format of correctly discriminating in every possible actually negative - actually positive pair. Although in ROC analysis there is a well known relationship of the proportion of correct discriminations and the area under a corresponding ROC curve, our indices permit a more general discrimination for which there may be no corresponding ROC curve. Furthermore, even

when there is a corresponding ROC curve (e.g., our procedures based on the maximum or average), when applied to an FROC process, a system which has no ability to discriminate between an individual abnormality and non-abnormality on a location level may not result in the area under the subject-based ROC curve of 0.5. The proposed indices will be equal to 0.5 if the actually positive and actually negative subjects have not only the same distribution of the ratings but also the same distribution of the number of the rated marks.

The simulation model that we used for the investigation of the properties of the developed statistical procedures was based on the models commonly used in the FROC analysis.¹⁰ Our simulation model is simplistic since it uses simple parametric distributions, fixed number of lesions within an actually positive subject, and does not describe the correlations that are likely to exist in real FROC data. However, the proposed statistical procedure is based on re-sampling subjects as a unit, and hence it is not directly affected by shape of the distribution of the ratings or by the within-subject correlations.

TABLE VI. Estimated type I error rates and power for detecting differences resulting from differing average numbers of FP marks (λ) for $AUC=0.8$ under a simulation model based on normal distributions with a sample size of 100 subjects and 1000 simulations. The symbols in the table heading are the following simulation parameters: AUC - is the degree of separation between the distribution of FP and TP ratings; b - is the ratio of the standard deviation of FP ratings to that of TP ratings; λ - is the average number of FP marks per subject; ν - is the average proportion of the marked lesions; t - is the number of lesions on an actually positive subject.

Type I error rate			$\lambda=0.5$			$\lambda=1.0$			$\lambda=2.0$		
			θ	A_1	A_2	θ	A_1	A_2	θ	A_1	A_2
$1/b$	t	ν									
1	1	0.5	0.049	0.048	0.046	0.066	0.066	0.065	0.056	0.053	0.056
		0.9	0.056	0.054	0.056	0.046	0.044	0.037	0.057	0.057	0.052
	3	0.5	0.049	0.055	0.061	0.062	0.051	0.052	0.048	0.039	0.038
		0.9	0.029	0.051	0.052	0.045	0.040	0.034	0.052	0.056	0.051
	2	0.5	0.060	0.053	0.054	0.061	0.055	0.056	0.055	0.046	0.051
		0.9	0.048	0.053	0.046	0.058	0.062	0.066	0.049	0.053	0.054
2	1	0.5	0.058	0.040	0.044	0.057	0.053	0.049	0.054	0.064	0.055
		0.9	0.039	0.045	0.049	0.055	0.065	0.063	0.051	0.047	0.041
	3	0.5	0.058	0.040	0.044	0.057	0.053	0.049	0.054	0.064	0.055
		0.9	0.039	0.045	0.049	0.055	0.065	0.063	0.051	0.047	0.041
	2	0.5	0.058	0.040	0.044	0.057	0.053	0.049	0.054	0.064	0.055
		0.9	0.039	0.045	0.049	0.055	0.065	0.063	0.051	0.047	0.041
Statistical power			$\lambda=1.0$ versus $\lambda=0.5$			$\lambda=2.0$ versus $\lambda=0.5$			$\lambda=2.0$ versus $\lambda=1.0$		
			θ	A_1	A_2	θ	A_1	A_2	θ	A_1	A_2
$1/b$	t	ν									
1	1	0.5	0.090	0.114	0.132	0.207	0.376	0.431	0.101	0.136	0.153
		0.9	0.268	0.411	0.512	0.805	0.945	0.973	0.312	0.457	0.547
	3	0.5	0.160	0.255	0.314	0.582	0.818	0.878	0.234	0.348	0.378
		0.9	0.234	0.333	0.401	0.783	0.859	0.914	0.312	0.310	0.368
	2	0.5	0.071	0.087	0.129	0.125	0.302	0.462	0.076	0.121	0.162
		0.9	0.178	0.315	0.488	0.569	0.880	0.976	0.166	0.367	0.541
2	1	0.5	0.093	0.219	0.345	0.295	0.725	0.916	0.132	0.282	0.407
		0.9	0.128	0.328	0.486	0.462	0.811	0.961	0.141	0.265	0.428
	3	0.5	0.093	0.219	0.345	0.295	0.725	0.916	0.132	0.282	0.407
		0.9	0.128	0.328	0.486	0.462	0.811	0.961	0.141	0.265	0.428
	2	0.5	0.093	0.219	0.345	0.295	0.725	0.916	0.132	0.282	0.407
		0.9	0.128	0.328	0.486	0.462	0.811	0.961	0.141	0.265	0.428

V. CONCLUSION

We investigated several specific indices from a family of proposed indices quantifying subject-based discriminative ability of a FROC system. Discriminative ability at the subject level is important to assess separately, especially since a system superior at location level is not necessarily good at subject level and vice versa. We proposed a nonparametric method for statistical analysis of this type of index. The proposed statistical approach can be used to compare two indices and our simulations demonstrate a reasonable type I error rate for the considered indices for sample sizes as small as 20 actually positive and 20 actually negative subjects. Different indices of the considered type, despite their apparent similarity, characterize slightly different features of the FROC data. In the analysis of the statistical power we demonstrated that even using simple models for FROC data it is possible to construct the scenarios where the use of different indices leads to different conclusions. Thus, there is no statistically superior index for comparing subject-based discriminative ability of two arbitrarily different FROC systems. The choice

of an index should be based on clinical considerations or by utilizing information from other studies that provide information on how correct localization is related to subject-level discrimination for the particular task being addressed.

ACKNOWLEDGMENTS

This research was supported in part by Grant Nos. EB006388, EB002106, and EB003503 (to the University of Pittsburgh) from the National Institute for Biomedical Imaging and Bioengineering (NIBIB), National Institute of Health.

^{a)} Author to whom correspondence should be addressed. Electronic mail: gurd@upmc.edu

¹ J. P. Egan, G. Z. Greenberg, and A. I. Schulman, "Operating characteristics, signal detectability, and the methods of free response," *J. Acoust. Soc. Am.* **33**(8), 993–1007 (1961).

² 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**(4), 165–171 (1978).

³ J. M. Irvine, "Assessing target search performance: The free-response

- operator characteristic model," *Opt. Eng. (Bellingham)* **43**(12), 2926–2934 (2004).
- ⁴D. P. Chakraborty, "Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data," *Med. Phys.* **16**(4), 561–568 (1989).
- ⁵D. C. Edwards, M. A. Kupinski, C. E. Metz, and R. M. Nishikawa, "Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model," *Med. Phys.* **29**(12), 2861–2870 (2002).
- ⁶F. W. Samuelson and N. Petrick, "Comparing image detection algorithms using resampling," *Biomedical Imaging: Macro to Nano, 3rd IEEE International Symposium*, Arlington, VA, April 6–9, 2006, pp. 1312–1315.
- ⁷F. W. Wagner, C. E. Metz, and G. Compbell, "Assessment of medical imaging systems and computer aids: A tutorial review," *Acad. Radiol.* **14**, 723–748 (2007).
- ⁸R. G. Swensson, "Unified measurement of observer performance in detecting and localizing target objects on images," *Med. Phys.* **23**, 1709–1725 (1996).
- ⁹D. P. Chakraborty and K. S. Berbaum, "Observer studies involving detection and localization: Modeling, analysis and validation," *Med. Phys.* **31**(8), 2313–2330 (2004).
- ¹⁰D. P. Chakraborty, "A search model and figure of merit for observer data acquired according to the free-response paradigm," *Phys. Med. Biol.* **51**(14), 3449–3462 (2006).
- ¹¹D. Bamber, "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *J. Math. Psychol.* **12**, 387–415 (1975).
- ¹²J. A. Hanley and B. J. McNeil, "The Meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology* **143**(11), 29–36 (1982).
- ¹³J. N. Arvesen, "Jackknifing U-statistics," *Ann. Math. Stat.* **40**, 2076–2100 (1969).
- ¹⁴E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the area under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics* **44**(3), 837–845 (1988).
- ¹⁵D. P. Chakraborty, "ROC curves predicted by a model of visual search," *Phys. Med. Biol.* **51**, 3463–3482 (2006).
- ¹⁶X. H. Zhou, N. A. Obuchowski, and D. K. McClish, *Statistical Methods in Diagnostic Medicine* (Wiley, New York, 2002).