Subject-centered free-response ROC (FROC) analysis

Andriy I. Bandos^{a)} and Howard E. Rockette

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, 130 DeSoto Street, Pittsburgh, Pennsylvania 15261

David Gur

Department of Radiology, School of Medicine, University of Pittsburgh, 3362 Fifth Avenue, Pittsburgh, Pennsylvania 15261

(Received 22 January 2013; revised 16 March 2013; accepted for publication 20 March 2013; published 18 April 2013)

Purpose: To develop an approach of estimating subject-centered free-response receiver operating characteristic (FROC) curve for providing patient-centered inferences regarding detection-localization characteristics of a diagnostic system.

Methods: The authors examine properties of the conventional, target-centered, FROC curve and demonstrate that in scenarios where the diagnostic performance correlates with the total number of targets on a subject, it may lead to inadequate inferences from the perspective of possible benefits to a patient. Following solutions to patient-centered approaches in other applications, the authors define a subject-centered FROC curve and develop its formulation as a covariate-adjusted FROC curve. The authors also conduct a numerical study illustrating the relative properties of the conventional and subject-centered approach and provide an example.

Results: A simple-to-implement approach for estimating the subject-centered FROC curve and its overall index can be formulated as a type of stratified FROC analysis. The authors demonstrate that when diagnostic performance is associated with the number of targets, the diagnostic system with apparently superior target-centered characteristics (conventional approach) can be actually inferior from the subject-centered perspective. The authors show that under some clinically reasonable conditions the magnitude of disagreement in results could be substantial. An example from an actual observer performance study illustrates the natural setting where the developed approach would be relevant and lead to conclusions that are contradictory to those obtained from conventional analysis.

Conclusions: The authors developed a subject-centered FROC curve and its overall index provides tools for inferences that may be relevant from a perspective of potential benefits to a patient. © 2013 American Association of Physicists in Medicine. [<http://dx.doi.org/10.1118/1.4799843>]

Key words: diagnostic tests, FROC curve, patient-centered inferences, cluster-specific analysis, population-averaged analysis

I. INTRODUCTION

In many fields including quality control, industrial and military monitoring, weather forecasting, medical diagnostics, and decision making the problem of classification and prediction is of great importance. Various topics related to the evaluation of diagnostic systems have been thoroughly discussed earlier^{1,2} and in more recent^{3–5} books. The performance evaluation of diagnostic systems is most commonly conducted within the general framework of a receiver operating characteristic (ROC) analysis. However, the complex nature of the diagnostic problems in different fields has led to a number of extensions of classical ROC analysis. One of such extensions is the “free-response” ROC (FROC) analysis which is designed to address tasks that combine detection and localization of multiple targets per subject (masses depicted on an imaging examination of the breast, specific sounds in a given temporal interval, etc.).

The tasks of detection and localization of multiple targets per subject result in randomly clustered data, where varying number of diagnostic outputs are clustered within a subject (e.g., identifications of suspected targets on an image).

The methodology for assessing performance in detection-and-localization tasks originated in problems related to acoustics and imaging^{6,7} and leads to the currently common FROC paradigm. Under the FROC protocol a diagnostic system can produce an unrestricted number of marks at any locations within a subject and supplement every mark with a confidence rating. A common basic tool for summarizing FROC data is the FROC curve.⁷

Several parametric approaches have been developed for analyzing FROC curves (e.g., Refs. 8 and 9). Conventional FROC approaches are driven by “target-centered,” or marginal, performance measures, which are interpretable in the context of the entire population of targets (ensemble of all targets from all subjects) rather than in the context of the population of subjects.

The problem of difference in interpretation of results of “target-centered” (interpretable for the population of targets) versus “subject-centered” (interpretable for the population of subjects) analyses has been acknowledged and addressed in other areas. In analyses of multireader ROC performance studies the analogous problem relates to the reader-centered (or reader-averaged) versus case-centered approaches, both of

which are used in practice (e.g., Refs. 10–13). Similar problems exist in analyses of multicenter studies (e.g., Ref. 14). In general, the difference between the two types of analysis can be attributed to the problem of the informative cluster size, methodological aspects of which were discussed by Hoffman *et al.*¹⁵ and Williamson *et al.*¹⁶

The idea of considering some type of per-subject characteristics in analyses of detection-localization studies has been exploited before. For example, one of the first approaches to analyze a simple (at most one target per subject) detection-localization task—localized receiver operating characteristics (LROC) curve¹⁷—uses the per-subject summary of false positive findings in the form of a frequency of “normal” subjects with FP findings. The same subject-based characteristic was later used in the analysis of more complex FROC data, by means of the alternative FROC (AFROC) analysis, proposed by Chakraborty and Winter.¹⁸ A subject-weighted version of the Wilcoxon statistics was used by Chakraborty and Berbaum¹⁹ in the development of JAFROC-1 and JAFROC-2 indices for summarizing conventional FROC data. Several indices based on the FROC data reduced to a per-subject level have been also considered by our group.²⁰ Subject-centered formulations have also been used for analysis of a simple “binary” (rating-free) task of detecting multiple targets (e.g., Ref. 21). However, subject-centered FROC curve itself and its corresponding summary indices have not been developed to date. This is the primary purpose of this paper.

I.A. Illustration of the two types of approaches in a simple binomial problem

One of the important problems stemming from the differences between target-centered and subject-centered approaches is that they could lead to contradictory study conclusions even in the simplest settings. In the context of a simple detection-localization problem this can be illustrated on the following example of relative magnitude of average ($\bar{p}$) and pooled ($\tilde{p}$) proportions of detected targets, which represent the trivial instances of the subject-centered and target-centered characteristics correspondingly. Consider a sample composed of subjects with a single abnormal target (n_1) and two abnormal targets (n_2). Let diagnostic system “I” have the same detection ability for both types of subjects ($p_1^{(I)} = p_2^{(I)} = 0.675$); consequently, it has the same overall pooled and averaged detection probability ($p_1^{(I)} = p_2^{(I)} = \bar{p}^{(I)} = \tilde{p}^{(I)} = 0.675$). Let diagnostic system “II” have superior (as compared with system “I”) ability to detect a single abnormality ($p_1^{(II)} = 0.9$), but a limited ability to detect subsequent abnormalities within the same subject ($p_2^{(II)} = 0.45$).

As illustrated in Fig. 1, the conclusion as to which system has higher detection ability can be contradictory for analyses based on pooled and average probability. The range where the qualitative difference can arise depends on the assumed probability parameters and sample composition. Figure 1 considers the sample composition where the subjects with multiple targets are less prevalent, i.e., $n_1/n_2 > 1$, which is typical di-

FIG. 1. The pooled and average detection probabilities of two diagnostic systems as a function of sample composition.

agnostic imaging. Real-life examples of contradictory conclusions of two approaches are discussed in Sec. V.

I.B. Significance of the problem and structure of the paper

As discussed further in the paper, the systematic discrepancy between target-centered and subject-centered characteristics can result when both numbers of disease targets within subjects and the performance characteristics for subjects with different number of targets are heterogeneous. The characteristics of both true positive and false positive marks with various ratings could differ among the groups of subjects (Fig. 2). However, as we discuss later, only variability of the former (i.e., detection rate achieved for adequately suspicious marks) is of concern in FROC analysis. These scenarios can happen very frequently as many experimental datasets include a mix of subjects with differing numbers of disease targets. Furthermore, in many fields performance characteristics for subjects with a different number of targets are frequently heterogeneous. Heterogeneity of performance characteristics for the images with different number of disease targets is frequently observed in medical imaging (e.g., Refs. 22 and 23) where one of the possible contributing factors is a “satisfaction of search” effect.²⁴ In these frequent settings when the numbers of targets covary with performance characteristics the patient-centered approach becomes more relevant from the perspective of possible benefit to a patient.

II. CONVENTIONAL FROC APPROACH

Given known locations of all targets, each mark (output) produced by a diagnostic system can be classified as a “true positive,” or TP (i.e., adequately localizes a true target) or “false positive,” or FP (does not adequately localize any target).

Similar to ROC analysis, in the FROC realm TP responses are quantified with a true positive fraction (or TPF), defined as a marginal probability to localize a true target, i.e., $TPF = P(H = 1)$, where $H = 1$ indicates an event of detection

FIG. 2. A graphic illustration of the true positive marks, false positives marks, and missed targets (or false negative results).

and adequate localization of a randomly selected target. Unlike ROC, however, in the FROC setting FP marks cannot be quantified in terms of a conventional FPF proportion since the number of true negative responses is undefined. Instead, a traditional FROC approach uses false positive rate (FPR), which is defined as the expected number of false positive responses per subject, i.e., $FPR = E(N)$, where N is the number of FP responses on a randomly selected subject. A traditional rating-based FROC protocol allows a diagnostic system to supplement each mark with a rating indicating the degree of suspicion regarding the status of the identified suspected (X for FP and Y for TP ratings). The ratings are then used to recover the inevitable tradeoff between the numbers of TP and FP responses by dichotomizing according to their ratings at multiple values of threshold ξ . $TPF(\xi)$ and $FPR(\xi)$ are the TPF and FPR computed using only the marks with ratings higher than ξ , in other words,

$$FPR(\xi) = E \left\{ \sum_{i=1}^N I(X_i > \xi) \right\} \quad \text{and} \quad (1)$$

$$TPF(\xi) = E \left\{ \sum_{j=1}^M I(Y_j > \xi) / T \right\},$$

where M is the total number of TP marks and T is the total number of targets. Accuracy of all marks regardless of their ratings is characterized with $TPF(-\infty)$ and $FPR(-\infty)$ and distinguished with subscript π , i.e., $FPR_{\pi} = FPR(-\infty) = E(N)$, $TPF_{\pi} = E(M/T)$. TPF_{π} and FPR_{π} are sometimes termed “pruning characteristics” and represent important characteristics of the diagnostic performance. The FROC curve summarizes operating characteristics for all threshold values, i.e., FROC: $[FPR(\xi), TPF(\xi)]$.

Several indices can be used for evaluation of diagnostic performance in FROC-type tasks. For example, one can estimate TPF characteristics corresponding to a given FPR rate,¹⁸ compare areas under the curves residing in a common range,²⁵ or evaluate and/or compare the average “superiority” over the guessing processes.²⁶ Other popular indices (e.g., Ref. 19) summarize FROC data circumventing the conventional FROC curve. Here, we focus on a summary index of the FROC curve²⁶ which can be computed in closed

form and offers a direct analogy to the area between the ROC curve and the diagonal reference line in conventional ROC analysis. This index can be expressed in the following manner:

$$\Lambda = \int TPF(\xi) dFPR(\xi) - FPR_{\pi} + TPF_{\pi} / \varphi, \quad (2)$$

where $\int TPF(\xi) dFPR(\xi)$ is the area under the empirical FROC curve and φ is the guessing parameter reflecting a degree of difficulty of detecting targets in a given sample of subjects. The integral in Eq. (2) can be simplified by noting that conditional on the number of marks the standardized quantities $TPF(\xi)/TPF_{\pi}$ and $FPR(\xi)/FPR_{\pi}$ are the survival distribution functions of the ratings for true positive $[P(Y > \xi)]$ and false positive marks $[P(X > \xi)]$ correspondingly. The simplified formula for Λ in the presence of continuous ratings can then be written as follows:

$$\Lambda = P(X < Y) FPR_{\pi} TPF_{\pi} - FPR_{\pi} + TPF_{\pi} / \varphi. \quad (3)$$

A conventional FROC study results in rating data that can be represented as follows:

$$\{ \{x_{si}\}_{i=1}^{n_s}, \{y_{sj}\}_{j=1}^{m_s}, t_s \}_{s=1}^S, \quad (4)$$

where s indicates one of S subjects used in the study; t_s denotes the number of targets contributed by the s th subject; m_s ($m_s \leq t_s$) indicates the observed number of TP marks with ratings $\{y_{sj}\}_{j=1}^{m_s}$; and n_s represents the observed number of FP marks with ratings $\{x_{si}\}_{i=1}^{n_s}$. When needed we use the subscript t to distinguish data related to the subject with exactly t targets. In these notations, the empirical estimate of the FROC curve has the following form:

$$\widehat{FPR}(\xi) = \sum_{s=1}^S \sum_{i=1}^{n_s} I(x_{si} > \xi) / S$$

$$\widehat{TPF}(\xi) = \sum_{s=1}^S \sum_{j=1}^{m_s} I(y_{sj} > \xi) / \sum_{s=1}^S t_s. \quad (5)$$

Similarly, the empirical pruning characteristics are $\widehat{FPR}_{\pi} = \sum_{s=1}^S n_s / S$, $\widehat{TPF}_{\pi} = \sum_{s=1}^S m_s / \sum_{s=1}^S t_s$, and the empirical estimate of Λ can be expressed as follows:

$$\hat{\Lambda} = \frac{\sum_{s=1}^S \sum_{i=1}^{n_s} \left\{ \sum_{s'=1}^{S-S_0} \sum_{j=1}^{m_{s'}} \psi(x_{si}, y_{s'j}) \right\}}{S \sum_{s=1}^S t_s}$$

$$- \frac{\sum_{s=1}^S n_s}{S} + \frac{1}{\varphi} \times \frac{\sum_{s=1}^S m_s}{\sum_{s=1}^S t_s}.$$

II.A. Target-centered nature of conventional FROC analysis

Conventional FROC analyses express the overall performance characteristics of a diagnostic system in a

target-centered manner, i.e., in the context of population of all targets in the ensemble of subjects. Indeed, one of the primary FROC characteristics at a fixed threshold, TPF, is the marginal probability of detection $TPF = P_{\Omega}(H = 1)$ which indicates the frequency of actually detected targets out of all known targets in all subjects, or more generally,

$$\begin{aligned} TPF(\xi) &= P_{\Gamma}\{H(\xi) = 1\} \\ &= \int P(H(\xi) = 1|\gamma) \times f_{\Gamma}(\gamma) d\gamma, \end{aligned}$$

where $f_{\Gamma}(\gamma) d\gamma$ reflects the probability of selecting a target from the population of targets Γ ($1/T$ is a finite population of T targets), and $H(\xi) = I(Y > \xi)$ if the target was marked and 0 otherwise.

Conventional TPF can be viewed as a weighted average of the per-subject characteristics, $P(H = 1|s)$, where the weights, $f_{\Gamma}(s) ds$, represent the proportion of targets that are contributed by a given subject s , i.e.,

$$\begin{aligned} TPF(\xi) &= \int P(H(\xi) = 1|\gamma) \times f_{\Gamma}(\gamma) d\gamma \\ &= \int \left\{ \underbrace{\int P(H(\xi) = 1|\gamma, s) \times f(\gamma|s) d\gamma}_{P(H(\xi)=1|s)} \right\} \\ &\quad \times f_{\Gamma}(s) ds = \int P\{H(\xi) = 1|s\} \times f_{\Gamma}(s) ds, \end{aligned} \quad (6)$$

where $f_{\Gamma}(s) ds$ reflects the probability that a randomly selected target belongs to the s th subject (e.g., $f_{\Gamma}(s) ds = t_s/\Sigma t_s$ in the finite population of targets from the S subjects). Equation (6) has a natural empirical counterpart,

$$\begin{aligned} \widehat{TPF}(\xi) &= \sum_{s=1}^S \sum_{j=1}^{m_s} I(y_{sj} > \xi) / \sum_{s=1}^S t_s \\ &= \sum_{s=1}^S \frac{t_s}{\sum_{s=1}^S t_s} \times \left\{ \frac{\sum_{j=1}^{m_s} I(y_{sj} > \xi)}{t_s} \right\} \\ &= \sum_{s=1}^S w_s \times \{\widehat{TPF}^s(\xi)\}. \end{aligned} \quad (7)$$

As a result of the target-centered nature of TPF characteristic of the conventional FROC curve, indices summarizing the curve, including the Λ index, also have a target-centered meaning.

III. SUBJECT-CENTERED FROC APPROACH

Alternatively to the conventional target-centered approach, the frequency of the correctly localized targets can be quantified in a subject-centered manner:

$$\begin{aligned} TPF^{sc}(\xi) &= P_{\Sigma}\{H(\xi) = 1\} \\ &= \int P(H = 1|s) \times f_{\Sigma}(s) ds, \end{aligned} \quad (8)$$

where superscript “sc” indicates a subject-centered type of a quantity and $f_{\Sigma}(s) ds$ is the probability that the s th subject is selected under a simple random sampling of the population of subjects Σ [e.g., $f_{\Sigma}(s) ds = 1/S$ in a finite population composed of S subjects]. The empirical estimator of patient-centered TPF can be formulated using per-subject characteristics as follows:

$$\widehat{TPF}^{sc}(\xi) = \sum_{s=1}^S \frac{1}{S} \times \widehat{TPF}^s(\xi) = \frac{1}{S} \sum_{s=1}^S \frac{\sum_{j=1}^{m_s} I(y_{sj} > \xi)}{t_s}. \quad (9)$$

Experimental FROC data typically include only a few groups of subjects with a different number of targets (e.g., mammograms with 1, 2, or 3 visible masses). This allows rewriting the estimators in Eqs. (8) and (9) in a form that is more amenable for statistical inferences.

The development of the subject-centered FROC curve, its summary index, and their empirical estimates is based on the sequence of straightforward, but essential steps summarized in the three following propositions.

Proposition 1:

For a population of subjects with the same number of targets the subject-centered TPF is equivalent to the target-centered TPF, i.e., $TPF^{sc}(\xi) = TPF(\xi)$.

Proof:

The equivalence of the two theoretical quantities follows from definitions (6) and (8) and equality between $f_{\Gamma}(s) ds$ and $f_{\Sigma}(s) ds$ when each of the subjects has the same number of targets. In other words, in the homogeneous population the “proportional to size” sampling process is equivalent to the “simple random sampling.” [The equality of the empirical estimator follows from the homogeneity of w_s weights in Eq. (7) in a sample where all subjects have the same number of targets.] ■

Proposition 1 enables us to establish an explicit relationship between the subject-centered and strata-specific target-centered detection rates in a general case. We formulate the corresponding result in the following proposition.

Proposition 2:

The subject-centered detection rate, $TPF^{sc}(\xi)$, is equivalent to the stratified target-centered detection rate where the stratification is done according to the number of targets per subject and strata are weighted according to the number of subjects in each stratum, i.e.,

$$TPF^{sc}(\xi) = \int TPF(\xi|t) \times g_{\Sigma}(t) dt, \quad (10)$$

where $g_{\Sigma}(t) dt$ is the probability to select a subject with t targets.

Proof:

Considering the number of targets per subject as a function of a subject, the subject-centered detection rate in Eq. (8)

could be rewritten in the following form:

$$\begin{aligned} \text{TPF}^{\text{sc}}(\xi) &= \int P(H = 1|s) \times f_{\Sigma}(s) ds \\ &= \int \left\{ \int P(H = 1|s) \times f_{\Sigma}(s|t) ds \right\} \times g_{\Sigma}(t) dt, \end{aligned}$$

where $g_{\Sigma}(t)dt$ is the fraction of subjects with t targets. As follows from Proposition 1, the inside integration is equivalent to the target-centered TPF in the subpopulation of subjects with t targets, i.e., $\int P(H(\xi) = 1|s) \times f_{\Sigma}(s|t) ds = \text{TPF}(\xi|t)$. Thus, the subject-centered TPF permits an interpretation as a weighted average of strata-specific instances of the conventional detection rates as in Eq. (10). ■

Since the number of targets is a discrete variable, Eq. (10) permits further simplification:

$$\text{TPF}^{\text{sc}}(\xi) = \sum_{t=1}^{t_{\max}} g_{\Sigma}(t) \times \text{TPF}(\xi|t). \quad (11)$$

In the above equation and later in the paper we, without loss of generality, assume that for any $t = 1, \dots, t_{\max}$ there are subjects with exactly t targets in the population. Representation (11) enables reformulating the empirical estimator for the subject-centered detection fraction as the weighted average of the group-specific target-centered sample detection rates [$\text{TPF}^t = \text{TPF}(\xi|t)$] with weights reflecting the overall composition of the sample, i.e.,

$$\begin{aligned} \widehat{\text{TPF}}^{\text{sc}}(\xi) &= \frac{\sum_{t=1}^{t_{\max}} \sum_{s=1}^{S_t} \widehat{\text{TPF}}_t^{\text{sc}}(\xi)}{\sum_{t=1}^{t_{\max}} S_t} \\ &= \sum_{t=1}^{t_{\max}} \left(\frac{S_t}{\sum_{\tau=1}^{t_{\max}} S_{\tau}} \right) \left(\frac{\sum_{s=1}^{S_t} \sum_{j=1}^{m_s} I(y_{tsj} > \xi)}{S_t} \right) \\ &= \sum_{t=1}^{t_{\max}} \left(\frac{S_t}{\sum_{\tau=1}^{t_{\max}} S_{\tau}} \right) \times \widehat{\text{TPF}}^t(\xi). \end{aligned} \quad (12)$$

Since the conventional $\text{FPR}(\xi)$ characteristic is already defined in the subject-centered manner, the subject-centered FROC curve describes a tradeoff between the conventional $\text{FPR}(\xi)$ and subject-centered $\text{TPF}^{\text{sc}}(\xi)$. Then, as an immediate consequence of Eq. (11), the subject-centered FROC curve is a weighted average of the FROC curves each of which is based on TP marks from the subjects with specific numbers of actual targets (t) and FP marks from all subjects in the sample. For averaging, these curves are calibrated by the rating (i.e., the combined operating points correspond to the same rating cutoff ξ). Using Eqs. (8) and (12) the empirical estimate of the subject-centered FROC curve can be written

as follows:

$$\begin{aligned} \widehat{\text{FPR}}^{\text{sc}}(\xi) &= \sum_{t=0}^{t_{\max}} \left(S_t / \sum_{\tau=0}^{t_{\max}} S_{\tau} \right) \times \widehat{\text{FPR}}^t(\xi) = \widehat{\text{FPR}}(\xi) \\ \widehat{\text{TPF}}^{\text{sc}}(\xi) &= \sum_{t=1}^{t_{\max}} \left(S_t / \sum_{\tau=1}^{t_{\max}} S_{\tau} \right) \times \widehat{\text{TPF}}^t(\xi). \end{aligned}$$

Using the above formulation of the subject-centered FROC curve we can define curve-related summary indices. In particular, one can easily compute the sensitivity corresponding to a given false positive rate (e.g., Ref. 18) or compute the partial area under the FROC curve (e.g., Ref. 25). One of these, the overall index of superiority over the guessing process Λ , which is equivalent to the area between the guessing-augmented FROC curve of a diagnostic system and the FROC curve of a pure guessing process, is considered below.

Proposition 3:

The area between the augmented subject-centered FROC curve and the guessing FROC curve can be expressed using conventional characteristics computed for subjects grouped according to the number of known targets, i.e.,

$$\begin{aligned} \Lambda^{\text{sc}} &= \sum_{t=1}^{t_{\max}} g_{\Sigma}(t) \times \left\{ \int \text{TPF}^t(\xi|t) d\text{FPR}(\xi) \right\} - \text{FPR}_{\pi} \\ &\quad + \sum_{t=1}^{t_{\max}} g_{\Sigma}(t) \times \text{TPF}^t \times \left(\frac{1}{\varphi} \right). \end{aligned} \quad (13)$$

Proof:

As a consequence of Proposition 1, the guessing parameter φ and corresponding guessing FROC curve are the same for both the target- and subject-centered approaches since the characteristics of the guessing process are not affected by the specific distribution of the targets across subjects. Hence, for the area between the augmented subject-centered FROC curve and the guessing FROC curve we can apply the original formula. Using the general definition in Eq. (2) the subject-centered Λ index can be written as follows:

$$\Lambda^{\text{sc}} = \int \text{TPF}^{\text{sc}}(\xi) d\text{FPR}(\xi) - \text{FPR}_{\pi} + \text{TPF}_{\pi}^{\text{sc}} \times \left(\frac{1}{\varphi} \right).$$

The second and third terms can be obtained by considering Eqs. (1) and (11) at $\xi = -\infty$. Following Eq. (11) the first term can be written as a weighted average of the areas under the conventional (target-centered) FROC curves each of which is constructed using the distribution of TP ratings in a subpopulation of subjects with a given number of targets and a distribution of FP ratings in the entire population, i.e.,

$$\begin{aligned} \int \text{TPF}^{\text{sc}}(\xi) d\text{FPR}(\xi) &= \sum_{t=1}^{t_{\max}} g_{\Sigma}(t) \times \left\{ \int \text{TPF}^t(\xi|t) d\text{FPR}(\xi) \right\}. \end{aligned} \quad (14)$$

By using Eqs. (1), (11) and (14) for combination in Eq. (2) we obtain Eq. (13). ■

We note that the subject-centered Λ is different from the correspondingly weighted average of the Λ 's computed for groups with fixed number of targets per subjects. The difference stems from the fact that the FPR used in Eq. (14) is based on all subjects rather than on a specific subgroup alone. For FROC data with continuous ratings we can obtain a subject-centered analog of a more detailed Λ representation (3) as follows:

$$\Lambda^{\text{sc}} = \text{FPR}_{\pi} \sum_{t=1}^{t_{\max}} g_{\Sigma}(t) \times P(X < Y^t) \times \text{TPF}_{\pi}^t - \text{FPR}_{\pi} + \left(\frac{1}{\varphi}\right) \sum_{t=1}^{t_{\max}} g_{\Sigma}(t) \times \text{TPF}_{\pi}^t, \quad (15)$$

where X is a rating for a false positive mark from a subject with any number of targets, and Y^t is a rating for a true positive marks from one of the subjects with t targets.

Proposition 3 provides a straightforward way for constructing an unbiased empirical estimator of the subject-centered Λ . The unbiased empirical estimators for the second and third summands are provided in Eqs. (5) and (12) correspondingly. The estimator for the first summand can be constructed as a weighted average of the estimators of the areas under the $t_{\max}$ empirical FROC curves, each of which is based on the TP marks from the subjects with t targets and False Positive marks from all subjects in the sample. Each area can be estimated by applying the previously developed U-statistic²⁶ to the corresponding sets of observations. As a result we obtain the following estimate for the subject-centered Λ :

$$\hat{\Lambda}^{\text{sc}} = \sum_{t=1}^{t_{\max}} \left(\frac{S_t}{\sum_{\tau=1}^{t_{\max}} S_{\tau}} \right) \frac{\sum_{s=1}^{S_t} \sum_{j=1}^{m_s} \left\{ \sum_{t'=0}^{t_{\max}} \sum_{s'=1}^{S_{t'}} \sum_{i=1}^{n_{s'}} \psi(x_{t's'i}, y_{tsj}) \right\}}{t S_t \left(\sum_{\tau=0}^{t_{\max}} S_{\tau} \right)} - \frac{\sum_{t=0}^{t_{\max}} \sum_{s=1}^{S_t} n_s}{\sum_{t=0}^{t_{\max}} S_t} + \frac{1}{\varphi} \sum_{t=1}^{t_{\max}} \left(\frac{S_t}{\sum_{\tau=1}^{t_{\max}} S_{\tau}} \right) \left(\frac{\sum_{s=1}^{S_t} m_s}{S_t t} \right). \quad (16)$$

For statistical inferences about the subject-centered Λ and other indices of the subject-centered FROC curve we propose using cluster-resampling techniques based on resampling subjects with all related observations, which were successfully used in analysis of FROC and clustered ROC data.^{25–27} Comparing to the subject-based bootstrap variance of the target-centered Λ we can expect a somewhat larger variability of the subject-centered indices in cases when within subject performance characteristics are homogeneous. For comparison of diagnostic modalities under the frequently employed paired design the permutation test can be used.^{25,28} We provide an illustration of the statistical properties of such a test in Sec. V. In Sec. V we provide an example in which we compare performances of two diagnostic modalities in detection and localization abnormal masses in breast.

IV. EXAMPLE

We provide herein a scenario from the actual diagnostic accuracy study where conditions favor contradictory results between the subject-centered and the conventional target-centered approaches, and in which the target-centered approach “inflates” performance level of a diagnostic technology.

This example is based on an observer performance study of digital breast tomosynthesis (DBT) technology.²³ This study included several modalities related to DBT and conventional full field digital mammography (FFDM), including a rudimentary modality “Frames” which was considered for secondary purposes in the original study.

“Frames” represented the unprocessed set of multiple low dose projection images acquired under the DBT procedure. Assessment of this image set for detecting and localizing “mass-like” breast abnormalities provides a natural environment for arising of contradictory conclusions of the target-centered and subject-centered analyses. Indeed, in mammography 3D technologies are known to help resolve confusing tissue overlap natural for the conventional 2D projection imaging. As an immediate result the corresponding techniques could improve the ability to detect secondary abnormalities within the same subjects. But, the underdeveloped 3D techniques could at the same time have lower detection ability for individual abnormalities, which is exactly the case for the “Frames” modality. Under these circumstances use of the conventional, target-centered techniques for diagnostic performance evaluation could place undue emphasis on the underdeveloped diagnostic system merely because of detection of secondary lesions in subjects with multiple abnormalities. At the same time the subject-centered approach would reveal the potential for decreased chances of detecting single abnormality in a subject, thereby increasing potential for missing patients with the disease.

Below we provide numerical results. Although the results of the comparison using target-centered and subject-centered estimates are of different signs, we note that the magnitude of the differences in this dataset is rather small due primarily to the small number of subjects with different numbers of mass-like breast abnormalities in this study.

For the purpose of this example we focus on the relative performance of radiologists in evaluating “mass-like” breast abnormalities in a subset of 100 cases using the following two imaging modes: (1) “Frames” and (2) conventional “FFDM”—a conventional 2D diagnostic technology. We use a subset of the data from the original study including 47 cases with no abnormalities, 44 cases with a single mass, and 9 cases with two abnormal masses. Cases were evaluated by each of eight participating radiologists under several paradigms including FROC. For brevity of presentation we pool observations from all readers and ignore reader-specific factors. Such data pooling is sometimes performed in multi-reader studies²⁹ but is generally not recommended. However, for illustration of the approach proposed in this paper we ignore related problems.

FIG. 3. The target-centered (left) and subject-centered (right) FROC curves for the “Frames” (dots) and “FFDM” (stars) modalities.

In agreement with Eqs. (5) and (12) we constructed the empirical estimates of the target-centered (conventional) and subject-centered (newly proposed) FROC curves. The empirical FROC curves are provided in Fig. 3. Overall, a lower fraction of all depicted disease targets were detected under FFDM mode than under “Frames” mode (TPF_π of 0.59 for FFDM versus 0.61 for Frames), while the total number of False Positive marks was approximately the same under both modalities ($\text{FPR}_\pi \approx 0.30$). In FROC tasks the detection of additional targets (i.e., higher TPF) is rather difficult to achieve and is often valued more than an improved rating separation between already identified FP and TP locations. Hence, even a small improvement in detection rate of 0.02 might be considered important. Thus, if the improvement is real, the “Frames” modality can be argued to be better than “FFDM” based on the conventional target-centered analysis. The improvement in detection rate is reflected in the Λ index for Frames being larger than that for FFDM for some reasonable values of the guessing parameter (including φ of 0.11, 0.38, and 1.00 considered later in Sec. V).

In contrast to conventional target-centered analysis, the subject-centered FROC curve for Frames terminates at a point lower than that for FFDM, without any crossing. Thus, regardless of sample size, from a subject-centered perspective “Frames” do not provide any advantage in the detection and localization task. Naturally, the subject-centered Λ index

for FFDM is higher regardless of the value of the guessing parameter.

V. NUMERICAL STUDY

In this section we discuss important tradeoffs between statistical properties of the target-centered and subject-centered analyses. In our numerical studies we use $S_0 = 120$ target-free subjects ($t = 0$), $S_1 = 80$ subjects with a single target ($t = 1$), and $S_2 = 40$ subjects with $t = 2$ disease targets. The results are presented for the three values of the guessing parameter φ : 0.11, 0.38, and 1.00—which, for given sample composition, correspond to the 10%, 30%, and 60% probability of hitting a target by a mark randomly placed on a target-depicted subject.

First, we consider scenario where the true target-centered and subject-centered summary index differ in one modality, but not in the other. Under this scenario the statistical test for comparison of the two modalities using the target-specific summary index will have statistical power close to 0.05 (nominal type I error rate); at the same time a consistent statistical test based on the subject-centered index will have statistical power approaching 1 with increasing sample size. Thus, for this scenario it is more informative to demonstrate the differences between subject-centered and target-centered summary indices for various combinations of FROC characteristics.

TABLE I. Percent differences between the target- and subject-centered Λ resulting from a given imbalance between performance characteristics in subjects with a single and two targets. [The assumed proportion of subjects $S_0:S_1:S_2$ is 3:2:1. For all scenarios, the FPR is 2.0 and the target-centered TPF and $P(X < Y)$ both equal to 0.7. The corresponding value of the target-centered Λ varies only with φ . Considered values of the guessing parameter φ correspond to the probabilities 0.1, 0.3, and 0.6 of hitting a target with a single mark in a sample with a given composition.]

Guessing parameter	Between-group TPF imbalance ($\text{TPF}_1 - \text{TPF}_2$)	$\text{TPF}_{\text{subj}} - \text{TPF}_{\text{targ}}$	Between-group imbalance of rating separation [$P(X^1 < Y^1) - P(X^2 < Y^2)$]				
			0	0.2	0.3	0.4	0.5
			$\{\sum g(t)P(X < Y^t)\} - P(X < Y)$				
			0.00	0.03	0.05	0.07	0.08
$\varphi = 0.11$	0	0.00	0.0%	0.4%	0.6%	0.7%	0.9%
	0.2	0.03	5.5%	5.9%	6.1%	6.3%	6.5%
	0.3	0.05	8.3%	8.7%	8.9%	9.1%	9.3%
	0.4	0.07	11.0%	11.5%	11.7%	11.9%	12.1%
	0.5	0.08	13.8%	14.2%	14.4%	14.6%	14.9%
$\varphi = 0.38$	0	0.00	0.0%	2.3%	3.5%	4.6%	5.8%
	0.2	0.03	9.5%	11.9%	13.1%	14.3%	15.6%
	0.3	0.05	14.2%	16.7%	18.0%	19.2%	20.4%
	0.4	0.07	19.0%	21.5%	22.8%	24.1%	25.3%
	0.5	0.08	23.7%	26.3%	27.6%	28.9%	30.2%
$\varphi = 1.00$	0	0.00	0.0%	14.6%	21.9%	29.2%	36.5%
	0.2	0.03	25.0%	40.3%	47.9%	55.6%	63.2%
	0.3	0.05	37.5%	53.1%	60.9%	68.8%	76.6%
	0.4	0.07	50.0%	66.0%	74.0%	81.9%	89.9%
	0.5	0.08	62.5%	78.8%	87.0%	95.1%	103.3%

The qualitative differences between target- and subject-centered inferences are illustrated in Table I. The effect of subject-specific performance heterogeneity in detection fraction (TPF_{π}) and rating separability [$P(X < Y)$] is summarized in rows and columns correspondingly. The results in the third column demonstrate the discrepancy in the estimates of the detection rate under the two approaches. These results are relevant for the analyses focused on a single point of the FROC curve (e.g., Ref. 18) or on the analysis of the rating-free detection-localization tasks.²¹ The results in cells show the difference between subject-centered and target-centered Λ 's. (The target-centered Λ is equivalent to the subject-centered Λ for the scenario of zero between subject heterogeneity.) From Table I it is clear that the qualitative difference between target and subject-centered indices could also have substantial magnitude.

Second, we illustrate the differences in statistical power for detecting differences between two modalities when the subjects are homogeneous with respect to their performance characteristics (i.e., when true target- and subject-centered indices are the same). Under this scenario we expect the target-centered approach to have larger statistical power due to the efficiency of the pooled estimate in the homogeneous case. Each generated sample consisted of three groups of subjects: 120 subjects without targets ($t = 0$), 80 subjects with a single target ($t = 1$), and 40 subjects with 2 targets ($t = 2$). Each of three groups of subjects (with 0, 1, and 2 targets) had false positive rates (FPR_{π}) of 2.0 and an equal TPF_{π} as well as an equal $P(X < Y)$. The parameters of the rating distributions for marks within each group are selected to maintain the overall separability between the ratings at the level corresponding to $P(X < Y)$ of 0.7 or 0.9. The detection and localization fraction for all subjects depicting actual targets was TPF_{π}

= 0.7 or 0.8. The correlation between ratings of multiple marks within the same subject, the correlation between ratings obtained under different modalities, and negative correlation between the ratings and the number of findings were generated by sharing subject-specific random effects. For comparison of the diagnostic modalities, in this simulation study, we used the permutation test for paired clustered data.²⁵ For each of 250 generated samples we used a Monte Carlo approximation to the exact permutation test based on 250 randomly selected interchanges of subject-specific data between the two modalities.

The results in Table II illustrate that in the considered scenarios where either the target-centered or the subject-centered approach can be used to evaluate the differences between the diagnostic systems, loss of statistical power due to the use of the subject-centered approach can be as high as 0.08. This difference in statistical power is usually important in practical studies.

VI. DISCUSSION

We proposed a subject-centered formulation of the FROC curve for performance assessment in increasingly common tasks of detection and localization of multiple targets per subject. The developed approach provides more practically relevant results in frequent scenarios when the number of targets is associated with diagnostic performance. In such scenarios the proposed subject-centered and the conventional target-centered approaches could often lead to contradictory conclusions.

Similar problems have been considered in the general analysis of clustered data, with applications in various field such as dental medicine, reproductive toxicology, and health care

TABLE II. Estimated power for a comparison of two diagnostic modalities under the paired design using the permutation test. [The estimates are based on 250 simulated dataset each of which is analyzed based on 250 random permutations. The number of subjects in the groups with 0, 1, and 2 targets are 120, 80, and 40, respectively. FROC parameters in the headings identify the simulation scenario for the second diagnostic modality. The distribution of the output of the first diagnostic modality corresponds to $TPF = 0.7$ and $P(X < Y) = 0.7$, hence the corresponding cell provides the estimates for the type I error rate. For all scenarios the FPR is 2.0, and all characteristics had the same value for both groups of subjects with 1 and 2 targets. Considered values of the guessing parameter φ correspond to the probabilities 0.1, 0.3, and 0.6 of hitting a target with a single mark in a sample with a given composition.]

Guessing parameter	TPF	Method	$P(X < Y)$	
			0.7	0.9
$\varphi = 0.11$	0.7	Target-centered	0.05	0.10
		Subject-centered	0.06	0.09
	0.8	Target-centered	0.56	0.76
		Subject-centered	0.48	0.70
$\varphi = 0.38$	0.7	Target-centered	0.04	0.25
		Subject-centered	0.05	0.24
	0.8	Target-centered	0.52	0.93
		Subject-centered	0.48	0.90
$\varphi = 1.00$	0.7	Target-centered	0.04	0.48
		Subject-centered	0.04	0.48
	0.8	Target-centered	0.45	0.98
		Subject-centered	0.41	0.97

research.^{15,16,32} The general problem is frequently described as a nonignorable cluster size, which cannot be adequately addressed by the conventional marginal models.¹⁵ One of the solutions to this problem is to consider subject-specific rather than population-averaged estimates,¹⁵ which is asymptotically equivalent to using marginal estimates weighted according to the cluster size.¹⁶ In this general context our approach can be viewed as a simple model-free subject-centered (i.e., subject-specific) method for analysis of FROC data (which is frequently associated with nonignorable cluster-size), which can be implemented using reweighted target-centered (population-averaged) FROC estimates.

An alternative to the conventional FROC method of performance analyses in detection and localization task is provided by the region-of-interest (ROI) approach.^{30,31} However, due to the use of marginal probabilities in defining rates for clustered ROC data, the summary indices of the ROI approach are also target-centered. The subject-centered reformulation of the ROI performance curve and related indices could be developed using the approach similar to the one applied in this paper.

The developed overall summary index of the subject-specific FROC curve can also be viewed as a special case of the general weighted index with weights reflecting relative importance of subset of subjects. However, while selection of weights for the general index could be arbitrary, weights of subject-centered FROC curve are determined by the *a priori* known cluster size, and consequently, has an objective probabilistic interpretation. Using the derived formulation of the subject-centered FROC curve, many conventional summary indices could be derived in a straightforward manner.

We discussed several important properties of the subject-centered approach. In scenarios where the association between the number of targets and performance characteristics is not substantial, the estimates of subject-centered quantities could be numerically similar to their target-centered counterparts. In these scenarios, the subject-centered and target-centered differences between modalities are likely to be in the same direction and of similar magnitude; and a lower statistical power of subject-centered techniques can make them less attractive than conventional target-centered approaches. When statistical power becomes an issue, the subject-centered approach could be used for verifying robustness of the inferences.

The approach presented in this paper combines FROC curves corresponding to groups of subjects with different number of targets by averaging the operating points corresponding to the same rating threshold. In practice, the decision threshold regarding a subject could naturally depend on the number of known targets (which often impacts the number of reported findings). Hence, it would be natural to combine the group-specific curves by averaging the points corresponding to the same decision threshold, which would most likely correspond to different ratings in different groups. Thus, a natural extension of the approach presented here would be to consider vertical averaging as well as other practically reasonable types of calibration approaches. Another practically useful direction for future research would be to develop an approach for the analysis of multireader studies, which are commonly employed in the evaluation of detection and localization systems.

The developed approach most directly impacts diagnostic test evaluation field. Assessment of diagnostic performance is increasingly more focused on the detection and localization techniques used for identifications of multiple targets per subject (e.g., nodules in the lungs, affected lymph nodes, masses in the breast). The analytical method we develop helps address a frequently relevant question that cannot be answered by the conventional FROC curve and helps avoid erroneous impression of the superiority of practically suboptimal diagnostic techniques. As such, this method can improve multiple stages of development, optimization, evaluation, regulatory approval, and practical utilization of diagnostic systems by helping identify diagnostic systems that are beneficial in tasks where patient-centered characteristics are important.

ACKNOWLEDGMENT

This research was in part supported by the National Institute of General Medical Science of the National Institutes of Health under Award No. R01GM098253.

^{a)} Author to whom correspondence should be addressed. Electronic mail: an661@pitt.edu; Telephone: +00-412-383-5738; Fax: +00-412-624-2183.

¹ J. P. Egan, *Signal Detection Theory and ROC Analysis* (Academic, New York, 1975).

² J. A. Swets and R. M. Pickett, *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory* (Academic, New York, 1982).

- ³X. H. Zhou, N. A. Obuchowski, and D. K. McClish, *Statistical Methods in Diagnostic Medicine* (Wiley, New York, 2002).
- ⁴M. S. Pepe, *The Statistical Evaluation of Medical Test for Classification and Prediction* (Oxford University Press, Oxford, 2003).
- ⁵K. H. Zou, A. Liu, A. I. Bandos, L. Ohno-Machado, and H. E. Rockette, *Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis*, CRC Biostatistics Series (CRC/Chapman and Hall, Boca Raton, 2011).
- ⁶J. P. Egan, G. Z. Greenberg, and A. I. Schulman, "Operating characteristics, signal detectability, and the methods of free response," *J. Acoust. Soc. Am.* **33**, 993–1007 (1961).
- ⁷P. 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**, 165–171 (1978).
- ⁸D. P. Chakraborty, "Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data," *Med. Phys.* **16**, 561–568 (1989).
- ⁹D. 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**, 2861–2870 (2002).
- ¹⁰Y. Zheng, W. E. Barlow, and G. Cutter, "Assessing accuracy of mammography in the presence of verification bias and intrareader correlation," *Biometrics* **61**, 259–268 (2005).
- ¹¹X. Song and X. H. Zhou, "A marginal model approach for analysis of multi-reader multi-test receiver operating characteristic (ROC) data," *Biostatistics* **6**, 303–312 (2005).
- ¹²B. D. Gallas, G. A. Penello, and K. L. Myers, "Multireader multicase variance analysis for binary data," *J. Opt. Soc. Am.* **24**, B70–B80 (2007).
- ¹³H. Ishwaran and C. A. Gatsonis, "A general class of hierarchical ordinal regression models with applications to correlated ROC analysis," *Can. J. Stat.* **28**, 731–750 (2000).
- ¹⁴P. B. Bach, L. D. Cramer, D. Schrag, R. J. Downey, S. E. Gelfand, and C. B. Begg, "The influence of hospital volume on survival after resection for lung cancer," *N. Engl. J. Med.* **345**, 181–188 (2001).
- ¹⁵E. B. Hoffman, P. K. Sen, and C. R. Weinberg, "Within-cluster resampling," *Biometrika* **88**, 1121–1134 (2001).
- ¹⁶J. M. Williamson, S. Datta, and G. A. Satten, "Marginal analyses of clustered data when cluster size is informative," *Biometrics* **59**, 36–42 (2003).
- ¹⁷S. J. Starr, C. E. Metz, L. B. Lusted, and D. J. Goodenough, "Visual detection and localization of radiographic images," *Radiology* **116**, 533–538 (1975).
- ¹⁸D. P. Chakraborty and L. H. L. Winter, "Free-response methodology: Alternate analysis and a new observer-performance experiment," *Med. Phys.* **17**, 873–881 (1990).
- ¹⁹D. P. Chakraborty and K. S. Berbaum, "Observer studies involving detection and localization: Modeling, analysis and validation," *Med. Phys.* **31**, 2313–2330 (2004).
- ²⁰T. Song, A. I. Bandos, H. E. Rockette, and D. Gur, "On comparing methods for discriminating between actually negative and actually positive subjects with FROC type data," *Med. Phys.* **35**(4), 1547–1558 (2008).
- ²¹A. H. Zwinderman, A. S. Glas, P. M. Bossuyt, J. Florie, S. Bipat, and J. Stoker, "Statistical models for quantifying diagnostic accuracy with multiple lesions per patient," *Biostatistics* **9**(3), 513–522 (2008).
- ²²R. G. Swensson, "Unified measurement of observer performance in detecting and localizing target objects on images," *Med. Phys.* **23**, 1709–1725 (1996).
- ²³D. Gur, G. S. Abrams, D. M. Chough, M. A. Ganott, C. M. Hakim, R. L. Perrin, G. Y. Rathfon, J. H. Sumkin, M. L. Zuley, and A. I. Bandos, "Digital breast tomosynthesis: An observer performance study," *Am. J. Roentgenol.* **193**, 586–591 (2009).
- ²⁴K. S. Berbaum, E. A. Franken, D. D. Dorfman, S. A. Rooholamini, M. H. Kathol, T. J. Barloon, F. M. Behlke, Y. Sato, C. C. Lu, G. Y. El-Khoury, F. W. Flickinger, and W. J. Montgomery, "Satisfaction of search in diagnostic radiology," *Invest. Radiol.* **25**, 133–140 (1990).
- ²⁵F. W. Samuelson and N. Petrick, "Comparing image detection algorithms using resampling," *Biomedical Imaging: Macro to Nano, 3rd IEEE International Symposium Arlington, Virginia, 6–9 April* (Rockville, MD, 2006), pp. 1312–1315.
- ²⁶A. I. Bandos, H. E. Rockette, T. Song, and D. Gur, "Area under the free-response ROC curve (FROC) and a related summary index," *Biometrics* **65**, 247–256 (2009).
- ²⁷C. M. Rutter, "Bootstrap estimation of diagnostic accuracy with patient-clustered data," *Acad. Radiol.* **7**, 413–419 (2000).
- ²⁸A. I. Bandos, H. E. Rockette, and D. Gur, "A permutation test sensitive to differences in areas for comparing ROC curves from a paired design," *Stat. Med.* **24**, 2873–2893 (2005).
- ²⁹M. L. T. Lee and B. A. Rosner, "The average area under correlated receiver operating characteristic curves: A nonparametric approach based on generalized two-sample Wilcoxon statistics," *Appl. Stat.* **50**, 337–344 (2001).
- ³⁰N. A. Obuchowski, "Nonparametric analysis of clustered ROC curve data," *Biometrics* **53**, 567–578 (1997).
- ³¹N. A. Obuchowski, M. L. Lieber, and K. A. Powel, "Data analysis for detection and localization of multiple abnormalities with application to mammography," *Acad. Radiol.* **7**, 516–525 (2000).
- ³²K. S. Panageas, D. Schrag, A. R. Localio, E. S. Venkatraman, and C. B. Begg, "Properties of analysis methods that account for clustering in volume-outcome studies when the primary predictor is cluster size," *Statistics in Medicine* **26**, 2017–2035 (2007).