A Nonparametric Procedure for Comparing the Areas Under Correlated LROC Curves

Adam Wunderlich* and Frédéric Noo, *Member, IEEE*

***Abstract*—In contrast to the receiver operating characteristic (ROC) assessment paradigm, localization ROC (LROC) analysis provides a means to jointly assess the accuracy of localization and detection in an observer study. In a typical multireader, multicase (MRMC) evaluation, the data sets are paired so that correlations arise in observer performance both between readers and across the imaging conditions (e.g., reconstruction methods or scanning parameters) being compared. Therefore, MRMC evaluations motivate the need for a statistical methodology to compare correlated LROC curves. In this paper, we suggest a nonparametric strategy for this purpose. Specifically, we find that seminal work of Sen on U-statistics can be applied to estimate the covariance matrix for a vector of LROC area estimates. The resulting covariance estimator is the LROC analog of the covariance estimator given by DeLong *et al.* for ROC analysis. Once the covariance matrix is estimated, it can be used to construct confidence intervals and/or confidence regions for purposes of comparing observer performance across imaging conditions. In addition, given the results of a small-scale pilot study, the covariance estimator may be used to estimate the number of images and observers needed to achieve a desired confidence interval size in a full-scale observer study. The utility of our methodology is illustrated with a human-observer LROC evaluation of three image reconstruction strategies for fan-beam X-ray computed tomography.**

***Index Terms*—Confidence intervals, image quality, receiver operating characteristic (ROC), U-statistics.**

I. Introduction

FOR the purpose of evaluating and optimizing medical imaging technology, task-based image quality assessments employing receiver operating characteristic (ROC) analysis [1]–[3] are widely utilized in the medical imaging community [4]. However, traditional ROC lesion detection studies have the limitation that they do not incorporate the results of lesion localization, i.e., knowledge of whether or not the lesion is correctly localized is not used in the analysis. This limitation has motivated the development of assessment methodologies that evaluate overall detection performance together with the accuracy of lesion localization.

Manuscript received April 20, 2012; revised June 04, 2012; accepted June 06, 2012. Date of publication June 18, 2012; date of current version October 26, 2012. This work was supported in part by the National Institute of Health under Grant R01 EB007236 and Grant R21 EB009168 and in part by a generous grant from the Ben B. and Iris M. Margolis Foundation. *Asterisk indicates corresponding author.*

*A. Wunderlich is with the Department of Radiology, University of Utah, Salt Lake City, UT 84108 USA (e-mail: awunder@ucair.med.utah.edu).

F. Noo is with the Department of Radiology, University of Utah, Salt Lake City, UT 84108 USA (e-mail: noo@ucair.med.utah.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2012.2205015

Two popular image quality evaluation strategies that assess both lesion localization and detection performance are the localization ROC (LROC) [5] and the free-response ROC (FROC) paradigms [6], [7, Ch. 8]. In an LROC experiment, each image contains at most one lesion, and the observer scores and marks the most suspicious location. The FROC paradigm is more general than the LROC approach in that an unknown number of lesions is present in each image, and the observer scores and marks a potentially unlimited number of suspicious locations. An advantage of the FROC paradigm over the LROC approach is discrimination power, which may be stronger in some settings due to the greater complexity of the task [8]. On the other hand, the LROC paradigm offers the advantages of simplicity and lower cost and may still provide better discrimination power than conventional ROC analysis, as demonstrated under specific modeling assumptions in [8] and [9]. Both approaches, FROC and LROC, are regularly used in the medical imaging community, e.g., see [10]–[13] for applications of FROC methodology and [14]–[18] for applications of LROC analysis. Note that aside from LROC and FROC methodology, other approaches have been proposed to analyze lesion localization and detection performance; for example, the ROI approach of Obuchowski *et al.* [19] can be considered. In this work, we focus on LROC analysis.

Just as the area under the ROC curve ($\mathcal{A}$) is a useful figure of merit for ROC studies [20], the area under the LROC curve ($\mathcal{A}_L$) is a suitable figure of merit for LROC analysis [21]. In an influential paper, Swensson [21] suggested a semiparametric "binormal" estimation strategy for the LROC curve and its area, $\mathcal{A}_L$. Unfortunately, because Swensson's estimation strategy makes relatively strong assumptions regarding the observer's search process, it is not always applicable. More recently, Popescu [9] introduced a family of nonparametric estimators for the LROC curve and $\mathcal{A}_L$ that avoid the restrictive assumptions of Swensson. Also, Popescu [9] provided corresponding variance estimators for his area estimates. In subsequent investigations, a nonparametric $\mathcal{A}_L$ estimator of Popescu was studied and generalized by Tang *et al.* [22], [23]. The papers [9] and [21]–[23] are limited to consideration of uncorrelated LROC curves. Because correlations in observer performance arise in typical multireader, multicase (MRMC) study designs, it is necessary to use statistical methodologies that can account for these correlations. One way to deal with this problem is to employ general ANOVA-based methodologies developed for MRMC analysis, e.g., [24]–[26].

Here, we present a simple, fully nonparametric approach for MRMC analysis of areas under LROC curves. To do this, we first observe that one nonparametric estimator of $\mathcal{A}_L$ introduced by Popescu [9] can be rewritten as a generalized U-statistic that

is similar to the Mann-Whitney U estimator for $\mathcal{A}$.[1] This observation enables us to apply the seminal work of Sen [28] on U-statistics to estimate the covariances between correlated estimates of $\mathcal{A}_L$. The resulting nonparametric covariance estimator is the LROC analog of the popular covariance estimator of DeLong *et al.* [29] for ROC analysis.

An important consideration in the design and analysis of an MRMC study is whether readers should be treated as a fixed or random effect [7, p. 132]. If readers are modeled as a random effect, they are assumed to be randomly drawn from a larger population. In this case, the statistical analysis must account for the variability due to reader sampling. On the other hand, when readers are modeled as a fixed effect, such variability is not included. Both types of analyses have their uses, depending on the goals of the study, i.e., whether inference is to be made for a large reader population or for a smaller, fixed pool of readers. From a practical viewpoint, treating readers as a random effect results in larger confidence intervals [7, p. 132] and therefore requires more readers and images to attain acceptable statistical power. For the purpose of early-stage evaluations, smaller study designs are more practical, and it is often preferable to treat readers as a fixed effect. Like the application of the DeLong *et al.* approach [29] to MRMC ROC analysis [30], the procedure presented in this paper for LROC analysis treats readers as a fixed effect.

Another aspect in the design of an LROC study is whether to use a continuous scale or an ordinal (discrete) scale for the observer ratings. In the ROC literature, there is evidence that using an ordinal scale can lead to unreliable, highly variable performance estimates if not used properly by the reader [31], [32]. Consequently, according to Metz [3], continuous rating scales are generally recommended for observer studies by many experts. For this reason and for expositional clarity, the development in this paper focuses on LROC studies with a continuous rating scale. However, it turns out that the present work also applies to ordinal rating data with only slight modifications. For the interested reader, the necessary modifications for ordinal ratings are discussed in Appendix A.

II. Background: LROC Curves

This section summarizes LROC curves together with material needed to present our results. For further details regarding LROC curves and their interpretation, see [5], [9], [21], and [33].

Consider a lesion detection and localization task in which, for a given image, the observer: 1) decides whether or not a lesion is present and 2) identifies the most likely position of a lesion. Here, each image either contains exactly one lesion or does not contain a lesion. For the purpose of characterizing observer performance on a detection and localization task, Starr *et al.* [5] introduced the localization ROC (LROC) curve. An LROC curve plots the joint probability of a true positive and correct lesion localization, i.e., the true positive localized fraction (TPLF), as a function of the probability of a false positive, i.e., the false positive fraction (FPF), [5]. An example of an LROC curve is plotted in Fig. 1.

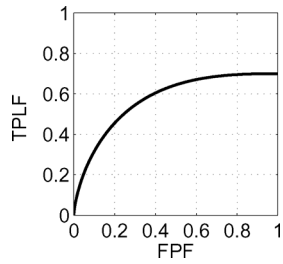

Fig. 1. Example of an LROC curve.

To measure an observer's LROC curve, an observer is presented with a sequence of images. For each image, the observer is asked to mark the most probable location of a lesion and to provide a rating regarding their confidence that a lesion is present. Suppose that the confidence rating is continuous-valued with higher ratings indicating a preference for a lesion-present image. According to the usual model, it is assumed that the observer decides in favor of lesion presence if the confidence rating is larger than a threshold c, and in favor of lesion absence otherwise [5].

Denote the set of images without a lesion as class 1 and the set of images with a lesion as class 2. In addition, denote the rating statistic for a class-1 image as X, the rating statistic for a class-2 image as Y, and the event that the lesion in a class-2 image is correctly localized as $\mathcal{L}$. Here, X and Y are independent, continuous random variables, and any event defined only with X is independent of $\mathcal{L}$. In general, $\mathcal{L}$ and any event involving Y are dependent.

In this setting, the FPF as a function of the decision threshold c is $\text{FPF}(c) = P(X \geq c)$. (Note that throughout the text we use the notation $P(A)$ to denote the probability of an event A.) Similarly, the TPLF as a function of c is $\text{TPLF}(c) = P(Y \geq c \cap \mathcal{L}) = P(Y \geq c \mid \mathcal{L})P(\mathcal{L})$, where $P(\mathcal{L})$ is the probability of correct localization. By definition, the LROC curve plots $\text{TPLF}(c)$ versus $\text{FPF}(c)$ as c varies over the range $(-\infty, \infty)$. To emphasize the functional dependence of TPLF on FPF, we will also use the notation $\text{TPLF}(\text{FPF})$ for the LROC curve. From the previous formula for TPLF, it follows that the value of the LROC curve at $\text{FPF} = 1$ is the probability of correct localization, i.e., $\text{TPLF}(\text{FPF} = 1) = \text{TPLF}(c = -\infty) = P(\mathcal{L})$. As an aside, observe that $P(\mathcal{L})$ is usually less than one, and therefore, the LROC curve does not generally end at the (1, 1) point, unlike the ROC curve; this is a fundamental difference between LROC and ROC curves.

The area under the LROC curve, $\mathcal{A}_L$, is a useful figure of merit that has a probabilistic interpretation as described by the following theorem. This well-known result was first observed without proof in an early paper by Metz *et al.* [34] and was later proved in [35] under restrictive assumptions. A simpler and fully general proof is given as follows.

Theorem 1: If X and Y are independent, continuous random variables, and $\mathcal{L}$ is independent of any event defined only with X, then $\mathcal{A}_L = P(Y > X \cap \mathcal{L})$.

[1]The Mann-Whitney statistic is a particular example of a U-statistic, which is a general type of unbiased, nonparametric statistic. For this reason, it is often called the "Mann-Whitney U statistic;" see [27].

Proof: Denote the probability density function (pdf) of X by $f_X(x)$. By definition, $\mathcal{A}_L = \int_0^1 \text{TPLF}\, d(\text{FPF})$. The expressions for TPLF and FPF given earlier imply that

$$\mathcal{A}_L = P(\mathcal{L}) \int_{-\infty}^{\infty} P(Y > c \mid \mathcal{L}) f_X(c) dc. \tag{1}$$

Since $\mathcal{L}$ and any event involving X are independent, and since X and Y are independent, $\mathcal{A}_L = P(\mathcal{L})P(Y > X \mid \mathcal{L}) = P(Y > X \cap \mathcal{L})$. ■

Recall that when X and Y are independent, continuous random variables, the area under the ROC curve is $\mathcal{A} = P(Y > X)$ [20, Result 4.6, p. 78]. Therefore, the previous theorem implies that $\mathcal{A} = \mathcal{A}_L + P(Y > X \cap \mathcal{L}^{\mathrm{C}})$, where the superscript "C" denotes the event complement. If additional modeling assumptions are made for the observer's search and decision process, then more specific relationships can be derived between $\mathcal{A}_L$ and $\mathcal{A}$, e.g., see [21]. Here, we avoid additional assumptions on the observer to maintain generality.

III. Nonparametric Estimation of the Area Under the LROC Curve

Over the course of an LROC study, suppose that an observer generates m independent, identically distributed (i.i.d.) class-1 ratings $X_1, X_2, \ldots, X_m$ and n i.i.d. class-2 ratings $Y_1, Y_2, \ldots, Y_n$. Denote the ratings of class-2 images with correct lesion localization as $Y_{l_1}, Y_{l_2}, \ldots, Y_{l_L}$, which is a length L subsequence of $Y_1, Y_2, \ldots, Y_n$ with $L \leq n$. Since lesion localization for a class-2 image results in either success or failure, it can be modeled as a Bernoulli trial with probability of success, $\lambda = P(\mathcal{L})$. Hence, it follows that the number of correctly localized class-2 images, L, is a binomial random variable with parameters n and λ. To estimate the area under the LROC curve, $\mathcal{A}_L$, Popescu [9] introduced the nonparametric estimator

$$\widehat{\mathcal{A}}_L = \frac{1}{mn} \sum_{j=1}^{L} \sum_{i=1}^{m} \mathcal{I}(Y_{l_j} > X_i) \tag{2}$$

where $\mathcal{I}(S) = 1$ if the proposition S is true and $\mathcal{I}(S) = 0$ otherwise. Popescu [9] gave an expression for the variance of $\widehat{\mathcal{A}}_L$ and showed that $\widehat{\mathcal{A}}_L$ is an unbiased estimator of $\mathcal{A}_L$, i.e., $\mathrm{E}[\widehat{\mathcal{A}}_L] = \mathcal{A}_L$. Additional investigations pertaining to the estimator in (2) were also performed by Tang *et al.* [22], [23].

Since U-statistics are a well-understood class of nonparametric estimators with nice optimality properties, it is desirable to estimate $\mathcal{A}_L$ with a U-statistic. Unfortunately, when $\widehat{\mathcal{A}}_L$ is written in the form of (2), it does not fit the definition of a U-statistic because

$$\mathrm{E}[\mathcal{I}(Y_{l_j} > X_i)] = \mathrm{E}[\mathcal{I}(Y > X) \mid \mathcal{L}] \tag{3}$$

$$= P(Y > X \mid \mathcal{L}) \neq \mathcal{A}_L \tag{4}$$

and because the summation over j in (2) involves a random number of terms, L, whereas U-statistics are only defined with fixed-length summations, i.e., sums that have a fixed (deterministic) number of terms.

However, we have found that $\widehat{\mathcal{A}}_L$ can be reformulated as a U-statistic, as explained in the following. For $j = 1, 2, \ldots, n$, let Q_j be equal to one if the lesion is correctly localized in the jth class-2 image and zero otherwise. The Q_j are assumed to be i.i.d., but note that Q_j and Y_j are generally dependent. With this definition for the localization results, $\widehat{\mathcal{A}}_L$ can be rewritten as

$$\widehat{\mathcal{A}}_L = \frac{1}{mn} \sum_{j=1}^{n} \sum_{i=1}^{m} \mathcal{I}(Y_j > X_i \,\&\, Q_j = 1). \tag{5}$$

Observe that the unbiasedness of $\widehat{\mathcal{A}}_L$ follows immediately from (5), since

$$\mathrm{E}[\widehat{\mathcal{A}}_L] = \mathrm{E}[\mathcal{I}(Y_j > X_i \,\&\, Q_j = 1)] \tag{6}$$

$$= P(Y > X \cap \mathcal{L}) = \mathcal{A}_L \tag{7}$$

where the last equality is due to Theorem 1. Because (5) involves summations of fixed length and $E[\mathcal{I}(Y_j > X_i \,\&\, Q_j = 1)] = \mathcal{A}_L$, the estimator $\widehat{\mathcal{A}}_L$ is a generalized U-statistic [36] with kernel $\phi(X_i, Y_j, Q_j) = \mathcal{I}(Y_j > X_i \,\&\, Q_j = 1)$, where Q_j is a Bernoulli random variable with probability of success $\lambda = P(\mathcal{L})$. More precisely, $\widehat{\mathcal{A}}_L$ is a two-sample U-statistic with a kernel that is a function of X_i and the vector $[Y_j, Q_j]$.

IV. Nonparametric Covariance Estimation

A. Estimator Definition

Suppose we wish to compare correlated LROC area estimates from q different LROC observer experiments. Denote the ratings and localization results for these experiments as X_i^k, Y_j^k, and Q_j^k for $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, q$, respectively. For ease of notation, let $\phi_{ij}^k = \phi(X_i^k, Y_j^k, Q_j^k) = \mathcal{I}(Y_j^k > X_i^k \,\&\, Q_j^k = 1)$. Also, denote the estimated area under the LROC curve for the kth experiment as $\widehat{\mathcal{A}}_L^k$, which takes the compact form

$$\widehat{\mathcal{A}}_L^k = \frac{1}{mn} \sum_{j=1}^{n} \sum_{i=1}^{m} \phi_{ij}^k. \tag{8}$$

In this section, we introduce an estimator for the covariance matrix of the vector $[\widehat{\mathcal{A}}_L^1, \widehat{\mathcal{A}}_L^2, \ldots, \widehat{\mathcal{A}}_L^q]^T$.

A formula for the covariance between two estimates $\widehat{\mathcal{A}}_L^u$ and $\widehat{\mathcal{A}}_L^v$ is given by the next theorem proved in Appendix B.

Theorem 2: Suppose that for experiment k, $\widehat{\mathcal{A}}_L^k$ is computed from m i.i.d. class-1 ratings and n i.i.d. class-2 ratings and localization results. Then the covariance between the LROC area estimates for experiments u and v is

$$\mathrm{Cov}[\widehat{\mathcal{A}}_L^u, \widehat{\mathcal{A}}_L^v] = \frac{(n-1)\zeta_{10}^{uv} + (m-1)\zeta_{01}^{uv} + \zeta_{11}^{uv}}{mn}$$

where $\zeta_{10}^{uv} = \mathrm{Cov}[\phi_{ij}^u, \phi_{ij'}^v]$ with $j \neq j'$, $\zeta_{01}^{uv} = \mathrm{Cov}[\phi_{ij}^u, \phi_{i'j}^v]$ with $i \neq i'$, and $\zeta_{11}^{uv} = \mathrm{Cov}[\phi_{ij}^u, \phi_{ij}^v]$.

Note that because the U-statistic kernel ϕ_{ij}^k is bounded above by 1 and below by 0, it follows that the covariances ζ_{10}^{uv}, ζ_{01}^{uv}, and ζ_{11}^{uv} all have values in the interval $[-1, 1]$. The covariance

formula in the previous theorem is similar to the covariance formula for the Mann-Whitney U statistic given by DeLong *et al.* [29] but with a different U-statistic kernel. As a special case, the proof in Appendix B also yields the Mann-Whitney result, which was not proved in [29].

Sen [28] introduced a very general nonparametric strategy that we use here to estimate ζ_{10}^{uv} and ζ_{01}^{uv}. First, define the "structural components"

$$V_{10}^{k}(i) = \frac{1}{n}\sum_{j=1}^{n}\phi_{ij}^{k}, \quad i = 1, 2, \ldots, m \tag{9}$$

$$V_{01}^{k}(j) = \frac{1}{m}\sum_{i=1}^{m}\phi_{ij}^{k}, \quad j = 1, 2, \ldots, n. \tag{10}$$

Observe that each structural component has mean $\mathcal{A}_L^k$. Also, it can be shown that $\mathrm{Cov}(V_{10}^u(i), V_{10}^v(i')) = \zeta_{01}^{uv}/n$ for $i \neq i'$ and $\mathrm{Cov}(V_{01}^u(j), V_{01}^v(j')) = \zeta_{10}^{uv}/m$ for $j \neq j'$. Hence, the structural components are asymptotically uncorrelated. Following the approach of Sen [28], [37, p. 82], we define

$$s_{10}^{uv} = \frac{1}{m-1}\sum_{i=1}^{m}[V_{10}^u(i) - \widehat{\mathcal{A}}_L^u][V_{10}^v(i) - \widehat{\mathcal{A}}_L^v] \tag{11}$$

$$s_{01}^{uv} = \frac{1}{n-1}\sum_{j=1}^{n}[V_{01}^u(j) - \widehat{\mathcal{A}}_L^u][V_{01}^v(j) - \widehat{\mathcal{A}}_L^v]. \tag{12}$$

The quantities s_{10}^{uv} and s_{01}^{uv} are the Sen estimators for ζ_{10}^{uv} and ζ_{01}^{uv}. Note that s_{10}^{uv} and s_{01}^{uv} are simply sample covariances computed from the structural components.

From Theorem 2, it is easy to see that for large m and n, $\mathrm{Cov}[\widehat{\mathcal{A}}_L^u, \widehat{\mathcal{A}}_L^v] \approx \zeta_{10}^{uv}/m + \zeta_{01}^{uv}/n$. Hence, the desired covariance between $\widehat{\mathcal{A}}_L^u$ and $\widehat{\mathcal{A}}_L^v$ can be estimated as

$$s^{uv} = \frac{s_{10}^{uv}}{m} + \frac{s_{01}^{uv}}{n}. \tag{13}$$

Let S, S_{10} and S_{01} be matrices with (u, v) entries given by s^{uv}, s_{10}^{uv} and s_{01}^{uv}, respectively, for $u, v \in \{1, 2, \ldots, q\}$. With this notation, the nonparametric Sen estimator for the $q \times q$ covariance matrix of the vector $[\widehat{\mathcal{A}}_L^1, \widehat{\mathcal{A}}_L^2, \ldots, \widehat{\mathcal{A}}_L^q]^T$ is

$$S = \frac{S_{10}}{m} + \frac{S_{01}}{n}. \tag{14}$$

From (11) and (12), it is clear that S_{10} and S_{01} can be written as sums of vector outer products. Hence, it follows that S_{10} and S_{01} are semipositive definite matrices and consequently, that S is semipositive definite.

In the context of ROC analysis, DeLong *et al.* [29] applied Sen's approach to develop a similar covariance estimator for the Mann-Whitney U statistic. The DeLong estimator is recovered from the above covariance estimator if the condition $Q_j^k = 1$ is removed from ϕ_{ij}^k. When $u = v$, the covariance estimator in (13) reduces to a variance estimator that is an alternative to the variance estimator of Popescu [9] for $\widehat{\mathcal{A}}_L$.

B. Theoretical Analysis

The bias of our covariance estimator, s^{uv}, is characterized by the following theorem that is proved in Appendix C.

Theorem 3: Let ζ_{10}^{uv}, ζ_{01}^{uv}, and ζ_{11}^{uv} be as in Theorem 2. Then

$$\mathrm{E}\left[s^{uv} - \mathrm{Cov}\left(\widehat{\mathcal{A}}_L^u, \widehat{\mathcal{A}}_L^v\right)\right] = \frac{1}{mn}\left[\zeta_{11}^{uv} - \zeta_{10}^{uv} - \zeta_{01}^{uv}\right].$$

The previous theorem implies that s^{uv} is asymptotically unbiased. Observe that the sign of this bias depends on the values of ζ_{10}^{uv}, ζ_{01}^{uv}, and ζ_{11}^{uv}, which are in the interval $[-1, 1]$.

A stronger characterization of the convergence of s^{uv} is provided by the next theorem. Its proof, which is given in Appendix D, generalizes an argument given by Sen [37, p. 80–82] for one-sample U-statistics to the present context.

Theorem 4: Suppose that $m/(m+n) = p_1 > 0$ and $n/(m+n) = p_2 > 0$, where p_1 and p_2 are fixed constants that are independent of $m + n$. Then as $m + n \to \infty$, the estimator s^{uv} converges with probability one to $\mathrm{Cov}[\widehat{\mathcal{A}}_L^u, \widehat{\mathcal{A}}_L^v]$, i.e.,

$$P\left(\lim_{m+n\to\infty} s^{uv} = \mathrm{Cov}[\widehat{\mathcal{A}}_L^u, \widehat{\mathcal{A}}_L^v]\right) = 1.$$

In the terminology of estimation theory, the previous theorem shows that s^{uv} is a *strongly consistent* estimator [38, p. 48] of $\mathrm{Cov}[\widehat{\mathcal{A}}_L^u, \widehat{\mathcal{A}}_L^v]$. The consistency of s^{uv} and a standard result regarding asymptotic normality for a vector of generalized U-statistics [39, Lemma 3.1] together yield the next theorem, which is proved in Appendix E.

Theorem 5: Let $\widehat{\boldsymbol{\mathcal{A}}}_L = [\widehat{\mathcal{A}}_L^1, \widehat{\mathcal{A}}_L^2, \ldots, \widehat{\mathcal{A}}_L^q]^T$ and $\boldsymbol{\mathcal{A}}_L = [\mathcal{A}_L^1, \mathcal{A}_L^2, \ldots, \mathcal{A}_L^q]^T$. Also, suppose that $m/(m+n) = p_1 > 0$ and $n/(m+n) = p_2 > 0$, where p_1 and p_2 are fixed constants that are independent of $m + n$. Then as $m + n \to \infty$, the random vector $S^{-1/2}(\widehat{\boldsymbol{\mathcal{A}}}_L - \boldsymbol{\mathcal{A}}_L)$ converges in distribution to multivariate normal vector with mean zero and identity covariance matrix.

As a corollary, note that any linear combination of $\widehat{\boldsymbol{\mathcal{A}}}_L = [\widehat{\mathcal{A}}_L^1, \widehat{\mathcal{A}}_L^2, \ldots, \widehat{\mathcal{A}}_L^q]^T$ is also asymptotically normal. Theorem 5 justifies the construction of confidence intervals or confidence regions with standard approaches based on asymptotic normality [20, p. 107],[40]. An example involving confidence intervals is given in Section VI-B.

V. Monte Carlo Evaluation

Supplementing the theoretical results of Section IV-B, we now present a Monte Carlo study of confidence intervals based on our covariance estimator. More specifically, coverage probabilities of confidence intervals for a difference of two $\mathcal{A}_L$ values are evaluated. This section begins with a description of the model used to generate random LROC ratings, followed by the Monte Carlo study results. In the following, if a $p \times 1$ random vector, $\mathbf{Z}$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, we write $\mathbf{Z} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$.

A. Ratings Model

To generate random, correlated ratings and localization results for two LROC observer experiments, we used a straightforward generalization of Swensson's binormal model for LROC data arising from a single LROC experiment [21]. In the explanation of this generalized Swensson model, note that we depart from Swensson's original notation [21].

Consider two LROC experiments, referred to as experiments A and B, respectively, and let superscripts A and B indicate

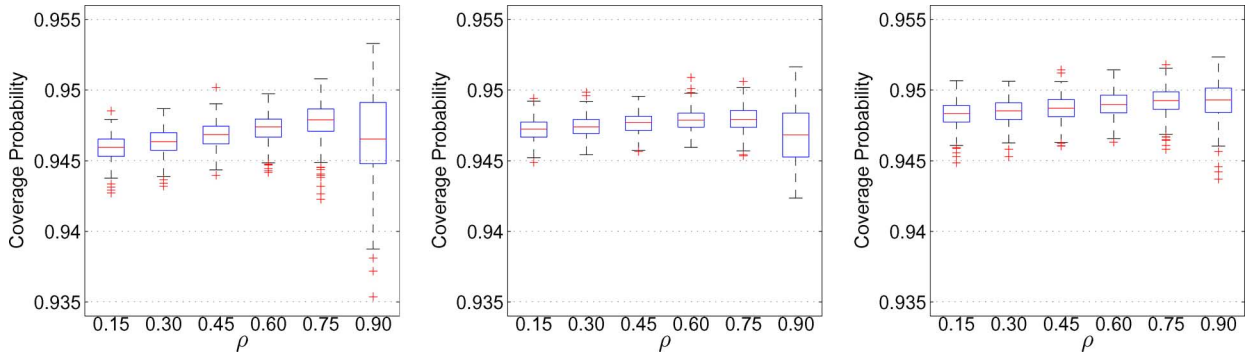

Fig. 2. Box plots of estimated coverage probability for approximate 95% confidence intervals for $\mathcal{A}_L^A - \mathcal{A}_L^B$. Each box plot summarizes coverage probability for 375 combinations of μ_A, μ_B, σ_A and σ_B. Left: $(m, n) = (25, 50)$. Center: $(m, n) = (75, 75)$. Right: $(m, n) = (50, 150)$.

quantities associated with these experiments. The class-1 and class-2 observer rating variables are collected in the vectors $\mathbf{X} = [X^A, X^B]^T$ and $\mathbf{Y} = [Y^A, Y^B]^T$, respectively, and the localization results are written as $\mathbf{Q} = [Q^A, Q^B]^T$.

For the generalized Swensson model, the class-1 ratings are assumed to follow a bivariate normal distribution with zero mean and covariance matrix, Σ_1, i.e., $\mathbf{X} \sim \mathcal{N}_2(\mathbf{0}, \Sigma_1)$, where

$$\Sigma_1 = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \tag{15}$$

and ρ is a correlation coefficient. To describe the class-2 ratings and localization results, it is necessary to first introduce two latent rating variables, $\mathbf{N} = [N^A, N^B]^T$ and $\mathbf{T} = [T^A, T^B]^T$. Here, $\mathbf{N}$ is interpreted as the maximum observer rating over all nonlesion locations and $\mathbf{T}$ is the observer rating for the lesion location. The latent variables $\mathbf{N}$ and $\mathbf{T}$ are assumed to be independent with $\mathbf{N} \sim \mathcal{N}_2(\mathbf{0}, \Sigma_1)$ and $\mathbf{T} \sim \mathcal{N}_2(\boldsymbol{\mu}, \Sigma_2)$, where $\boldsymbol{\mu} = [\mu^A, \mu^B]^T$ and

$$\Sigma_2 = \begin{bmatrix} (\sigma^A)^2 & \rho\,\sigma^A\sigma^B \\ \rho\,\sigma^A\sigma^B & (\sigma^B)^2 \end{bmatrix}. \tag{16}$$

The class-2 ratings are then defined as $Y^A = \max(N^A, T^A)$ and $Y^B = \max(N^B, T^B)$, respectively. Further, for a given image, the localization result Q^A is defined to be one if $T^A > N^A$ and zero otherwise. Similarly, Q^B is one if $T^B > N^B$ and zero otherwise.

The previous model thus describes the correlated LROC ratings for two experiments with five parameters: μ^A, μ^B, σ^A, σ^B and ρ. For the case of a single LROC experiment, the previous model reduces to the two-parameter Swensson binormal model [21]. Using formulas in Swensson's paper [21], it is straightforward to calculate the LROC areas, $\mathcal{A}_L^A$ and $\mathcal{A}_L^B$, and the localization probabilities, λ^A and λ^B, for the two experiments from μ^A, μ^B, σ^A, and σ^B.

B. Confidence Interval Validation

Given a random sample of observer ratings and localization results, an approximate, Wald-style 95% confidence interval for the difference of LROC areas, $\Delta\mathcal{A}_L = \mathcal{A}_L^A - \mathcal{A}_L^B$, was calculated as

$$\Delta\widehat{\mathcal{A}}_L \pm 1.96\sqrt{\mathrm{Var}(\Delta\widehat{\mathcal{A}}_L)} \tag{17}$$

where $\Delta\widehat{\mathcal{A}}_L = \widehat{\mathcal{A}}_L^A - \widehat{\mathcal{A}}_L^B$. To get $\mathrm{Var}(\Delta\widehat{\mathcal{A}}_L)$, the covariance estimator of Section IV-A was used to estimate each term in

$$\mathrm{Var}(\Delta\widehat{\mathcal{A}}_L) = \mathrm{Var}(\widehat{\mathcal{A}}_L^A) + \mathrm{Var}(\widehat{\mathcal{A}}_L^B) - 2\mathrm{Cov}(\widehat{\mathcal{A}}_L^A, \widehat{\mathcal{A}}_L^B). \tag{18}$$

For the Monte Carlo evaluation, random LROC ratings and localization results were generated with the generalized Swensson model described previously, for a large number parameter combinations. Specifically, for $(m, n) = (25, 50)$, $(m, n) = (75, 75)$, and $(m, n) = (50, 150)$, and for $\rho \in \{.15, .3, .45, .6, .75, .9\}$, we took $\mu^A \in \{.5, .75, 1, 1.25, 1.5\}$ with $\mu^B = \mu^A - \delta$, where $\delta \in \{.5, .4, .3, .2, .1\}$, and $\sigma^A \in \{.5, .75, 1, 1.25, 1.5\}$ with $\sigma^B = \gamma\sigma^A$, where $\gamma \in \{.8, 1, 1.2\}$. Thus, for three choices of m and n, and six choices of ρ, we evaluated the coverage probability for $5 \times 5 \times 5 \times 3 = 375$ different combinations of μ^A, μ^B, σ^A, and σ^B.

For each parameter combination, the coverage probability for the approximate 95% confidence interval in (17) was estimated from 100 000 Monte Carlo trials, so that each estimate of the coverage probability had a standard deviation of 0.0007. Fig. 2 contains box plots summarizing the estimated coverage probabilities for $(m, n) = (25, 50), (75, 75)$, and $(50, 150)$ for different values of ρ. The box plots were generated with the MATLAB command *boxplot*. To interpret the plots, note that the edges of each box are the 25% and 75% percentiles, and the horizontal line inside the box is the median. The length of each whisker is at most 1.5 times the distance between the 75% and 25% percentiles, and data points outside this range are plotted individually.

The plots in Fig. 2 indicate that the coverage probabilities are generally reliable, with better accuracy as the number of images increases. Moreover, even for the case with 75 images ($m = 25, n = 50$), the coverage probabilities are relatively close to 95%. The largest variations in the coverage probabilities were observed for $\rho = 0.90$. However, even in this case, the plots indicate that for each pair of m and n, the majority of the 375 parameter combinations yielded coverage probabilities within ± 0.005 of the desired 0.95.

VI. Application

The covariance matrix estimator defined by (14) can be applied to estimate confidence intervals (or regions) for a vector of figures of merit. In addition, it can be used together with the

results of a pilot study to estimate the sample size and number of observers required to attain a particular confidence interval size. These topics are discussed further and are illustrated with an example.

A. Confidence Intervals and Study Design

As before, suppose that we wish to compare correlated results from q LROC observer experiments. Denote the vector of (true) LROC areas corresponding to the q experiments as $\mathbf{c} = [\mathcal{A}_L^1, \mathcal{A}_L^2, \ldots, \mathcal{A}_L^q]^T$. Also, define F to be a $p \times q$ matrix so that the $p \times 1$ vector $\mathbf{d} = F\mathbf{c}$ contains the desired figures of merit that are to be used for inference. Now, let $\widehat{\mathbf{c}} = [\widehat{\mathcal{A}}_L^1, \widehat{\mathcal{A}}_L^2, \ldots, \widehat{\mathcal{A}}_L^q]^T$ be the estimate of $\mathbf{c}$ and let $\widehat{\mathbf{d}} = F\widehat{\mathbf{c}}$ be the estimate of $\mathbf{d}$. Estimating the covariance matrix of $\widehat{\mathbf{c}}$ with S from (14), we can estimate the covariance matrix for $\widehat{\mathbf{d}}$ as $W = FSF^T$. The diagonal entries of W are variance estimates for the corresponding components of $\widehat{\mathbf{d}}$. Confidence intervals (or regions) for the components of $\mathbf{d}$ can be constructed with these variance estimates using standard approaches based on asymptotic normality with or without logit transformation [20, p. 107], [40].

For purposes of study design, it is often advantageous to conduct a preliminary pilot study with a small number of images and observers. For example, suppose that an LROC study is to be executed to evaluate reader-averaged performance for t different imaging conditions with a fixed pool of r readers examining m class-1 images and n class-2 images, where r, m, and n need to be determined. Denote the $t \times 1$ vector of LROC area estimates for the ith reader as $\widehat{\mathbf{g}}_i = [\widehat{\mathcal{A}}_L^1, \widehat{\mathcal{A}}_L^2, \ldots, \widehat{\mathcal{A}}_L^t]^T$ and the vector of reader-averaged LROC area estimates as $\widehat{\mathbf{g}} = (1/r)\sum_{i=1}^r \widehat{\mathbf{g}}_i$. Also, let Γ_i and Γ be the $t \times t$ covariance matrices for $\widehat{\mathbf{g}}_i$ and $\widehat{\mathbf{g}}$, respectively. If Γ can be reliably estimated from a pilot study as a function of m, n, and r, then the values of m, n, and r required to achieve a desired statistical precision can be predicted.

In the case of a partially paired design, in which the image sets for distinct observers are independent (which is preferable for assessing reader-averaged performance), the covariance matrix for $\widehat{\mathbf{g}}$ takes the form

$$\Gamma = \frac{1}{r^2}\sum_{i=1}^{r}\Gamma_i = \left(\frac{1}{r}\right)\bar{\Gamma} \tag{19}$$

where $\bar{\Gamma} = (1/r)\sum_{i=1}^r \Gamma_i$ is the reader-averaged covariance matrix. Suppose that a pilot study is carried out in which w readers are each given t (generally correlated) sets of images, with each image set corresponding to one of the t imaging conditions. From the results of this pilot study, we can use (11) and (12) to compute the $t \times t$ matrices S_{10} and S_{01} for the jth observer, which we denote as $S_{10}^{(j)}$ and $S_{01}^{(j)}$, respectively. Applying (14) to estimate Γ_j as $G_j = S_{10}^{(j)}/m + S_{01}^{(j)}/n$, the reader-averaged covariance matrix, $\bar{\Gamma}$, can be estimated as $(1/w)\sum_{j=1}^w G_j$. Inserting this expression into (19), we see that for any values of m, n, and r, the covariance matrix Γ can be estimated with

$$G = \frac{1}{rw}\sum_{j=1}^{w}\left(\frac{S_{10}^{(j)}}{m} + \frac{S_{01}^{(j)}}{n}\right). \tag{20}$$

Confidence intervals for the desired figures of merit can be constructed from G and the reader-averaged LROC area estimates from the pilot study. Thus, the procedure can be used to predict the values of m, n, and r needed to achieve a desired confidence interval length.

If a fully paired design is to be used in which the readers all look at the same images, then detection performance will necessarily be correlated between readers. In this case, if the desired number of readers is known, then the estimator in (14) can still be utilized in conjunction with a pilot study to estimate the number of images, m and n, needed to achieve a given confidence interval size. Namely, a pilot study with t imaging conditions can be conducted with all r readers to be used for the fixed-reader study, and the full $rt \times rt$ matrices S_{10} and S_{01} can be estimated with (11) and (12). Hence, for any choice of m and n, (14) yields an $rt \times rt$ covariance matrix, S, that can be utilized together with reader-averaged LROC area estimates to estimate confidence interval sizes.

B. Example

To illustrate the utility of the LROC covariance estimator, we conducted a fixed-reader human observer LROC study to compare three filtered backprojection (FBP) image reconstruction algorithms for fan-beam CT [41], [42]:

1) full-scan direct FBP reconstruction;
2) short-scan direct FBP reconstruction (240°);
3) full-scan indirect FBP reconstruction using rebinning to parallel-beam geometry.

The full-scan indirect FBP was implemented with factor of two upsampling in the rebinning step so that resolution was matched near the center of the imaging field of view for all three algorithms. In the LROC evaluation, algorithms A and C were anticipated to yield comparable observer performance, whereas observer performance for algorithm B was expected to be worse due to the fact that it uses less CT data, and therefore produces noisier images.

Fan-beam data sets were simulated for the head phantom shown in Fig. 3, which has the same dimensions as the FORBILD head phantom, but with a uniform attenuation value (50 HU) for the region inside the skull. The CT data simulation was carried out as described in [43], including Poisson noise, modeling of finite X-ray source and detector sizes, and a bowtie filter model. The photon level was chosen to be 85 000 and the bowtie filter was designed for a circular water cylinder of radius 14 cm. Lesion-present sinograms were generated by inserting a 5-mm-diameter circular lesion with a random contrast in the range of [25, 35] HU at a random location inside the skull. Note that the lesion location was such that the lesion never intersected the skull. All reconstructions were produced on a 400×400 grid with pixel size $\Delta x = \Delta y = 0.7$ mm.

LROC training and testing was conducted with a MATLAB graphical user interface, a screen capture of which is shown in Fig. 3. The images were displayed with a grayscale window of $[-25, 125]$ HU, centered on the brain tissue of the head phantom. A continuous rating scale in the range [0, 100] was used, with higher values corresponding to more certainty that a lesion was present. In addition, a lesion was deemed to be correctly localized if the reader correctly marked its location within 10 pixels; this criterion was determined by inspecting plots of the readers' true localized fraction versus localization radius, as

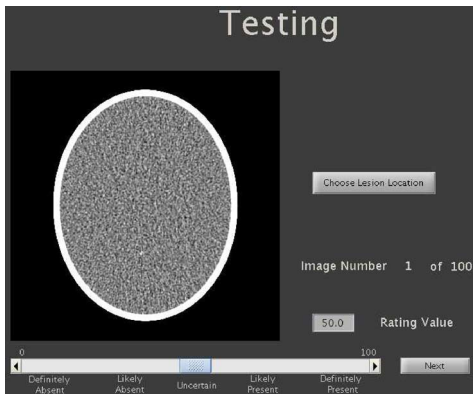

Fig. 3. Screen capture of observer study software. Grayscale range of displayed image is [−25, 125] HU.

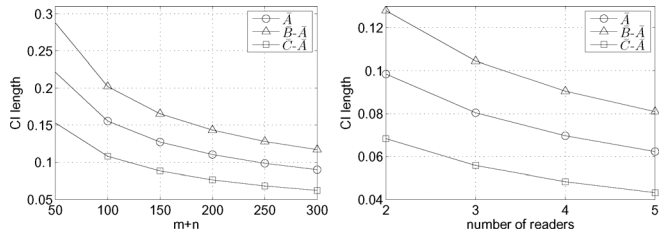

Fig. 4. Lengths of 98.33% confidence intervals predicted by pilot study. Left: confidence interval length for two readers plotted versus $m + n$ with $n = 2m$. Right: confidence interval length for $r = 2, 3, 4, 5$ readers with $m = 83$ and $n = 167$.

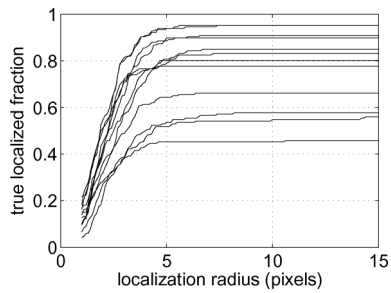

Fig. 5. Plots of true localized fraction versus localization radius for full-scale study. The twelve plots shown correspond to all combinations of the four readers and three reconstruction strategies.

in [14] and [15]. Plots of true localized fraction versus localization radius are provided later in this section for the full-scale study (see Fig. 5).

1) Pilot Study: Before conducting the full-scale LROC evaluation, a pilot study was performed with two readers. Each reader was shown 40 training images followed by 60 testing images for each reconstruction algorithm with a 5 min break between each of the three reading sessions. In both the training and testing sets, two-thirds of the images contained a lesion. The testing images in the pilot study were generated in a partially paired manner so that the images sets shown to distinct readers were statistically independent, but the images sets for each algorithm shown to the same reader were statistically dependent. Namely, two independent groups of 60 fan-beam data sets were created for testing, with each of the 60 fan-beam sinograms subsequently reconstructed with the three algorithms, yielding two independent groups of 3×60 images, i.e., one group for each reader.

Denote the LROC areas for algorithms A, B, and C and reader j as A_j, B_j and C_j, respectively. Also, define the reader-averaged LROC areas as $\bar{A} = (A_1 + A_2)/2$, $\bar{B} = (B_1 + B_2)/2$, and $\bar{C} = (C_1 + C_2)/2$. Letting $\mathbf{g} = [\bar{A}, \bar{B}, \bar{C}]^T$, our vector of figures of merit was chosen to be $\mathbf{d} = [\bar{A}, \bar{B} - \bar{A}, \bar{C} - \bar{A}]^T$, i.e., $\mathbf{d} = H\mathbf{g}$, where H is a 3×3 matrix.

The pilot study results were used to compute the 3×3 covariance matrix estimate, G, as given by (20), for various values of m, n, and r. Confidence intervals for the entries of $\mathbf{d}$ were subsequently estimated from the diagonal entries of HGH^T and $\widehat{\mathbf{d}}$, the estimate of $\mathbf{d}$ based on $\widehat{\mathcal{A}}_L$. More specifically, an asymmetric confidence interval for the first entry was estimated with the logit-transformation approach advocated by Pepe [20, p. 107], and symmetric confidence intervals for the second and third entries were estimated with the conventional Wald method [20, p. 107]. The coverage probability for each confidence interval was selected to be 98.33% so that the joint coverage probability for the three intervals together was at least 95% by the Bonferroni inequality.[2]

For the case of two readers, Fig. 4 (left) contains a plot of the estimated confidence interval lengths as a function of the total number of images, $m + n$, where $n = 2m$. From this plot, we see that even with 250 images per reconstruction method, the confidence interval length for $\bar{B} - \bar{A}$ was predicted to be greater than 0.1. Since it was desired to have all confidence interval lengths less than 0.1 with 250 images or less, more than two readers were evidently needed. Fig. 4 (right) plots the estimated confidence interval lengths for 250 images ($m = 83$ and $n = 167$) as a function of the number of readers, r. This plot predicted that four readers should yield confidence intervals less than 0.1 in length for all figures of merit.

2) Full-Scale Study: Based on the results of the pilot study, it was decided that the full-scale study would use four readers and 250 images per reconstruction method, with two-thirds of the images containing a lesion. Like the pilot study, the testing images in the full-scale study were generated in a partially paired manner so that the images sets shown to distinct readers were statistically independent, but the image sets for each algorithm shown to the same reader were statistically dependent. Specifically, four independent groups of 250 fan-beam data sets were created for testing, with each of the 250 fan-beam sinograms reconstructed with the three algorithms, yielding four independent groups of 3×250 images, i.e., one group for each reader. The image reading sessions for each observer were split over two days, where each day of reading consisted of three sessions, one reconstruction method per session. Each reading session was comprised of 40 training and 125 testing images, followed by a 10 min break.

Plots of true localized fraction versus localization radius for the full-scale study are shown in Fig. 5. As mentioned earlier in this section, a lesion was deemed to be correctly localized if it was marked within ten pixels, based on the observation that the true localized fraction plateaued for localization radii larger than ten pixels for all reader evaluations. Using this localization criterion, we estimated the area under the LROC curve for each reader and reconstruction algorithm combination. Table I

[2]For arbitrary events $E_1, E_2, \ldots, E_l$, the Bonferroni inequality [44, eq. (1.2.10), p. 13] takes the form $P\left(\bigcap_{i=1}^{l} E_i\right) \geq \sum_{i=1}^{l} P(E_i) - (l-1)$.

TABLE I
ESTIMATED $\mathcal{A}_L$ VALUES AND STANDARD DEVIATIONS (IN PARENTHESES) FOR EACH READER AND RECONSTRUCTION STRATEGY IN EXAMPLE

	recon A	recon B	recon C
reader 1	0.887 (0.022)	0.531 (0.036)	0.902 (0.021)
reader 2	0.694 (0.033)	0.371 (0.036)	0.682 (0.033)
reader 3	0.752 (0.031)	0.443 (0.036)	0.774 (0.030)
reader 4	0.834 (0.026)	0.493 (0.037)	0.860 (0.025)

lists each LROC area estimate, together with an estimate of its standard deviation, which was obtained with the method of Section IV-A.

Denote the vector of $\mathcal{A}_L$ values for each reader and reconstruction method as

$$\mathbf{c} = [A_1, B_1, C_1, A_2, B_2, C_2, A_3, B_3, C_3, A_4, B_4, C_4]^T \quad (21)$$

where the letters correspond to the reconstruction method, and the subscripts denote the reader number. Using this notation, define the reader-averaged $\mathcal{A}_L$ values for methods A, B, and C as $\bar{A} = (A_1 + A_2 + A_3 + A_4)/4$, $\bar{B} = (B_1 + B_2 + B_3 + B_4)/4$, and $\bar{C} = (C_1 + C_2 + C_3 + C_4)/4$, respectively. Confidence intervals were estimated for $\bar{A}$, $\bar{B} - \bar{A}$, and $\bar{C} - \bar{A}$ ($\bar{A}$ was included to provide a reference value). Let $\mathbf{d} = [\bar{A}, \bar{B} - \bar{A}, \bar{C} - \bar{A}]^T$, so that $\mathbf{d} = F\mathbf{c}$ where F is a 3×12 matrix.

Each entry in $\mathbf{c}$ was estimated with (5) to get the vector of estimated LROC areas, $\widehat{\mathbf{c}}$. Next, the 12×12 covariance matrix for $\widehat{\mathbf{c}}$ was estimated using (14). Because the study was designed with independent image sets for distinct readers, the covariance matrix for $\hat{\mathbf{c}}$ had a block-diagonal structure, with four 3×3 blocks on the diagonal. Therefore, the off-diagonal blocks of the covariance matrix estimate, S, were replaced with zeros to remove unnecessary statistical variability. (Note that since each estimated block was semipositive definite, the resulting block diagonal estimate was also guaranteed to be semipositive definite.) Confidence intervals for the entries of $\mathbf{d}$ were estimated from $\widehat{\mathbf{d}} = F\widehat{\mathbf{c}}$ and the diagonal elements of $W = FSF^T$. More specifically, an asymmetric confidence interval for $\bar{A}$ was estimated with the logit transformation approach described by Pepe [20, p. 107], and symmetric confidence intervals for $\bar{B} - \bar{A}$ and $\bar{C} - \bar{A}$ were estimated using the conventional Wald method [20, p. 107]. As in the pilot study, the coverage probability for each confidence interval was selected to be 98.33% so that the joint coverage probability for the three intervals together was at least 95% by the Bonferroni inequality.

The estimated confidence intervals for $\bar{A}$, $\bar{B} - \bar{A}$, and $\bar{C} - \bar{A}$ are shown in Fig. 6. The numerical values of the 98.33% confidence intervals were $[0.752, 0.823]$ for $\bar{A}$, $[-0.380, -0.285]$ for $\bar{B} - \bar{A}$, and $[-0.016, 0.042]$ for $\bar{C} - \bar{A}$. Note that the confidence intervals all had lengths less than 0.1, which is consistent with the prediction of the pilot study. Examining the results, we see that the reader-averaged performance for method B was worse than method A with statistical significance. However, no statistically significant difference was observed between methods A and C. These observations are in agreement with the aforementioned expectations.

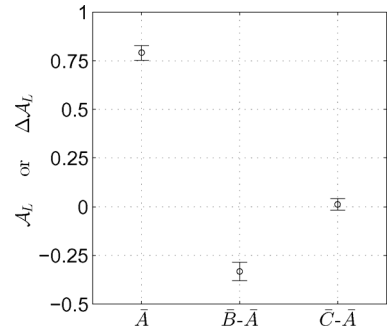

Fig. 6. Confidence intervals for reader-averaged performance in the example. Circles denote point estimates of the figures of merit. Each confidence interval has a coverage probability of 98.33% so that the joint coverage probability of all three intervals is at least 95%.

VII. DISCUSSION AND CONCLUSION

We introduced a nonparametric strategy to compare the areas under correlated LROC curves. Our approach relied on reformulating an LROC area estimator of Popescu [9] as a generalized U-statistic so that we could apply the theory of Sen [28] to construct a covariance estimator. Furthermore, additional results of Sen [37], [45] were applied to prove that the covariance estimator is strongly consistent. The application of our covariance estimator to confidence interval estimation was supported with a theorem on asymptotic normality and with a Monte Carlo simulation study. Last, our methodology was illustrated with a human observer study comparing lesion detectability for three fan-beam CT reconstruction algorithms.

The present work may be viewed as a generalization of the covariance estimator of DeLong *et al.* [29] developed for ROC analysis. Our proof of strong consistency implies the strong consistency of DeLong's estimator as a special case. (Note that strong consistency was not proved in [29].) Like DeLong's method, our covariance estimator is applicable to the analysis of performance estimates for any fixed set of readers. As mentioned in Section I, studies with a fixed reader pool are best suited for early stage evaluations, i.e., when extensive studies with large numbers of readers and images are not practical. Nevertheless, extension of our approach to random-reader inference is of high interest and will be investigated in the future. Results in [46]–[48] may prove useful in the derivation of such an extension.

For the special case of a single LROC area estimate, the covariance estimator in (13) with $u = v$ reduces to a variance estimator that may be seen as an alternative to the variance estimator of Popescu [9] for $\widehat{\mathcal{A}}_L$. In a limited set of Monte Carlo simulation studies, we have observed that the variance estimator defined by (13) performs similarly to that of Popescu [9]. However, unlike the variance estimator of Popescu [9], the estimator in (13) is known to be strongly consistent.

In addition to strong consistency, our covariance matrix estimator has the satisfying property that it is semipositive definite, which guarantees that any variance estimate obtained as described in Section VI-A will be nonnegative. Moreover, this property implies that both rectangular and ellipsoidal confidence regions can be reliably constructed for a vector of figures of merit. Additional positive features of our covariance estimator are its conceptual simplicity and computational efficiency, which together enable straightforward sample size predictions for purposes of study design. Although not discussed here, our estimator can also be used to estimate the optimal ratio of lesion-absent to lesion-present images that minimizes variability in an LROC study. We hope to report on

this issue in the future, thereby generalizing a result of [9] for a single LROC curve.

Last, as mentioned in Section I, the estimation theory presented here for continuous-valued ratings also applies to ordinal ratings with only minor modifications. The extension of our approach to ordinal rating data is given in Appendix A.

Appendix A
Extension to Ordinal Ratings

In Section II, the LROC curve was defined as the plot of TPLF(c) versus FPF(c), where TPLF(c) $= P(Y \geq c \cap \mathcal{L})$ and FPF(c) $= P(X \geq c)$. For continuous-valued ratings, X and Y are continuous random variables and the threshold c takes values in the range $(-\infty, \infty)$. Instead, if a finite number of ordinal rating categories are used, then X and Y are discrete random variables and $c \in \mathcal{D}$, where $\mathcal{D}$ is a finite subset of the real line. In this case, the LROC "curve" is a discrete function, consisting of a finite set of points. The discrete LROC function for ordinal data is analogous to the discrete ROC function, which is discussed in [49] and [20, Sec. 4.5.5]. As in the ROC case [20], it is possible to define a continuous LROC curve for ordinal ratings by assuming a semiparametric model like that of Swensson [21]. Alternatively, motivated by the following theorem, which is closely related to Theorem 1, one can define a figure of merit for discrete LROC functions without making such assumptions. This theorem is analogous to a well-known result for discrete ROC functions [20, Result 4.10, p. 92], [49].

Theorem 6: Suppose that X and Y are independent, discrete random variables, and that $\mathcal{L}$ is independent of any event defined only with X. Then the area under the curve formed by linearly interpolating the discrete LROC function is $\mathcal{B}_L = P(Y > X \cap \mathcal{L}) + (1/2)P(X = Y \cap \mathcal{L})$.

Proof: The proof is essentially the same as that of Bamber [49] for discrete ROC functions and is therefore omitted. ■

Note that $\mathcal{B}_L = \mathcal{A}_L$ when X and Y are continuous random variables, since in that case, $P(X = Y) = 0$. Motivated by its close relationship to the area under the continuous LROC curve, $\mathcal{B}_L$ can be adopted as a figure of merit for discrete LROC functions.

For the case of ordinal ratings, we define an unbiased estimator of $\mathcal{B}_L$ for the kth LROC experiment as

$$\widehat{\mathcal{B}}_L^k = \frac{1}{mn} \sum_{j=1}^{n} \sum_{i=1}^{m} \varphi_{ij}^k \tag{22}$$

where $\varphi_{ij}^k = \mathcal{I}(Y_j > X_i \;\&\; Q_j = 1) + (1/2)\mathcal{I}(X_i = Y_j \;\&\; Q_j = 1)$. It is straightforward to see that the results of Sections III and IV for $\widehat{\mathcal{A}}_L^k$ also apply to $\widehat{\mathcal{B}}_L^k$, with ϕ_{ij}^k replaced by φ_{ij}^k. Therefore, ordinal ratings can be handled with only minor modifications to our approach. (A similar observation was made by Popescu [9] for his estimators.)

Appendix B
Proof of Theorem 2

From (8) and the linearity of the expectation operator

$$\mathrm{E}(\widehat{\mathcal{A}}_L^u \widehat{\mathcal{A}}_L^v) = \frac{1}{m^2 n^2} \sum_{j=1}^{n} \sum_{i=1}^{m} \sum_{j'=1}^{n} \sum_{i'=1}^{m} \mathrm{E}(\phi_{ij}^u \phi_{i'j'}^v). \tag{23}$$

Let δ_{ij} denote the Kronecker delta function for which $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. Since

$$\begin{aligned}\delta_{ii'}\delta_{jj'} + (1-\delta_{ii'})\delta_{jj'} + \delta_{ii'}(1-\delta_{jj'}) + (1-\delta_{ii'})(1-\delta_{jj'}) \\ = (\delta_{ii'} + 1 - \delta_{ii'})(\delta_{jj'} + 1 - \delta_{jj'}) = 1\end{aligned}$$

it follows that

$$\begin{aligned}\mathrm{E}(\widehat{\mathcal{A}}_L^u \widehat{\mathcal{A}}_L^v) = \frac{1}{m^2 n^2} \sum_{j=1}^{n} \sum_{i=1}^{m} \sum_{j'=1}^{n} \sum_{i'=1}^{m} \mathrm{E}(\phi_{ij}^u \phi_{i'j'}^v)[\delta_{ii'}\delta_{jj'} \\ + (1-\delta_{ii'})\delta_{jj'} + \delta_{ii'}(1-\delta_{jj'}) \\ + (1-\delta_{ii'})(1-\delta_{jj'})]\end{aligned} \tag{24}$$

which implies

$$\begin{aligned}\mathrm{E}(\widehat{\mathcal{A}}_L^u \widehat{\mathcal{A}}_L^v) = \frac{1}{m^2 n^2} \sum_{j=1}^{n} \sum_{i=1}^{m} \Bigg[\mathrm{E}(\phi_{ij}^u \phi_{ij}^v) + \sum_{\substack{i'=1 \\ i' \neq i}}^{m} \mathrm{E}(\phi_{ij}^u \phi_{i'j}^v) \\ + \sum_{\substack{j'=1 \\ j' \neq j}}^{n} \mathrm{E}(\phi_{ij}^u \phi_{ij'}^v) + \sum_{\substack{j'=1 \\ j' \neq j}}^{n} \sum_{\substack{i'=1 \\ i' \neq i}}^{m} \mathrm{E}(\phi_{ij}^u \phi_{i'j'}^v) \Bigg].\end{aligned} \tag{25}$$

Using the definitions of ζ_{10}^{uv}, ζ_{01}^{uv}, and ζ_{11}^{uv} and the fact that $\mathrm{E}(\phi_{ij}^r) = \mathcal{A}_L^r$ yields

$$\begin{aligned}\mathrm{E}(\widehat{\mathcal{A}}_L^u \widehat{\mathcal{A}}_L^v) &= \frac{1}{mn}[\zeta_{11}^{uv} + \mathcal{A}_L^u \mathcal{A}_L^v + (m-1)(\zeta_{01}^{uv} + \mathcal{A}_L^u \mathcal{A}_L^v) \\ &\quad + (n-1)(\zeta_{10}^{uv} + \mathcal{A}_L^u \mathcal{A}_L^v) + (m-1)(n-1)\mathcal{A}_L^u \mathcal{A}_L^v] \qquad (26) \\ &= \frac{1}{mn}[\zeta_{11}^{uv} + (m-1)\zeta_{01}^{uv} + (n-1)\zeta_{10}^{uv}] + \mathcal{A}_L^u \mathcal{A}_L^v. \qquad (27)\end{aligned}$$

The desired covariance formula follows from (27).

Appendix C
Proof of Theorem 3

The proof of Theorem 3 relies on the following two lemmas. The definitions of ζ_{10}^{uv}, ζ_{01}^{uv}, and ζ_{11}^{uv} can be found in the statement of Theorem 2.

Lemma 1: Suppose that $V_{10}^k(i)$ and $V_{01}^k(j)$ are computed from m i.i.d. class-1 ratings and n i.i.d. class-2 ratings and localization results from the kth LROC experiment. Then for fixed i and j

(a)

$$\mathrm{E}[V_{10}^u(i) V_{10}^v(i)] = \frac{(n-1)\zeta_{10}^{uv} + \zeta_{11}^{uv}}{n} + \mathcal{A}_L^u \mathcal{A}_L^v.$$

(b)

$$\mathrm{E}[V_{01}^u(j) V_{01}^v(j)] = \frac{(m-1)\zeta_{01}^{uv} + \zeta_{11}^{uv}}{m} + \mathcal{A}_L^u \mathcal{A}_L^v.$$

Proof: We prove part (a). The proof of part (b) is similar

$$\mathrm{E}[V_{10}^u(i) V_{10}^v(i)] = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{j'=1}^{n} \mathrm{E}(\phi_{ij}^u \phi_{ij'}^v) \tag{28}$$

$$= \frac{1}{n^2}\sum_{j=1}^{n}\left[\mathrm{E}(\phi_{ij}^u\phi_{ij}^v) + \sum_{\substack{j'=1\\ j'\neq j}}^{n}\mathrm{E}(\phi_{ij}^u\phi_{ij'}^v)\right] \quad (29)$$

$$= \frac{1}{n}[\zeta_{11} + \mathcal{A}_L^u\mathcal{A}_L^v + (n-1)(\zeta_{10}^{uv} + \mathcal{A}_L^u\mathcal{A}_L^v)] \quad (30)$$

$$= \frac{1}{n}[(n-1)\zeta_{10}^{uv} + \zeta_{11}^{uv}] + \mathcal{A}_L^u\mathcal{A}_L^v. \quad (31)$$

■

Lemma 2: Suppose that for experiment k,$\widehat{\mathcal{A}}_L^k$, $V_{10}^k(i)$ and $V_{01}^k(j)$ are computed from m i.i.d. class-1 ratings and n i.i.d. class-2 ratings and localization results. Then for fixed i and j

$$\mathrm{E}[\widehat{\mathcal{A}}_L^u(i)V_{10}^v(i)] = \mathrm{E}[\widehat{\mathcal{A}}_L^u(i)V_{01}^v(j)] = \frac{(n-1)\zeta_{10}^{uv} + (m-1)\zeta_{01}^{uv} + \zeta_{11}^{uv}}{mn} + \mathcal{A}_L^u\mathcal{A}_L^v.$$

Proof: The proof is analogous to that for Theorem 2 and is therefore omitted. ■

Next, observe that

$$\mathrm{E}[s_{10}^{uv}] = \frac{1}{m-1}\sum_{i=1}^{m}\mathrm{E}[(V_{10}^u(i) - \widehat{\mathcal{A}}_L^u)(V_{10}^v(i) - \widehat{\mathcal{A}}_L^v)] \quad (32)$$

$$= \frac{1}{m-1}\sum_{i=1}^{m}[\mathrm{E}(V_{10}^u(i)V_{10}^v(i)) - \mathrm{E}(\widehat{\mathcal{A}}_L^uV_{10}^v(i)) - \mathrm{E}(V_{10}^u(i)\widehat{\mathcal{A}}_L^v) + \mathrm{E}(\widehat{\mathcal{A}}_L^u\widehat{\mathcal{A}}_L^v)]. \quad (33)$$

Using Lemma 1(a), Lemma 2, and Theorem 2 together yields

$$\mathrm{E}[s_{10}^{uv}] = \frac{1}{n}\left[(n-1)\zeta_{10}^{uv} - \zeta_{01}^{uv} + \zeta_{11}^{uv}\right]. \quad (34)$$

Similarly, it is straightforward to show that

$$\mathrm{E}[s_{01}^{uv}] = \frac{1}{m}\left[(m-1)\zeta_{01}^{uv} - \zeta_{10}^{uv} + \zeta_{11}^{uv}\right]. \quad (35)$$

Finally, applying (13), (34), and (35) together with Theorem 2 gives the stated result for Theorem 3.

Appendix D
Proof of Theorem 4

Recall that a sequence of random variables $Z_1, Z_2, \ldots,$ is said to converge with probability one to a limit, Z, if and only if $P(\lim_{i\to\infty} Z_i = Z) = 1$ [38, p. 6]. In this case, we write $Z_i \xrightarrow{wp1} Z$ as $i \to \infty$. Our proof of Theorem 4 depends on the following lemma.

Lemma 3: As $m \to \infty$ and $n \to \infty$

$$s_{10}^{uv} \xrightarrow{wp1} \zeta_{10}^{uv} \qquad \text{and} \qquad s_{01}^{uv} \xrightarrow{wp1} \zeta_{01}^{uv}.$$

Proof: The convergence proof for s_{10}^{uv} is given as follows. For the proof, we apply an argument introduced by Sen [37, p. 80–82] in the context of one-sample U-statistics to our two-sample setting. The convergence proof for s_{01}^{uv} follows similar steps, and is therefore omitted.

By definition, s_{10}^{uv} is such that

$$(m-1)\,s_{10}^{uv} = \sum_{i=1}^{m}[V_{10}^u(i) - \widehat{\mathcal{A}}_L^u][V_{10}^v(i) - \widehat{\mathcal{A}}_L^v] = \sum_{i=1}^{m}[V_{10}^u(i)V_{10}^v(i) - \widehat{\mathcal{A}}_L^uV_{10}^v(i) - V_{10}^u(i)\widehat{\mathcal{A}}_L^v + \widehat{\mathcal{A}}_L^u\widehat{\mathcal{A}}_L^v].$$

Since $\widehat{\mathcal{A}}_L^r = (1/m)\sum_{i=1}^{m}V_{10}^r(i)$, the previous equation becomes

$$s_{10}^{uv} = \frac{1}{m-1}\sum_{i=1}^{m}V_{10}^u(i)V_{10}^v(i) - \left(\frac{m}{m-1}\right)\widehat{\mathcal{A}}_L^u\widehat{\mathcal{A}}_L^v. \quad (36)$$

The first term can be rewritten as

$$\frac{1}{m-1}\sum_{i=1}^{m}V_{10}^u(i)V_{10}^v(i) = \frac{1}{(m-1)n^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{j'=1}^{n}\phi_{ij}^u\phi_{ij'}^v = \frac{1}{(m-1)n^2}\left[\sum_{i=1}^{m}\sum_{j=1}^{n}\phi_{ij}^u\phi_{ij}^v + \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\phi_{ij}^u\phi_{ij'}^v\right] = \frac{m}{(m-1)n}U_{11} + \frac{m(n-1)}{(m-1)n}U_{10} \quad (37)$$

where

$$U_{11} = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m}\phi_{ij}^u\phi_{ij}^v \quad (38)$$

$$U_{10} = \frac{1}{mn(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j'\neq j}}^{n}\phi_{ij}^u\phi_{ij'}^v. \quad (39)$$

To apply Sen's convergence argument, we need to interpret U_{11} and U_{10} as U-statistics. Recall that a generalized two-sample U-statistic has a kernel that is a function of two sets of arguments, where the kernel is symmetric with respect to elements of each set, although the roles of the two sets need not be symmetric [36, p. 64–65]. Define the vectors $\mathbf{X}_i = [X_i^u, X_i^v]^T$ and $\mathbf{Y}_j = [Y_j^u, Q_j^u, Y_j^v, Q_j^v]^T$. Then U_{11} can be interpreted as a two-sample U-statistic with the (trivially) symmetric kernel $\Phi(\mathbf{X}_i, \mathbf{Y}_j) = \phi_{ij}^u\phi_{ij}^v$ when written as

$$U_{11} = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m}\Phi(\mathbf{X}_i, \mathbf{Y}_j). \quad (40)$$

For U_{10}, define the kernel $\Psi(\mathbf{X}_i, \mathbf{Y}_j, \mathbf{Y}_{j'}) = (\phi_{ij}^u\phi_{ij'}^v + \phi_{ij'}^u\phi_{ij}^v)/2$. Note that $\Psi(\mathbf{X}_i, \mathbf{Y}_j, \mathbf{Y}_{j'}) = \Psi(\mathbf{X}_i, \mathbf{Y}_{j'}, \mathbf{Y}_j)$, so Ψ is a symmetric. Hence, U_{10} can be interpreted as a two sample U-statistic when written in the form

$$U_{10} = \frac{2}{mn(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{j'=j+1}^{n}\Psi(\mathbf{X}_i, \mathbf{Y}_j, \mathbf{Y}_{j'}). \quad (41)$$

Furthermore, observe that $\mathrm{E}[U_{11}] = \mathrm{E}[\Phi(\mathbf{X}_i, \mathbf{Y}_j)] = \zeta_{11}^{uv} + \mathcal{A}_L^u \mathcal{A}_L^v$ and $\mathrm{E}[U_{10}] = \mathrm{E}[\Psi(\mathbf{X}_i, \mathbf{Y}_j, \mathbf{Y}_{j'})] = \zeta_{10}^{uv} + \mathcal{A}_L^u \mathcal{A}_L^v$.

Because the kernels $\Phi(\mathbf{X}_i, \mathbf{Y}_j)$ and $\Psi(\mathbf{X}_i, \mathbf{Y}_j, \mathbf{Y}_{j'})$ are bounded above by 1, it follows that $\mathrm{E}[\Phi(\mathbf{X}_i, \mathbf{Y}_j)^2] < \infty$ and $\mathrm{E}[\Psi(\mathbf{X}_i, \mathbf{Y}_j, \mathbf{Y}_{j'})^2] < \infty$. Therefore, results of Sen [45] on generalized U-statistics imply that U_{11} and U_{10} converge with probability one to their means, i.e., as $m, n \to \infty$

$$U_{11} \xrightarrow{wp1} \zeta_{11}^{uv} + \mathcal{A}_L^u \mathcal{A}_L^v \quad \text{and} \quad U_{10} \xrightarrow{wp1} \zeta_{10}^{uv} + \mathcal{A}_L^u \mathcal{A}_L^v. \tag{42}$$

Hence, $U_{11}/n \xrightarrow{wp1} 0$ as $m, n \to \infty$, and (37) and (42) thus imply that as $m, n \to \infty$

$$\frac{1}{m-1} \sum_{i=1}^{m} V_{10}^u(i) V_{10}^v(i) \xrightarrow{wp1} \zeta_{10}^{uv} + \mathcal{A}_L^u \mathcal{A}_L^v. \tag{43}$$

Also, since $\mathrm{E}[(\phi_{ij}^r)^2] = \mathcal{A}_L^r < \infty$, it follows from [45] that as $m, n \to \infty$, $\widehat{\mathcal{A}}_L^u \xrightarrow{wp1} \mathcal{A}_L^u$ and $\widehat{\mathcal{A}}_L^v \xrightarrow{wp1} \mathcal{A}_L^v$. These results with (36), (43), and the continuous transformation property for sequences that converge with probability one [38, p. 24], allow us to conclude that $s_{10}^{uv} \xrightarrow{wp1} \zeta_{10}^{uv}$ as $m, n \to \infty$. ■

Recall that $p_1 = m/(m+n)$ and $p_2 = n/(m+n)$ are fixed positive constants. Lemma 3 and the continuous transformation property for sequences that converge with probability one [38, p. 24] imply that

$$\frac{s_{10}^{uv}}{p_1} + \frac{s_{01}^{uv}}{p_2} \xrightarrow{wp1} \frac{\zeta_{10}^{uv}}{p_1} + \frac{\zeta_{01}^{uv}}{p_2} \quad \text{as} \quad m+n \to \infty. \tag{44}$$

Using the expression for $\mathrm{Cov}[\widehat{\mathcal{A}}_L^u, \widehat{\mathcal{A}}_L^v]$ in Theorem 2, it is then straightforward to show that as $m+n \to \infty$

$$\frac{s_{10}^{uv}}{p_1} + \frac{s_{01}^{uv}}{p_2} \xrightarrow{wp1} (m+n)\mathrm{Cov}[\widehat{\mathcal{A}}_L^u, \widehat{\mathcal{A}}_L^v] \tag{45}$$

and hence, $s^{uv} \xrightarrow{wp1} \mathrm{Cov}[\widehat{\mathcal{A}}_L^u, \widehat{\mathcal{A}}_L^v] \quad \text{as} \quad m+n \to \infty$.

Appendix E
Proof of Theorem 5

In the following, $\mathbf{Z}_i \xrightarrow{d} \mathbf{Z}$ denotes a sequence of random vectors $\mathbf{Z}_1, \mathbf{Z}_2, \ldots$ that converges in distribution to a limit $\mathbf{Z}$, as $i \to \infty$ [38, p. 8]. Following standard convention, if $\mathbf{Z}_i \xrightarrow{d} \mathbf{Z}$ with $\mathbf{Z} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$, we write $\mathbf{Z}_i \xrightarrow{d} \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$.

First, recall that $\widehat{\boldsymbol{\mathcal{A}}}_L = [\widehat{\mathcal{A}}_L^1, \widehat{\mathcal{A}}_L^2, \ldots, \widehat{\mathcal{A}}_L^q]^T$, $\boldsymbol{\mathcal{A}}_L = [\mathcal{A}_L^1, \mathcal{A}_L^2, \ldots, \mathcal{A}_L^q]^T$, $p_1 = m/(m+n)$ and $p_2 = n/(m+n)$, where p_1 and p_2 are fixed positive constants. Also, let Ω be the $q \times q$ matrix with (u, v) entry $\Omega_{uv} = \zeta_{10}^{uv}/p_1 + \zeta_{01}^{uv}/p_2$. Our proof for Theorem 5 relies on the following lemma.

Lemma 4: $\sqrt{m+n}(\widehat{\boldsymbol{\mathcal{A}}}_L - \boldsymbol{\mathcal{A}}_L) \xrightarrow{d} \mathcal{N}_q(\mathbf{0}, \Omega)$ as $m+n \to \infty$.

Proof: Since $\widehat{\boldsymbol{\mathcal{A}}}_L$ is a vector of two-sample U-statistics, the statement results from a standard theorem on generalized U-statistics [39, Lemma 3.1],[50, pp. 142–143]. ■

Now, from (44), we have $(m+n)S \xrightarrow{wp1} \Omega \quad \text{as} \quad m+n \to \infty$. Using the fact that convergence with probability one implies convergence in distribution [38], together with the continuous transformation property for sequences that converge in distribution [38, p. 24], we obtain

$$(m+n)^{-1/2} S^{-1/2} \Omega^{1/2} \xrightarrow{d} I \quad \text{as} \quad m+n \to \infty \tag{46}$$

where I is the $q \times q$ identity matrix. Also, the continuous transformation property for convergence in distribution [38, p. 24] and Lemma 4 imply that as $m+n \to \infty$

$$(m+n)^{1/2} \Omega^{-1/2} (\widehat{\boldsymbol{\mathcal{A}}}_L - \boldsymbol{\mathcal{A}}_L) \xrightarrow{d} \mathcal{N}_q(\mathbf{0}, I). \tag{47}$$

Again, using the continuous transformation property for sequences that converge in distribution [38, p. 24] with (46) and (47), we find that

$$\left[(m+n)^{-1/2} S^{-1/2} \Omega^{1/2}\right] \left[(m+n)^{1/2} \Omega^{-1/2} (\widehat{\boldsymbol{\mathcal{A}}}_L - \boldsymbol{\mathcal{A}}_L)\right] \xrightarrow{d} \mathcal{N}_q(\mathbf{0}, I) \quad \text{as} \quad m+n \to \infty$$

i.e., $S^{-1/2} (\widehat{\boldsymbol{\mathcal{A}}}_L - \boldsymbol{\mathcal{A}}_L) \xrightarrow{d} \mathcal{N}_q(\mathbf{0}, I) \quad \text{as} \quad m+n \to \infty$.

Acknowledgment

The authors would like to thank D. Heuscher and Z. Yu for acting as image readers for the full-scale LROC example. The contents of this paper are solely the responsibility of the authors.

References

[1] H. H. Barrett and K. J. Myers, *Foundations of Image Science*. New York: Wiley, 2004.
[2] R. F. Wagner, C. E. Metz, and G. Campbell, "Assessment of medical imaging systems and computer aids: A tutorial review," *Acad. Radiol.*, vol. 14, no. 6, pp. 723–748, Jun. 2007.
[3] C. E. Metz, "ROC analysis in medical imaging: A tutorial review of the literature," *Radiol. Phys. Technol.*, vol. 1, no. 1, pp. 2–12, Jan. 2008.
[4] J. Shiraishi, L. L. Pesce, C. E. Metz, and K. Doi, "Experimental design and data analysis in receiver operating characteristic studies: Lessons learned from reports in *Radiology* from 1997 to 2006," *Radiology*, vol. 253, no. 3, pp. 822–830, Dec. 2009.
[5] S. J. Starr, C. E. Metz, L. B. Lusted, and D. J. Goodenough, "Visual detection and localization of radiographic images," *Radiology*, vol. 116, no. 3, pp. 533–538, Sep. 1975.
[6] P. Bunch, J. Hamilton, G. Sanderson, and A. Simmons, "Free-response approach to the measurement and characterization of radiographic observer performance," *J. Appl. Photogr. Eng.*, vol. 4, no. 4, pp. 166–171, 1978.
[7] K. H. Zou, A. Liu, A. I. Bandos, L. Ohno-Machado, and H. E. Rockette, *Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis*. Boca Raton, FL: CRC Press, 2011.
[8] L. M. Popescu, "Nonparametric signal detectability evaluation using an exponential transformation of the FROC curve," *Med. Phys.*, vol. 38, no. 10, pp. 5690–5702, Oct. 2011.
[9] L. M. Popescu, "Nonparametric ROC and LROC analysis," *Med. Phys.*, vol. 34, no. 5, pp. 1556–1564, May 2007.
[10] Q. Li, F. Li, and K. Doi, "Computerized detection of lung nodules in thin-section CT images by use of selective enhancement filters and an automated rule-based classifier," *Acad. Radiol.*, vol. 15, pp. 165–175, 2008.
[11] M. F. McEntee and S. Dunnion, "A FROC analysis of radiographers performance in identification of distal radial fractures," *Eur. J. Radiol.*, vol. 1, pp. 90–94, 2009.
[12] V. A. Fisichella, M. Bath, A. A. Johnsson, F. Jaderling, T. Bergsten, U. Persson, K. Mellingen, and M. Hellstrom, "Evaluation of image quality and lesion perception by human readers on 3D CT colonography: Comparison of standard and low radiation dose," *Eur. Radiol.*, vol. 20, pp. 630–639, 2010.

[13] D. Gur, A. I. Bandos, H. E. Rockette, M. L. Zuley, J. H. Sumkin, D. M. Chough, and M. H. Christiane, "Localized detection and classification of abnormalities on FFDM and tomosynthesis examinations rated under an FROC paradigm," *Amer. J. Roentgenol.*, vol. 196, pp. 737–741, Mar. 2011.

[14] H. C. Gifford, P. Kinahan, C. Lartizien, and M. A. King, "Evaluation of multiclass model observers in PET LROC studies," *IEEE Trans. Nucl. Sci.*, vol. 54, no. 1, pp. 116–123, Feb. 2007.

[15] D. J. Kadrmas, E. Casey, Michael, N. F. Black, J. J. Hamill, V. Y. Panin, and M. Conti, "Experimental comparison of lesion detectability for four fully-3D PET reconstruction schemes," *IEEE Tran. Med. Imag.*, vol. 28, no. 4, pp. 523–534, Apr. 2009.

[16] A. Lehovich, P. P. Bruyant, H. S. Gifford, P. B. Schneider, S. Squires, R. Licho, G. Gindi, and M. A. King, "Impact on reader performance for lesion-detection/localization tasks of anatomical priors in SPECT reconstruction," *IEEE Trans. Med. Imag.*, vol. 28, no. 9, pp. 1459–1467, Sep. 2009.

[17] S. Surti, J. Scheuermann, G. E. Fakhri, M. E. Daube-Witherspoon, R. Lim, N. Abi-Hatem, E. Moussallem, F. Benard, D. Mankoff, and J. S. Karp, "Impact of time-of-flight PET on whole-body oncologic studies: A human observer lesion detection and localization study," *J. Nucl. Med.*, vol. 52, no. 5, pp. 712–719, May 2011.

[18] M. Das, H. C. Gifford, J. M. O'Conner, and S. J. Glick, "Penalized maximum likelihood reconstruction for improved microcalcification detection in breast tomosynthesis," *IEEE Trans. Med. Imag.*, vol. 30, no. 4, pp. 904–914, Apr. 2011.

[19] N. A. Obuchowski, M. L. Lieber, and K. A. Powell, "Data analysis for detection and localization of multiple abnormalities with application to mammography," *Acad. Radiol.*, vol. 7, no. 7, pp. 516–525, Jul. 2000.

[20] M. S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford, U.K.: Oxford Univ. Press, 2003.

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

[22] L. Tang, "A family of nonparametric statistics for LROC curves," in *Proc. Int. Conf. Biomed. Eng. Informat.*, May 2008, pp. 758–762.

[23] L. L. Tang and N. Balakrishnan, "A random-sum Wilcoxon statistic and its application to analysis of ROC and LROC data," *J. Stat. Plan. Infer.*, vol. 141, no. 1, pp. 335–344, Jan. 2011.

[24] D. D. Dorfman, K. S. Berbaum, and C. E. Metz, "ROC rating analysis: Generalization to the population of readers and cases with the jackknife method," *Invest. Radiol.*, vol. 27, no. 9, pp. 723–731, Sep. 1992.

[25] N. A. Obuchowski and H. E. Rockette, "Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: An ANOVA approach with dependent observations," *Commun. Stat. Simulat.*, vol. 24, no. 2, pp. 285–308, 1995.

[26] S. L. Hillis, K. S. Berbaum, and C. E. Metz, "Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis," *Acad. Radiol.*, vol. 15, no. 5, pp. 647–661, May 2008.

[27] J. Kowalski and X. M. Tu, *Modern Applied U-Statistics*. New York: Wiley, 2008.

[28] P. K. Sen, "On some convergence properties of U-statistics," *Calcutta Stat. Assoc.*, vol. 10, no. 37-38, pp. 1–18, Nov. 1960.

[29] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characterstic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, Sep. 1988.

[30] H. H. Song, "Analysis of correlated ROC areas in diagonostic testing," *Biometrics*, vol. 53, pp. 370–382, Mar. 1997.

[31] R. F. Wagner, S. V. Beiden, and C. E. Metz, "Continuous versus categorical data for ROC analysis: Some quantitative considerations," *Acad. Radiol.*, vol. 8, no. 4, pp. 328–334, Apr. 2001.

[32] L. Hadjiiski, H.-P. Chan, B. Sahiner, M. A. Helvie, and M. A. Roubidoux, "Quasi-continuous and discrete confidence rating scales for observer performance studies: Effects on ROC analysis," *Acad. Radiol.*, vol. 14, no. 1, pp. 38–48, Jan. 2007.

[33] R. G. Swensson, "Using localization data from image interpretations to improve estimates of performance accuracy," *Med. Decis. Making*, vol. 20, no. 2, pp. 170–185, Apr. 2000.

[34] C. Metz, S. Starr, L. Lusted, and K. Rossmann, "Progress in evaluation of human observer visual detection performance using the ROC curve approach," in *Proc. 4th Int. Conf. Inf. Process. Scintigraphy*, Jul. 1975, pp. 420–436.

[35] B. Liu, S. Kulkarni, and G. Gindi, "The efficiency of the human observer for lesion detection and localization in emission tomography," *Phys. Med. Biol.*, vol. 54, no. 9, pp. 2651–2666, May 2009.

[36] M. L. Puri and P. K. Sen, *Nonparametric Methods in Multivariate Analysis*. New York: Wiley, 1971.

[37] P. K. Sen, *Sequential Nonparametrics*. New York: Wiley, 1981.

[38] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*. New York: Wiley, 1980.

[39] V. P. Bhapkar, "A nonparametric test for the problem of several samples," *Ann. Math. Stat.*, vol. 32, no. 4, pp. 1108–1117, Dec. 1961.

[40] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 5th ed. Upper Saddle River, NJ: Prentice-Hall, 2002.

[41] A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*. Philadelphia, PA: SIAM, 2001.

[42] T. M. Buzug, *Computed Tomography: From Photon Statistics to Modern Cone-Beam CT*. New York: Springer, 2008.

[43] A. Wunderlich and F. Noo, "Image covariance and lesion detectability in direct fan-beam X-ray computed tomography," *Phys. Med. Biol.*, vol. 53, no. 10, pp. 2471–2493, May 2008.

[44] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Duxbury, MA: Duxbury, 2001.

[45] P. K. Sen, "Almost sure convergence of generalized U-statistics," *Ann. Probab.*, vol. 5, no. 2, pp. 287–290, 1977.

[46] B. D. Gallas, "One-shot estimate of MRMC variance: AUC," *Acad. Radiol.*, vol. 13, no. 3, pp. 353–362, Mar. 2006.

[47] B. D. Gallas, A. Bandos, F. W. Samuelson, and R. F. Wagner, "A framework for random-effects ROC analysis: Biases with the bootstrap and other variance estimators," *Commun. Stat. A-Theor.*, vol. 38, no. 15, pp. 2586–2603, 2009.

[48] D. Gur, A. I. Bandos, H. E. Rockette, M. L. Zuley, C. M. Hakim, D. M. Chough, M. A. Ganott, and J. H. Sumkin, "Is an ROC-type response truly always better than a binary response in observer performance studies?," *Acad. Radiol.*, vol. 17, no. 5, pp. 639–645, May 2010.

[49] D. Bamber, "The area above the ordinal dominance graph and the area below the receiver operating characterstic graph," *J. Math. Psych.*, vol. 12, pp. 387–415, 1975.

[50] A. J. Lee, *U-Statistics: Theory and Practice*. New York: Marcel Dekker, 1990.