Confidence intervals for performance assessment of linear observers

Adam Wunderlich[a)] and Frédéric Noo[b)]
*Utah Center for Advanced Imaging Research, Department of Radiology, University of Utah,
Salt Lake City, Utah 84108*

(Received 27 June 2010; revised 20 January 2011; accepted for publication 20 January 2011;
published 20 July 2011)

Purpose: This work seeks to develop exact confidence interval estimators for figures of merit that describe the performance of linear observers, and to demonstrate how these estimators can be used in the context of x-ray computed tomography (CT). The figures of merit are the receiver operating characteristic (ROC) curve and associated summary measures, such as the area under the ROC curve. Linear computerized observers are valuable for optimization of parameters associated with image reconstruction algorithms and data acquisition geometries. They provide a means to perform assessment of image quality with metrics that account not only for shift-variant resolution and nonstationary noise but that are also task-based.

Methods: We suppose that a linear observer with fixed template has been defined and focus on the problem of assessing the performance of this observer for the task of deciding if an unknown lesion is present at a specific location. We introduce a point estimator for the observer signal-to-noise ratio (SNR) and identify its sampling distribution. Then, we show that exact confidence intervals can be constructed from this distribution. The sampling distribution of our SNR estimator is identified under the following hypotheses: (i) the observer ratings are normally distributed for each class of images and (ii) the variance of the observer ratings is the same for each class of images. These assumptions are, for example, appropriate in CT for ratings produced by linear observers applied to low-contrast lesion detection tasks.

Results: Unlike existing approaches to the estimation of ROC confidence intervals, the new confidence intervals presented here have exactly known coverage probabilities when our data assumptions are satisfied. Furthermore, they are applicable to the most commonly used ROC summary measures, and they may be easily computed (a computer routine is supplied along with this article on the Medical Physics Website). The utility of our exact interval estimators is demonstrated through an image quality evaluation example using real x-ray CT images. Also, strong robustness is shown to potential deviations from the assumption that the ratings for the two classes of images have equal variance. Another aspect of our interval estimators is the fact that we can calculate their mean length exactly for fixed parameter values, which enables precise investigations of sampling effects. We demonstrate this aspect by exploring the potential reduction in statistical variability that can be gained by using additional images from one class, if such images are readily available. We find that when additional images from one class are used for an ROC study, the mean AUC confidence interval length for our estimator can decrease by as much as 35%.

Conclusions: We have shown that exact confidence intervals can be constructed for ROC curves and for ROC summary measures associated with fixed linear computerized observers applied to binary discrimination tasks at a known location. Although our intervals only apply under specific conditions, we believe that they form a valuable tool for the important problem of optimizing parameters associated with image reconstruction algorithms and data acquisition geometries, particularly in x-ray CT. © *2011 American Association of Physicists in Medicine*.
[DOI: 10.1118/1.3577764]

Key words: image quality, ROC, CT

I. INTRODUCTION

The development of new reconstruction algorithms and data acquisition strategies often requires image quality assessment to optimize system parameters. A rigorous way to measure image quality is with a task-based approach.[1] This approach requires the clear definition of (1) a task, (2) an observer, and (3) a meaningful figure of merit.[1] Unfortunately, image quality assessment with human observers is not practical for purposes of system optimization. Indeed, because human observer studies are expensive and time-consuming, they are cumbersome for system optimization over a large set of possible system and signal parameters.[2] To overcome this problem, the use of computerized observers has been advocated for image quality assessment for the purpose of system optimization.[1] The results presented in this work are specifically intended for linear computerized observers; they should not be blindly applied to human observers (our assumptions are not likely to be met in this case).

Task-based assessments of image quality using linear computerized observers often involve binary classification. For example, studies of medical image quality frequently evaluate a task in which an observer attempts to discriminate between two classes of images: those images that contain a feature of interest (later called lesion, following customary usage) and those that do not. Observer performance for a binary classification task can be expressed using a receiver operating characteristic (ROC) curve, which plots the true positive fraction (TPF) as a function of false positive fraction (FPF).[1] For purposes of simplification, the whole ROC curve is often reduced to a single number, called an ROC summary measure.[1,3] In practice, ROC curves and ROC summary measures must be estimated from observer rating data obtained experimentally. Consequently, they suffer from statistical variability that must be characterized in order to make inferences about observer performance.

One way that estimator variability may be summarized is through the use of confidence intervals. As opposed to point estimates, confidence intervals provide a probabilistic guarantee of covering the parameter of interest.[4] Moreover, as observed in Ref. 5 a virtue of confidence intervals is that they convey more information than hypothesis testing (together with *p*-values) in two ways. First, confidence intervals communicate the amount of statistical precision involved in an experiment. Second, they communicate the relative size of an experimental effect, i.e., they show how significant experimentally observed differences are in terms of their magnitude. The importance of confidence intervals for ROC analysis of diagnostic image quality has been previously emphasized by Metz.[6]

Previous work that examined confidence intervals with application to ROC analysis has primarily focused on estimation of the area under the ROC curve (AUC), a widely used summary measure; see, e.g., Refs. 7 and 9 for overviews of the literature on this topic. Also, confidence bands for the entire ROC curve have been investigated by.[10,11] The majority of the previously investigated ROC confidence intervals are based on either nonparametric or semiparametric estimation techniques. Such distribution-free methods have the advantage that they are broadly applicable because they make very weak assumptions regarding the distributions of the observer ratings; this makes them suitable for assessment of human observers. However, a drawback of the previously investigated interval and band estimators is a reliance on either asymptotic approximations or resampling techniques. Because these methods are not appropriate for small samples, they can yield confidence intervals with inaccurate coverage probabilities.[8] By contrast, the coverage probabilities for the confidence intervals that we propose in this work are known exactly when our assumptions are satisfied.

In this work, we present fully parametric estimators that yield exact confidence intervals for ROC summary measures and exact confidence bands for ROC curves. Our new estimators are designed for continuous-valued observer ratings under the dual assumptions that (i) the observer ratings are normally distributed for each class of images and (ii) the variance of the observer ratings is the same for each class of images.

Although the aforementioned assumptions appear to be restrictive, they are generally satisfied for image evaluation tasks involving detection of small, low-contrast lesions with linear computerized observers. The reasons are as follows. First, most linear computerized observers compute each observer rating as a linear combination of a large number of image pixel values. Therefore, a general formulation of the central limit theorem[12] implies that the ratings will tend to be normally distributed for each class of images. For x-ray computed tomography (CT), the normality of the observer ratings is further re-enforced by the near-normality of measured data[13] and the linearity of reconstruction algorithms. Second, the absence or presence of a small, low-contrast lesion usually has little impact on the image covariance matrix, so that the variance of ratings produced by linear computerized observers is practically the same for each class of images. This second observation has been made by Barrett and Myers[1(p. 1209)] in the context of nuclear medicine. We carefully analyze its applicability for CT in Sec. IV.

After reviewing relevant background material, we present our new confidence interval estimators. Subsequently, we evaluate two aspects of the AUC interval estimator. The first aspect regards the statistical utility of using more images from one class than the other. Knowledge of this utility is important, as it is often possible to get more images from the class without a lesion; see e.g., Ref. 14. The second aspect regards the robustness of our estimator for application to rating data with unequal variances for the two classes. We examine cases that are likely to be extreme for linear computerized observers performing low-contrast lesion detection tasks with CT images. Finally, the paper illustrates the usefulness of our estimators in the context of image quality evaluation with real x-ray CT images.

II. PRELIMINARIES

This section introduces our notation and reviews important background material. First, we remind the reader of the definition of the noncentral *t* distribution. The remainder of the section reviews summary measures of observer performance.

Throughout the text, the probability density function (pdf) of a continuous random variable X, will be written as $f_X(x)$, and its cumulative distribution function (cdf) will be denoted as $F_X(x)$. If $f_X(x)$ and $F_X(x)$ depend on a parameter θ, then they will be written as $f_X(x\,;\,\theta)$ and $F_X(x;\theta)$, respectively. Similarly, if $f_X(x)$ and $F_X(x)$ depend on several parameters $\theta_1, \theta_2, \ldots, \theta_m$, then they will be written as $f_X(x\,;\,\theta_1, \theta_2, \ldots, \theta_m)$ and $F_X(x; \theta_1, \theta_2, \ldots, \theta_m)$, respectively.

We assume that the reader is familiar with the normal and χ^2 probability distributions.[4] If a random variable X is normally distributed with mean μ and variance σ^2, this fact will be written as $X \sim \mathcal{N}(\mu, \sigma^2)$. Similarly, a χ^2 distributed random variable Y with ν degrees of freedom is denoted by $Y \sim \chi^2_\nu$. For only the particular case of the standard normal distribution, $\mathcal{N}(0,1)$, we depart from the notation introduced above for the cdf and employ the usual notations $\Phi(x)$ and $\Phi^{-1}(p)$ for the cdf and inverse cdf, respectively.

II.A. The noncentral *t* distribution

If $X \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_\nu$ are independent random variables, then for any $\delta \in (-\infty, \infty)$, the ratio

$$T = \frac{X+\delta}{\sqrt{Y/\nu}} \tag{1}$$

has a noncentral *t* distribution with ν degrees of freedom and noncentrality parameter δ.[15] In this case, we write $T \sim t'_\nu(\delta)$. The mean of a noncentral *t* random variable is[15(p. 513)]

$$\mathrm{E}[T] = \delta B((\nu-1)/2, 1/2)\sqrt{\nu/2\pi} \quad \text{for} \quad \nu > 1, \tag{2}$$

where $B(x,y)$ is the Euler beta function. Expressions for the pdf, cdf, and higher moments of the noncentral *t* distribution may be found in Ref. 15.

II.B. ROC figures of merit

We consider assessment of linear computerized observer performance for a binary classification task in which the observer must discriminate between two classes of images, denoted as class 1 and class 2. For a detection task, these classes could correspond to normal and diseased conditions, respectively. Given an image, a linear computerized (model) observer[1] computes a scalar, continuous-valued rating statistic y, as the inner product of a fixed (nonrandom) $q \times 1$ template, $\mathbf{w}$, with the image, $\mathbf{p}$, represented as a $q \times 1$ column vector, i.e., $y = \mathbf{w}^T\mathbf{p}$. The observer then compares y to a threshold, c, to classify the image. If $y > c$, the observer decides the image is from class 2. Otherwise, the image is classified as belonging to class 1.

For each threshold, c, the observer's performance is fully characterized by two quantities, called the true positive fraction (TPF) and the false positive fraction (FPF).[1,3] The TPF is the probability that the observer correctly classifies a class-2 image as belonging to class 2, whereas the FPF is the probability that the observer incorrectly classifies a class-1 image as belonging to class 2. Since each value of c results in a different TPF and FPF, observer performance over all thresholds is completely described by the curve of (FPF, TPF) values parameterized by c. This curve is called the receiver operating characteristic (ROC) curve.[1,3] To denote the TPF as a function of the FPF, we will write TPF(FPF).

Throughout this paper, we assume that y is normally distributed with equal variances for each class, i.e, $y \sim \mathcal{N}(\mu_1, \sigma^2)$ and $y \sim \mathcal{N}(\mu_2, \sigma^2)$ for images from classes 1 and 2, respectively. In this case, the ROC curve takes the form [Ref. 3 (p. 82), Result 4.7]

$$\mathrm{TPF(FPF)} = \Phi(\mathrm{SNR} + \Phi^{-1}(\mathrm{FPF})), \tag{3}$$

where

$$\mathrm{SNR} = \frac{\mu_2 - \mu_1}{\sigma} \tag{4}$$

is the observer signal-to-noise ratio. The SNR is sometimes used as figure of merit, since it may be interpreted as a measure of the distance between the distributions of classes 1 and 2. It must be remembered, however, that the SNR is useful only when the variance is a good measure of the spread of the distribution for y (Ref. 1, p. 819); this is the case when y is normal under each class.

Another useful figure of merit for observer performance is the area under the ROC curve, denoted as AUC. The AUC generally falls in the range [0, 1], with larger values signifying greater discrimination ability, and values less than 0.5 indicating that, on average, the observer performs worse than guessing. The AUC may be interpreted as the average TPF, averaged over the entire range of FPF values.[3] Under our distributional assumptions, the AUC may be calculated as[1(p. 819),3(p. 84)]

$$\mathrm{AUC} = \Phi(\mathrm{SNR}/\sqrt{2}). \tag{5}$$

If only a restricted range of FPF values is considered relevant for observer performance, then the partial area under the ROC curve, defined as

$$\mathrm{pAUC}(\mathrm{FPF}_0, \mathrm{FPF}_1) = \int_{\mathrm{FPF}_0}^{\mathrm{FPF}_1} \mathrm{TPF(FPF)}\, d(\mathrm{FPF}) \tag{6}$$

may be used as a summary measure. Observe that under our assumptions for y, TPF at fixed FPF, AUC, and pAUC are strictly increasing functions of SNR only. We use this property in the next section to construct our confidence interval estimators.

It is well-known that the ROC curve is invariant under any strictly increasing transformation of y [Ref. 3 (p. 69), Result 4.1]. Therefore, AUC and pAUC are also invariant under any such transformation. Likewise, if SNR is defined from AUC via Eq. (5), then it too is invariant under any strictly increasing transformation of y. However, it is important to recognize that if SNR is computed using Eq. (4), then the resulting value is *not* invariant under any such transformation, since this relation depends on the first and second moments of y in each class. The confidence interval estimators that we propose in the next section each rely on a point estimate of SNR, motivated by Eq. (4). Therefore, our interval estimators are not invariant under arbitrary strictly increasing transformations of y. Nonetheless, they are invariant under any strictly increasing affine transformation of y, and we will see that they possess attractive properties.

The figures of merit discussed above are widely used and accepted summary measures for observer performance.[1,3] For additional examples of summary measures, see (Ref. 3, Sec. 4.3.3) and (Ref. 16).

III. CONSTRUCTION OF INTERVAL ESTIMATORS

Suppose that an observer rates n_1 images from class 1 and n_2 images from class 2. Denote these ratings for classes 1 and 2 as $y_1^{(1)}, y_2^{(1)}, \ldots, y_{n_1}^{(1)}$ and $y_1^{(2)}, y_2^{(2)}, \ldots, y_{n_2}^{(2)}$, respectively. We wish to estimate confidence intervals for summary measures of observer performance from this finite sample of rating data. In this section, we introduce our estimators assuming that $y \sim \mathcal{N}(\mu_1, \sigma^2)$ and $y \sim \mathcal{N}(\mu_2, \sigma^2)$ for images from classes 1 and 2, respectively, where μ_1, μ_2, and σ^2 are unknown. Each of our interval estimators is based on a point estimator for

SNR, which is introduced first. For a clear presentation, all mathematical proofs are deferred to the appendices.

III.A. Point estimation of SNR

Let the sample mean and the sample variance for class k be $\bar{y}_k = (1/n_k)\sum_{i=1}^{n_k} y_i^{(k)}$ and $s_k^2 = [1/(n_k-1)]\sum_{i=1}^{n_k}(y_i^{(k)} - \bar{y}_k)^2$, respectively. Also, define a pooled estimator for σ^2 as $s^2 = [1/(n_1+n_2-2)]\big((n_1-1)s_1^2 + (n_2-1)s_2^2\big)$. With these definitions in place, we define an estimator of SNR as

$$\widehat{\mathrm{SNR}} = \gamma\,(\bar{y}_2 - \bar{y}_1)/s, \tag{7}$$

with

$$\gamma = \frac{\sqrt{2\pi/(n_1+n_2-2)}}{B((n_1+n_2-3)/2,1/2)}, \tag{8}$$

where $B(x,y)$ is the Euler beta function. The multiplicative factor γ is chosen to make the estimator unbiased. We have the following characterization of $\widehat{\mathrm{SNR}}$, which is proved in Appendix A.

Theorem 1. Suppose that $y \sim \mathcal{N}(\mu_1, \sigma^2)$ for images from class 1 and that $y \sim \mathcal{N}(\mu_2, \sigma^2)$ for images from class 2. Also, suppose that $\widehat{\mathrm{SNR}}$ is computed from independent samples of y, denoted as $y_1^{(1)}, y_2^{(1)}, \ldots, y_{n_1}^{(1)}$ and $y_1^{(2)}, y_2^{(2)}, \ldots, y_{n_2}^{(2)}$, corresponding to classes 1 and 2, respectively. Then

(i) $\eta\widehat{\mathrm{SNR}} \sim t'_v(\delta)$ with $\eta = \frac{1}{\gamma}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$, $v = n_1 + n_2 - 2$, and $\delta = \mathrm{SNR}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$

(ii) $\widehat{\mathrm{SNR}}$ is the unique uniformly minimum variance unbiased (UMVU) estimator of SNR.

III.B. Confidence interval estimation

Given a random variable, X, with a distribution depending on a parameter θ, one may define a random interval estimate $[\theta_L(X),\ \theta_U(X)]$ for θ. This interval is said to be a $1-\alpha$ confidence interval for θ if $P(\theta \in [\theta_L(X),\ \theta_U(X)]) = 1-\alpha$ for any value of θ.[17]

Our knowledge of the sampling distribution for $\widehat{\mathrm{SNR}}$ implies the next theorem, which is proved in Appendix B. It allows us to compute confidence intervals for SNR with exact coverage probabilities.

Theorem 2. Suppose that the hypotheses of Theorem 1 are satisfied. Let $\alpha_1, \alpha_2 \in (0,1)$ be such that $\alpha_1 + \alpha_2 = \alpha$ for some $\alpha \in (0,1)$, and let $T = \eta\widehat{\mathrm{SNR}}$ with $\eta = \frac{1}{\gamma}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$. Then

(i) For each observation t of T, there exist unique values $\delta_L(t)$ and $\delta_U(t)$ in $(-\infty,\infty)$ satisfying $F_T(t\,;\,\nu,\delta_L(t)) = 1-\alpha_1$ and $F_T(t\,;\,\nu,\delta_U(t)) = \alpha_2$, where $F_T(t\,;\,\nu,\delta)$ is the cdf of the noncentral t distribution with $\nu = n_1 + n_2 - 2$.

(i) The random interval $[\delta_L(T)/(\gamma\eta),\ \delta_U(T)/(\gamma\eta)]$ is an exact $1-\alpha$ confidence interval for SNR.

Hence, we can calculate a $1-\alpha$ confidence interval for SNR from a realization of $\widehat{\mathrm{SNR}}$ by numerically solving the relations in Theorem 2(i) for δ_L and δ_U and then substituting these values into the expression of Theorem 2(ii). Above, if $\alpha_1 = 0$, then $\delta_L(t) = -\infty$ and if $\alpha_2 = 0$, then $\delta_U(t) = \infty$. In either of these cases, the confidence interval is said to be one-sided. Otherwise, the interval is said to be two-sided.[17]

The following corollary shows that we can also calculate exact confidence intervals for TPF(FPF), AUC, and pAUC (FPF_0, FPF_1). It follows from Theorem 2 and the strictly increasing transformation property of confidence intervals, which is stated and proved in Appendix B as Lemma 3.

Corollary 1. Suppose that the hypotheses of Theorem 2 are satisfied. Let $\mathrm{SNR_L(T)}=\delta_\mathrm{L}(\mathrm{T})/(\gamma\eta)$ and $\mathrm{SNR_U(T)}=\delta_\mathrm{U}(\mathrm{T})/(\gamma\eta)$. Then the random intervals

$$[\mathrm{TPF}(\mathrm{FPF}\,;\,\mathrm{SNR}_L(\mathrm{T})),\ \mathrm{TPF}(\mathrm{FPF}\,;\,\mathrm{SNR}_U(\mathrm{T}))],$$

$$[\mathrm{AUC}(\mathrm{SNR}_L(\mathrm{T})),\ \mathrm{AUC}(\mathrm{SNR}_U(\mathrm{T}))],$$

and

$$[\mathrm{pAUC}(\mathrm{FPF}_0, \mathrm{FPF}_1\,;\,\mathrm{SNR}_L(\mathrm{T})),\ \mathrm{pAUC}(\mathrm{FPF}_0, \mathrm{FPF}_1\,;\,\mathrm{SNR}_U(\mathrm{T}))],$$

defined by substituting $\mathrm{SNR}_L(T)$ and $\mathrm{SNR}_U(T)$ for SNR in Eqs. (3), (5), and (6) are exact $1-\alpha$ confidence intervals for TPF(FPF), AUC, and pAUC(FPF_0, FPF_1), respectively.

A MATLAB® routine that calculates the confidence intervals for SNR, TPF, AUC, and pAUC is provided in along with the article on the Medical Physics Website. Note that because $\widehat{\mathrm{SNR}}$ is invariant under strictly increasing affine transformations of y, it follows that our confidence intervals also share this invariance.

A confidence band for the entire ROC curve may be found from a confidence interval for SNR in the sense of the next theorem, which is proved in Appendix C. We denote the collection of points on the ROC curve as $\Omega_{\mathrm{ROC}} = \{(\mathrm{FPF}, \mathrm{TPF})\ :\ \mathrm{FPF} \in [0,1]\}$.

Theorem 3. Suppose that $y \sim \mathcal{N}(\mu_1, \sigma^2)$ for images from class 1 and $y \sim \mathcal{N}(\mu_2, \sigma^2)$ for images from class 2. Let $[\mathrm{SNR}_L,\ \mathrm{SNR}_U]$ be a $1-\alpha$ confidence interval for SNR, and define the set

$$\widehat{\Omega}_{\mathrm{ROC}} = \{(\mathrm{FPF}, \mathrm{T})\ :\ \mathrm{FPF} \in [0,1] \text{ and } \mathrm{T} \in \mathcal{I}\},$$

where

$$\mathcal{I} = [\mathrm{TPF}(\mathrm{FPF}\,;\,\mathrm{SNR}_L),\ \mathrm{TPF}(\mathrm{FPF}\,;\,\mathrm{SNR}_U)].$$

Then $\widehat{\Omega}_{\mathrm{ROC}}$ is a $1-\alpha$ confidence band for the ROC curve in the sense that, for any value of SNR, Ω_{ROC} is contained in $\widehat{\Omega}_{\mathrm{ROC}}$ with probability $1-\alpha$, i.e., $P(\Omega_{\mathrm{ROC}} \subset \widehat{\Omega}_{\mathrm{ROC}}) = 1-\alpha$.

Observe that the $1-\alpha$ confidence band defined in the above theorem is equivalent to the union over all FPF values of $1-\alpha$ confidence intervals for TPF. This construction of a simultaneous $1-\alpha$ confidence band for the ROC curve is possible because our assumptions imply that the ROC curve is parameterized by only SNR. More generally, when the ROC curve is parameterized by more than one parameter (as in the binormal model[3]), the confidence band formed from the union of $1-\alpha$ TPF intervals will have a coverage probability smaller than $1-\alpha$ for the whole ROC curve simultaneously.[10]

IV. PROPERTIES OF THE AUC CONFIDENCE INTERVALS

In this section, we examine two aspects of the previously introduced AUC confidence intervals. First, we explore the potential advantage offered by using additional images from class 1 if such images are readily available. Second, for situations relevant to CT image quality assessment, we evaluate the robustness of the proposed intervals to violation of the equal-variance assumption on the rating data.

IV.A. Advantage gained by using additional images from one class

In some circumstances, it may be possible to obtain additional images from one class of images at low cost; see, e.g., Ref. 14. Therefore, it is desirable to examine the potential decrease in statistical variability that may be gained by using such extra images. Below, we consider the case when more images are available for class 1. However, due to the symmetric role of n_1 and n_2 in our confidence interval estimators, the same conclusions also hold when there are more images for class 2.

For our evaluations, we assessed the mean 95% AUC confidence interval length, defined as $\text{MCIL}_{.95} = \text{E}[\text{AUC}_U - \text{AUC}_L]$ for fixed values of n_1, n_2, and AUC, where AUC_U and AUC_L are the upper and lower endpoints, respectively, of the 95% ($\alpha_1 = \alpha_2 = 0.025$) confidence interval estimator for AUC. To compute this expected value, we numerically evaluated the integral

$$\text{MCIL}_{.95}(n_1, n_2, \text{AUC}) = \int_{-\infty}^{\infty} [\text{AUC}_U(x) - \text{AUC}_L(x)] f_{\widehat{\text{SNR}}}(x\,;\, n_1, n_2, \text{AUC})\, dx \tag{9}$$

$$= \int_{-\infty}^{\infty} [\text{AUC}_U(x) - \text{AUC}_L(x)] \eta f_T(\eta x\,;\, \nu, \delta)\, dx, \tag{10}$$

where η, ν, and δ are as given in Theorem 1(i). In Eq. (10), the pdf of $\widehat{\text{SNR}}$, $f_{\widehat{\text{SNR}}}$, was rewritten in terms of the noncentral t pdf, f_T, using Theorem 1(i) and a standard result for the pdf of a monotonic transformation of a random variable [Ref. 4 (p. 51), Theorem 2.1.5].

Figure 1 contains plots of the relative decrease in mean 95% AUC confidence interval length that is obtained by increasing n_1 relative to n_2. In these plots, the relative percentage decrease (RPD) in $\text{MCIL}_{.95}$, relative to the $n_1 = n_2$ case, was calculated as

$$\text{RPD}(n_1, n_2, \text{AUC}) = \frac{\text{MCIL}_{.95}(n_2, n_2, \text{AUC}) - \text{MCIL}_{.95}(n_1, n_2, \text{AUC})}{\text{MCIL}_{.95}(n_2, n_2, \text{AUC})} \times 100. \tag{11}$$

The plots in Fig. 1 indicate that the mean AUC confidence interval length can shrink by as much as 35% when n_1 is increased relative to n_2. In particular, the relative decrease is greatest for small values of n_2 and for large values of AUC. Moreover, the plots illustrate that the advantage gained by increasing n_1 relative to n_2 flattens out for large values of n_1.

IV.B. Applicability for CT image quality evaluation: Robustness to violation of the equal-variance assumption

The ROC confidence intervals introduced in Sec. III assume that the variance of the ratings is the same for each class of images, i.e., $\sigma_1^2 = \sigma_2^2$, where σ_1^2 and σ_2^2 are the rating variances for classes 1 and 2, respectively. As discussed in the introduction, this assumption should be a good approximation for linear computerized observers applied to tasks involving the detection of small, low-contrast lesions in CT images. We now take a closer look at the quality of the $\sigma_1^2 = \sigma_2^2$ approximation in the context of CT image quality evaluation, and then we examine the coverage probability of the AUC interval estimator in extreme cases.

Consider a uniform circular cylinder B of diameter D, which may, or may not, contain a small spherical lesion L of diameter d at its center; see Fig. 2. Denote the linear x-ray

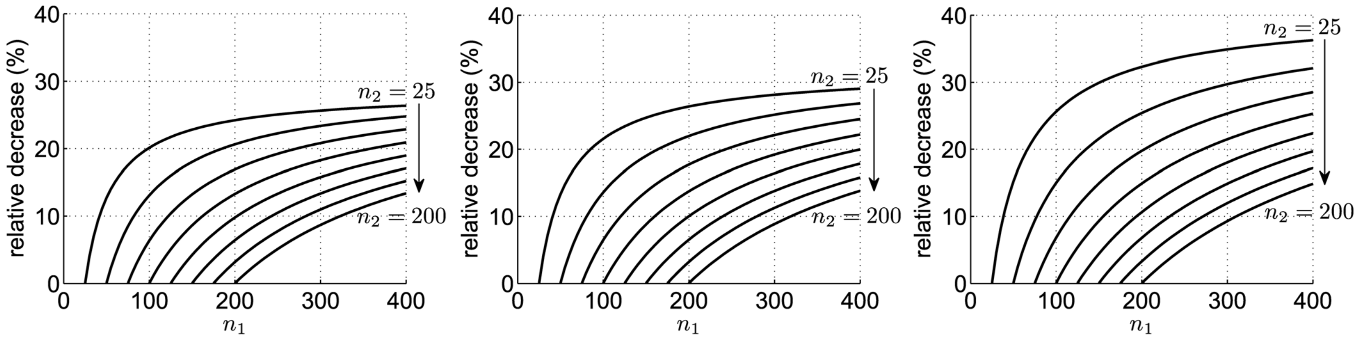

FIG. 1. Percentage decrease of mean (95%) AUC confidence interval length, relative to the $n_1 = n_2$ case, plotted as a function of n_1, with n_2 and AUC held fixed. From top to bottom, the curves correspond to n_2 values of 25, 50, 75, 100, 125, 150, 175, and 200, respectively. The plots are for AUC values of 0.6 (left), 0.75 (center), and 0.9 (right).

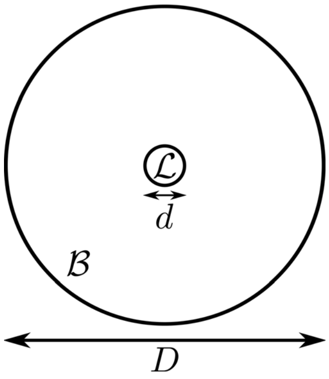

FIG. 2. Depiction of a small lesion embedded in a larger, uniform cylinder.

attenuation coefficients for the cylinder and for the lesion as μ_B and μ_L, respectively. Assuming a conventional Poisson noise model and a monochromatic beam,[18,19] the variance of the measured CT data for a ray passing through the center of B when the lesion is not present is

$$\eta_1^2 = 1/[N_i \exp(-D\mu_B)], \tag{12}$$

where N_i is the number of photons entering the cylinder. When the lesion L is present, the variance of the measured CT data for a ray passing through the center of B becomes

$$\eta_2^2 = 1/[N_i \exp(-D\mu_B)\exp(-dC\mu_W/1000)], \tag{13}$$

where μ_W is the linear x-ray attenuation coefficient for water, and the lesion contrast C is defined in Hounsfield units as $C = (\mu_L - \mu_B)1000/\mu_W$. Forming the ratio of the above variances, we have

$$\eta_1^2/\eta_2^2 = \exp(-dC\mu_W/1000) \tag{14}$$

Hence, the ratio of the variances for rays passing through the lesion always falls between $\exp(-dC\mu_W/1000)$ and one. It is straightforward to see that this result is very general and is, for example, applicable to noncircularly symmetric background objects and arbitrarily located lesions. Now, suppose that image reconstruction is performed using the classical 2-D filtered backprojection (FBP) algorithm. Then, expressions given for the image variance and covariance in[18,19] show that the entry-wise ratio of the image covariance matrices for each class is, to a good approximation, bounded between $\exp(-dC\mu_W/1000)$ and one. Consequently, it follows that for linear computerized observers, the ratio of the rating variances, σ_1^2/σ_2^2, is approximately bounded between $\exp(-dC\mu_W/1000)$ and one. The bound $\exp(-dC\mu_W/1000)$ is tabulated in Table I for various lesion diameters, d, and contrasts, C, assuming a linear x-ray attenuation coefficient of $\mu_W = 0.183\,\text{cm}^{-1}$ for water.

Examining Table I, we see that for a large variety of lesion detection tasks, it is safe to assume that $0.95 \leq \sigma_1^2/\sigma_2^2 \leq 1.05$. Moreover, the values $\sigma_1^2/\sigma_2^2 = 0.95$ and $\sigma_1^2/\sigma_2^2 = 1.05$ correspond to extreme cases in which the lesion must either be very large or have a high contrast. In light of these results, we next examine the coverage probability of the AUC interval estimator when the equal-variance assumption is violated with either $\sigma_1^2/\sigma_2^2 = 0.95$ or $\sigma_1^2/\sigma_2^2 = 1.05$. We consider two scenarios: (i) $n_1 = n_2$ and (ii) $n_1 = 2\,n_2$.

TABLE I. Bounds on the variance ratio of the rating data, σ_1^2/σ_2^2, for a lesion of diameter d mm with contrast C HU. (The other bound on σ_1^2/σ_2^2 is always 1).

d(mm)	C(HU)	$\exp(-dC\mu_W/1000)$
1	100	0.9982
1	–100	1.0018
5	500	0.9553
5	10	0.9991
5	–10	1.0009
5	–500	1.0468
10	50	0.9909
10	–50	1.0092
15	50	0.9864
15	–50	1.0138
20	100	0.9641
20	75	0.9729
20	–75	1.0278
20	–100	1.0373

For each choice of the parameters n_1, n_2, AUC, and $\kappa = \sigma_1^2/\sigma_2^2$, we performed 10^7 Monte Carlo trials to estimate the coverage probability of the 95% ($\alpha_1 = \alpha_2 = .025$) AUC confidence intervals. Without loss of generality, we assumed that the ratings for class 1 came from a standard normal distribution since our confidence interval estimators are invariant under strictly increasing affine transformations of the ratings. An individual trial was carried out by first randomly generating n_1 values of $y \sim \mathcal{N}(0,1)$ and n_2 values of $y \sim \mathcal{N}(\mu, \sigma^2)$, with $\mu = \Phi^{-1}$ (AUC)$\sqrt{(\kappa+1)/\kappa}$ and $\sigma^2 = 1/\kappa$. [The formula for μ was found using the AUC expression given in Ref. 1 (p. 819) for the general case when σ_1^2 may not equal σ_2^2.] Then, a single interval estimate for AUC was calculated from the ratings using the steps described in Sec. III. After running the trials, the coverage probability was estimated as the proportion of the 10^7 trials for which the estimated interval covered the true AUC value. In addition, 95% confidence intervals for the coverage probability were estimated using the Wilson score method advocated in (Ref. 20) for binomial proportions. In all cases, the Wilson score intervals indicated that the upper (respectively lower) bound of a conservative 95% confidence interval for the coverage probability may be obtained by adding 0.014% to (respectively subtracting 0.014% from) each point estimate expressed in percent.

Table II contains the estimated coverage probabilities for the $\sigma_1^2/\sigma_2^2 = 0.95$ and $\sigma_1^2/\sigma_2^2 = 1.05$ cases in the $n_1 = n_2$ scenario. From this table, we see that for all AUC values and choices of $n_1 = n_2$, the coverage probabilities are very close to 95%. The estimated coverage probabilities in the unbalanced $n_1 = 2n_2$ scenario are shown in Table III for the $\sigma_1^2/\sigma_2^2 = 0.95$ and $\sigma_1^2/\sigma_2^2 = 1.05$ cases. For all of the tested AUC values and choices of $n_1 = 2n_2$, the coverage probabilities for the $\sigma_1^2/\sigma_2^2 = 1.05$ case are conservative. On the other hand, the coverage probabilities for the $\sigma_1^2/\sigma_2^2 = 0.95$ case are all slightly less than 95% in the $n_1 = 2n_2$ scenario.

TABLE II. Estimated coverage probabilities (in percent) for two-sided 95% AUC confidence intervals generated from normally distributed rating data with $n_1 = n_2 = n$. The tables correspond to variance ratios of $\sigma_1^2/\sigma_2^2 = 0.95$ and $\sigma_1^2/\sigma_2^2 = 1.05$. In all cases, the upper and lower bounds of a conservative 95% confidence interval for the coverage probability may be obtained by adding and subtracting 0.014% to/from each point estimate, respectively.

AUC \ n	25	50	75	100
$\sigma_1^2/\sigma_2^2 = 0.95$				
0.52	95.00	95.00	95.00	94.99
0.55	95.00	95.01	95.00	94.99
0.60	94.99	95.00	94.99	95.00
0.70	94.99	95.00	95.00	95.01
0.80	95.00	95.00	95.00	95.01
0.90	94.99	94.99	95.00	95.01
0.95	94.99	94.99	94.99	94.99
0.98	95.00	95.00	95.01	95.00
$\sigma_1^2/\sigma_2^2 = 1.05$				
0.52	95.01	95.00	95.00	95.01
0.55	95.00	95.01	95.00	95.00
0.60	94.99	95.01	94.99	95.00
0.70	94.99	95.00	95.01	95.00
0.80	95.00	95.00	95.00	95.00
0.90	94.99	95.01	95.00	95.00
0.95	94.99	95.00	94.99	95.00
0.98	95.00	95.01	95.01	94.99

Overall, these results indicate that our AUC interval estimators maintain accurate coverage probabilities even in the extreme cases $\sigma_1^2/\sigma_2^2 = 0.95$ and $\sigma_1^2/\sigma_2^2 = 1.05$. They also exhibit a difference in behavior between having $n_1 = n_2$ or not—the error in coverage probability is smaller when $n_1 = n_2$. Also, when n_1 is different from n_2, the coverage probability may be conservative or not depending on the value of the variance ratio.

TABLE III. Estimated coverage probabilities (in percent) for two-sided 95% AUC confidence intervals generated from normally distributed rating data with $n_1 = 2n$ and $n_2 = n$. The tables correspond to variance ratios of $\sigma_1^2/\sigma_2^2 = 0.95$ and $\sigma_1^2/\sigma_2^2 = 1.05$). In all cases, the upper and lower bounds of a conservative 95% confidence interval for the coverage probability may be obtained by adding and subtracting 0.014% to/from each point estimate, respectively.

AUC \ n	25	50	75	100
$\sigma_1^2/\sigma_2^2 = 0.95$				
0.52	94.81	94.80	94.81	94.81
0.55	94.80	94.80	94.80	94.81
0.60	94.81	94.81	94.81	94.81
0.70	94.81	94.81	94.82	94.80
0.80	94.83	94.84	94.83	94.82
0.90	94.87	94.85	94.86	94.85
0.95	94.90	94.88	94.86	94.86
0.98	94.90	94.92	94.90	94.88
$\sigma_1^2/\sigma_2^2 = 1.05$				
0.52	95.19	95.18	95.19	95.19
0.55	95.18	95.18	95.18	95.19
0.60	95.18	95.18	95.18	95.19
0.70	95.17	95.16	95.17	95.15
0.80	95.14	95.14	95.14	95.12
0.90	95.10	95.09	95.10	95.08
0.95	95.08	95.05	95.04	95.04
0.98	95.02	95.04	95.02	95.01

Last, note that although the evaluations in this section were performed only for AUC confidence intervals, the same conclusions hold for the SNR interval estimator discussed in (Sec. III B), since SNR and AUC are related through a continuous, strictly increasing transformation that does not depend on σ_1^2/σ_2^2.

V. APPLICATION TO TASK-BASED IMAGE QUALITY EVALUATION

We present here an example of how our estimators can be used in the context of task-based image quality evaluation. This example involves real x-ray computed tomography (CT) images but is not meant to recommend a specific methodology for assessment of image quality in CT. Specifically, our choices for the task and for the observer are not optimal for CT images. Furthermore, the example should not be taken as evidence in favor of one reconstruction strategy over another. Our purpose is simply to demonstrate the usefulness of the tools that we have developed in this paper for interval estimation of ROC summary measures and of ROC curves.

A Siemens SOMATOM® Sensation™ 64 CT scanner was employed to repeatedly scan a thorax phantom 186 times over a circular source trajectory. The phantom consisted of a torso constructed by QRM (Möhrendorf, Germany)[21] together with two different water bottles attached to the sides to simulate arms. A mean image of the whole phantom estimated from 186 reconstructions is shown in Fig. 3(left). The scans were executed in a thorax scan mode using a two-slice acquisition with a slice thickness of 1 mm and a rotation speed of 3 revolutions per second. The x-ray tube settings were 25 mAs and 120 kVp, and the data acquisition was performed with no tube current modulation to accentuate noise correlation in the image. The measurements for the first of the two slices over the 186 repeated scans constituted 186 fan-beam data sets that were used for the image quality evaluation.

The CT data was read using software supplied by Siemens and fed directly into our implementation of the

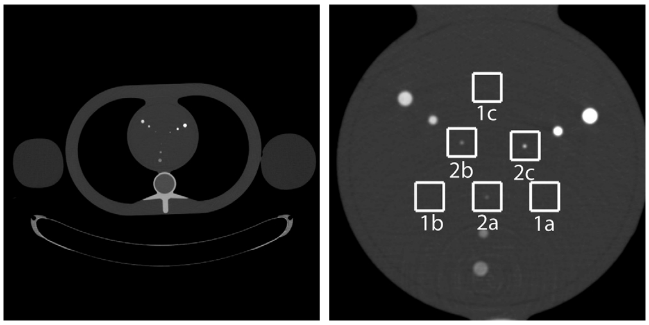

FIG. 3. Mean images of the QRM phantom displayed with a grayscale window of [–200, 600] HU. Whole phantom (left) and reconstruction focused on the heart insert (right) with regions of interest marked with white boxes. ROI-1a, ROI-1b, and ROI-1c contain no lesion. ROI-2a, ROI-2b, and ROI-2c contain a low-contrast, and a medium-contrast, and high-contrast lesion, respectively.

classical filtered backprojection (FBP) algorithm for direct reconstruction from either short-scan or full-scan fan-beam data; see Ref. 19 for a description of the algorithm as we used it. The short-scan reconstructions were performed using only 230° of the source trajectory. The reconstructions were consistently performed on a grid of 550 × 550 square pixels, each of size 0.02 × 0.02 cm, that was centered on the heart insert, as shown in Fig. 3 (right).

For both short-scan and full-scan reconstructions, we considered a lesion detection task that amounts to determining whether or not a lesion is present at the center of a region-of-interest (ROI). To train and test an observer for this task, we identified six regions of interest (ROI) in the heart insert: three regions without lesions, labeled as ROI-1a, ROI-1b, and ROI-1c, and three regions with a lesion at their center, labeled as ROI-2a, ROI-2b, and ROI-2c; see Fig. 3(right). Each ROI when viewed as an image consists of 50 × 50 pixels. The lesion diameter in each of the lesion-present images is 1 mm, but the contrast varies. Specifically, the lesion contrast is 210 HU for ROI-2a, 452 HU for ROI-2b, and 997 HU for ROI-2c. In each case, the background value is 40 HU. We assumed that the images of the six ROIs obtained from one given CT data set are independent. This is justified by a previous study of direct fan-beam FBP reconstruction from simulated CT data, which indicated that correlations between image pixels are negligible over the distance (1–2 cm) that separates the ROIs.[19] For examples of practical image quality studies that employ multiple ROIs in a similar fashion, see Refs. 2 and 14.

For our observer, we used a trained channelized Hotelling observer (CHO), which is a popular type of linear computerized observer; for details, see Ref. 1. The CHO was implemented with 40 Gabor channels using the parameters given in Ref. 19. A specific template was built for each type of reconstruction (full-scan and short-scan). These two templates were each estimated (trained) using the first 50 CT data sets. The class-1 images used for the training were the images defined by ROI-1a, ROI-1b, and ROI-1c, pooled together, and the class-2 images used for the training were the images for ROI-2a, ROI-2b, and ROI-2c, pooled together. In total, 150 class-1 and 150 class-2 images were thus used to train the observer for each reconstruction type.

Once the observer was trained, we tested its performance on a lesion detection task. The class-1 images for this task were defined using only ROI-1a, and the class-2 images were defined using only ROI-2a. The observer performance for the detection task was estimated (tested) using the remaining 136 CT data sets in a fully paired design, i.e., the same CT data sets were used for the short-scan and full-scan testing. Hence, for each type of scan, $n_1 = 136$ class-1 images and $n_2 = 136$ class-2 images were used for testing the observer.

For both the short-scan and full-scan reconstructions, our two assumptions that (i) the observer ratings are normally distributed for each class of images, and (ii) the variance of the observer ratings is the same for each class of images, were further verified with hypothesis tests. The first assumption was checked by performing the Lilliefors normality test[22] at the $\alpha = 0.10$ significance level, using the built-in MATLAB® function *lillietest*. In all cases, there was not enough evidence at the 0.10 significance level to reject the null hypothesis that the ratings for each class came from a normal distribution. The second assumption was checked by performing the two-sample F-test for equal variances at the $\alpha = 0.10$ significance level using the built-in MATLAB® function *vartest*2. For both types of reconstruction, there was not enough evidence at the 0.10 significance level to reject the null hypothesis that the ratings were normally distributed in each class with equal variances. The p-values for these tests are reported in Table IV. Note that because all of the p-values are very large, they cast little doubt on the validity of the null hypothesis for each test.

TABLE IV. Estimated p-values for the example. The p-values for the Lilliefors normality test and for the two-sample F-test of equal variances.

	Short-scan	Full-scan
The p-values for the Lilliefors normality test		
class 1	0.744	0.370
class 2	0.905	0.544
The p-values for the two-sample F-test of equal variances		
	0.279	0.710

We estimated the observer performance by first applying Eq. (7) with $n_1 = 136$ and $n_2 = 136$ to obtain the value of SNR. Then, we applied the MATLAB® routine supplied along with the article on the Medical Physics Website with the following parameters: $\alpha_1 = \alpha_2 = 0.025$, $\mathrm{FPF}_0 = 0$, $\mathrm{FPF}_1 = 0.2$ and a fine sampling of FPF values over the range [0,1].

Table V gives the estimated 95% confidence intervals for SNR, AUC, and pAUC corresponding to the short-scan and full-scan reconstructions, respectively. In addition, the estimated 95% confidence bands for the entire ROC curve are displayed in Fig. 4.

Table V indicates that the 95% confidence intervals for observer performance overlap for the short-scan and full-scan reconstructions, with the lower (respectively upper) interval bound for full-scan reconstruction being above that for short-scan reconstruction. Care has to be taken with the statistical interpretation of the results. To compare the results for the two reconstruction strategies against each other, recall that the testing used a fully paired design so that the 95% confidence interval (band) estimates obtained for each task are dependent. In this case, we can use the Bonferroni inequality to determine a lower bound on the joint coverage probability of the intervals for observer performance.[23(p. 232)] For arbitrary

TABLE V. Comparison of 95% confidence intervals estimated for observer performance on short-scan and full-scan reconstructions. The intervals were estimated from $n_1 = 136$ class-1 ratings and $n_2 = 136$ class-2 ratings.

	Short-scan	Full-scan
SNR	[1.2939 1.8377]	[1.7982 2.3905]
AUC	[0.8199 0.9031]	[0.8982 0.9545]
pAUC(0, 0.2)	[0.0935 0.1320]	[0.1294 0.1634]

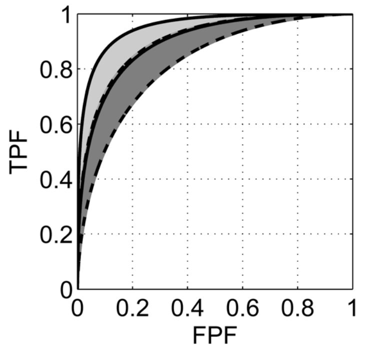

FIG. 4. Ninety-five percent confidence bands for the ROC curves corresponding to observer performance on short-scan and full-scan reconstructions. The band for short-scan reconstruction is shown in dark gray, delimited by dashed lines and the band for full-scan reconstruction is shown in light gray, delimited by solid lines. Note that the bands slightly overlap.

events A_1, A_2, …, A_n, the Bonferroni inequality takes the form [Ref. 4 (p. 13) Eq. (1.2.10), and Ref. 23]

$$P\left(\bigcap_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} P(A_i) - (n-1). \qquad (15)$$

Suppose, for example, that we wish to compare the SNR values obtained for the short-scan and full-scan reconstruction strategies. Let SNR_{ss} and SNR_{fs} be these two values, and let $[L_{ss},U_{ss}]$ and $[L_{fs},U_{fs}]$ be their 95% confidence intervals, respectively. Then by the Bonferroni inequality, the region $[L_{ss}, U_{ss}] \times [L_{fs}, U_{fs}]$ covers the pair (SNR_{ss}, SNR_{fs}) with a probability of at least $0.95 + 0.95 - 1 = 0.90$, i.e., with 90% confidence; see Fig. 6. Both the size and the position of the confidence region determine how the results should be interpreted. First, a smaller region covering a given pair of SNR values indicates a higher statistical precision. Second, if the confidence region does not intersect the line at 45° in the plane of possible values for (SNR_{ss}, SNR_{fs}), then there is evidence that the two tasks correspond to dissimilar detection performance. Going back to the example, because the SNR confidence intervals overlap, the SNR confidence region intersects the 45° line, as shown in Fig. 5, and there is not enough evidence at the 90% confidence level to reject the hypothesis that $\mathrm{SNR}_{ss} = \mathrm{SNR}_{fs}$. Likewise, the same conclusion can be made for the other figures of merit. Conversely, if the confidence region had instead not intersected the 45° line, there would have been evidence at the 90% confidence level to reject the hypothesis that $\mathrm{SNR}_{ss} = \mathrm{SNR}_{fs}$. Moreover, note, from the large size of the confidence region, that our conclusion is drawn with a fairly poor statistical precision.

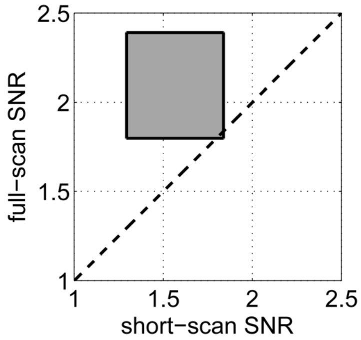

FIG. 5. Ninety percent confidence region for (SNR_{ss}, SNR_{fs}).

VI. DISCUSSION AND CONCLUSIONS

In this work, we proposed confidence interval estimators that may be used in ROC evaluations of task-based image quality studies employ in linear computerized observers defined with a fixed (nonrandom) template. All ratings provided by such an observer are linear combinations of image pixel values. A strength of the new interval estimators is that they have exactly known coverage probabilities. This property is particularly relevant for small sample sizes, since approximate confidence intervals have been found to be problematic in this case.[8] The price of our approach is a reliance on two assumptions, which are usually satisfied by ratings produced by linear computerized observers performing lesion detection tasks involving small, low-contrast lesions at a known location: (i) the observer ratings are normally distributed for each class of images, and (ii) the variance of the observer ratings is the same for each class of images. For x-ray CT, the normality assumption on the ratings is commonly accepted, and we used Monte Carlo simulations to investigate the reliability of the proposed AUC confidence intervals when the equal-variance assumption is violated for the rating data. In cases that are extreme for CT, we observed that the AUC interval estimators maintain highly accurate coverage probabilities. In addition, we discovered that the error in coverage probability was very small ($\leq 0.01\%$) when $n_1 = n_2$. When $n_1 \neq n_2$, we found that the error in coverage probability was about 0.2% in extreme cases, and that the confidence intervals may or may not be conservative, depending on the variance ratio of the ratings. These results demonstrate the practicality of our ROC confidence intervals for CT image quality evaluation.

When our assumptions for the observer ratings are strongly violated (which would be the case for human observer studies), the interval estimators introduced here are no longer appropriate. In this situation, it is preferable to use confidence intervals based on either nonparametric or semiparametric estimators, Ref. 3 (Chap. 5) which often rely on resampling techniques. Such approaches have the advantage that they do not require distributional assumptions on the observer ratings. However, construction of confidence intervals based on these estimators requires asymptotic assumptions that are violated for small samples.[8]

Note that our approach is equivalent to explicit utilization of a binormal model with *a priori* knowledge that the second binormal parameter (usually denoted as *b*) is equal to one. This knowledge was a critical component in our construction of confidence bands for the ROC curve with exactly known coverage probabilities. The ability to build such confidence bands is an attractive feature of our parametric approach. Another attractive feature is the ability to evaluate sample size effects without expensive Monte Carlo trials. As an example, we explored the potential decrease in statistical variability that can be gained by increasing the number of images from one class, which is an important issue, since it is often possible to get many additional class-1 images at low cost, as previously discussed in Ref. 14. Our results indicated that the mean AUC confidence interval length can shrink by as much as 35% when using this strategy.

Last, we illustrated the use of the new confidence interval estimators with an example involving a trained CHO applied to a lesion detection task with real x-ray CT images. This example demonstrated how different reconstruction methods can be compared with two-sided confidence intervals. Although it was not discussed in the example, our interval estimators can also be used to calculate one-sided confidence intervals, which occur when either α_1 or α_2 is zero.[17] One-sided intervals are useful for inferences involving statements such as "The AUC for method 1 is higher than the AUC for method 2." The MATLAB® routine supplied along with this article on the Medical Physics Website works for both two-sided and one-sided confidence intervals.

When designing an image evaluation study, it is sometimes possible to use the same data set for each scenario of interest. The design of the study is then called *paired* (as opposed to *unpaired*). The example we considered is compatible with pairing, and so we used a fully paired study design. Other comparisons, such as a study of the effect of different bowtie filters or dose-modulation strategies, would not allow pairing. In a paired situation, it is important to realize that the SNR estimates obtained for each reconstruction scenario are correlated and this must be taken into account when generating a joint confidence region for these estimates. We used the Bonferroni inequality to obtain a conservative rectangular confidence region. Another approach would have been to look for a nonrectangular confidence region. It is not clear that such a region can be found without assuming that the SNR estimates follow a joint multivariate normal distribution, which is only true for large sample sizes. In any case, such nonrectangular regions are difficult to visualize when more than two reconstruction scenarios have to be compared, unlike the rectangular Bonferroni-based regions. The interested reader will find a discussion on Bonferroni-based joint confidence regions and their attractive properties in Johnson and Wichern (Ref. 23, Sec. 5.4).

Because paired studies offer higher statistical power than unpaired studies, they should be considered whenever possible. The gain in statistical power results from two effects due to the pairing. First, more images are available to assess each scenario. Second, the pairing is likely to induce a positive correlation between the ratings associated to the scenarios under comparison. The Bonferroni-based confidence region approach makes full use of the first effect, but does not take advantage of large positive correlations in the ratings. One way to take advantage of large positive correlations between scenarios is to construct a nonrectangular confidence region, with the associated drawbacks discussed in the last paragraph. Another approach is to instead form a confidence interval for the difference of summary measures; see, e.g., Ref. 3 (Chap. 5). However, this approach typically requires an asymptotic normality assumption (which is not satisfied for small samples) to construct the confidence interval. In addition, the relative importance of an observed difference can be meaningfully interpreted only if the baseline is known, i.e, if the nominal value of one of the summary measures is known. As a concrete example, an observed gain in AUC value of 0.05 carries different meanings when the AUC value of the reference approach is 0.55 as opposed to 0.95; in the first case, the gain may seem marginal, whereas in the second case, it is as large as it could possibly be. For the example we used, the correlations between ratings can be shown to offer little benefit.

Finally, it should be emphasized that our choices for the task and for the observer in our image quality evaluation example were not optimal for image quality assessment in CT. There is large flexibility in the way that the task and the observer template may be defined. Investigation of more sophisticated tasks and observers suitable for CT images is an important topic for future research.

ACKNOWLEDGMENTS

This work was partially supported by NIH Grants R01 EB007236, R21 EB009168 and by a generous grant from the Ben B. and Iris M. Margolis Foundation. Its contents are solely the responsibility of the authors.

APPENDIX A: PROOF OF THEOREM 1

Here, we prove Theorem 1, which characterizes $\widehat{\mathrm{SNR}}$ when the rating data are normally distributed with equal variances for each class, i.e., $y_i^{(1)} \sim \mathcal{N}(\mu_1, \sigma^2)$ and $y_j^{(2)} \sim \mathcal{N}(\mu_2, \sigma^2)$, for $i = 1, 2, \ldots, n_1$ and $j = 1, 2, \ldots, n_2$.

Part 1. Since $\bar{y}_1 \sim \mathcal{N}(\mu_1, \sigma^2/n_1)$ and $\bar{y}_2 \sim \mathcal{N}(\mu_2, \sigma^2/n_2)$ are independent, it follows that

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}}\,\frac{(\bar{y}_2 - \bar{y}_1)}{\sigma} \sim \mathcal{N}\left(\mathrm{SNR}\sqrt{\frac{n_1 n_2}{n_1 + n_2}},\, 1\right). \tag{A1}$$

Also, since

$$\frac{(n_1 - 1)}{\sigma^2} s_1^2 \sim \chi^2_{n_1 - 1} \quad \text{and} \quad \frac{(n_2 - 1)}{\sigma^2} s_2^2 \sim \chi^2_{n_2 - 1} \tag{A2}$$

are independent, we have

$$\frac{(n_1 + n_2 - 2)}{\sigma^2} s^2 \sim \chi^2_{n_1 + n_2 - 2}. \tag{A3}$$

Thus, Eqs. (A1) and (A3) imply that

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}}\,\frac{(\bar{y}_2 - \bar{y}_1)}{s} \sim t'_\nu(\delta), \tag{A4}$$

where $\nu = n_1 + n_2 - 2$ and $\delta = \mathrm{SNR}\sqrt{n_1 n_2/(n_1 + n_2)}$. The desired relation then follows from the definitions of η and $\widehat{\mathrm{SNR}}$.

Part 2. From Part 1 and the expression for the mean of a noncentral t random variable given in Eq. (2), it is easy to see that $\widehat{\mathrm{SNR}}$ is an unbiased estimator of SNR. The joint pdf of the rating data is

$$f(y_1^{(1)}, y_2^{(1)}, \ldots, y_{n_1}^{(1)}, y_1^{(2)}, y_2^{(2)}, \ldots, y_{n_2}^{(2)}) = (2\pi)^{-(n_1+n_2)/2}\sigma^{-(n_1+n_2)} \times \exp[-1/(2\sigma^2)(\sum_{i=1}^{n_1}(y_i^{(1)} - \mu_1)^2 + \sum_{j=1}^{n_2}(y_j^{(2)} - \mu_2)^2)]. \tag{A5}$$

After some algebra, one may show that

$$\sum_{i=1}^{n_r}\left(y_i^{(r)}-\mu_r\right)^2=(n_r-1)s_r^2+n_r\bar{y}_r^2-2n_r\bar{y}_r\mu_r+n_r\mu_r^2, \tag{A6}$$

for $r=1$, 2. Using Eq. (A6) and the definition of s^2, we may rewrite Eq. (A5) in the form

$$\begin{aligned} f(y_1^{(1)},y_2^{(1)},\ldots,y_{n_1}^{(1)},y_1^{(2)},y_2^{(2)},\ldots,y_{n_2}^{(2)}) &\\ =(2\pi)^{-(n_1+n_2)/2}\sigma^{-(n_1+n_2)} &\\ \times\exp\{-1/(2\sigma^2)[n_1\mu_1^2+n_2\mu_2^2]\} &\\ \times\exp\{-1/(2\sigma^2)[(n_1+n_2-2)s^2+n_1\bar{y}_1^2+n_2\bar{y}_2^2]\} &\\ \times\exp\{(1/\sigma^2)[n_1\mu_1\bar{y}_1+n_2\mu_2\bar{y}_2]\}. & \end{aligned} \tag{A7}$$

By the Fisher–Neyman factorization theorem [Ref. 24, Theorem 6.5 (p. 35)] the statistic

$$W=[(n_1+n_2-2)s^2+n_1\bar{y}_1^2+n_2\bar{y}_2^2,\ \bar{y}_1,\ \bar{y}_2] \tag{A8}$$

is sufficient. Moreover, because the expression in Eq. (A7) has the form of a full rank exponential family,[24(p. 23,24)] W is a complete statistic [Ref. 24, Theorem 6.22 (p. 42)]. Since (i) W is a complete sufficient statistic, (ii) $\widehat{\text{SNR}}$ is an unbiased estimator of SNR, and (iii) $\widehat{\text{SNR}}=\text{E}[\widehat{\text{SNR}}|W]$, i.e., $\widehat{\text{SNR}}$ is a function of W only, the Lehmann–Scheffé theorem [Ref. 24, Theorem 1.11 (p. 88) and Ref. 25 (p. 164)] implies that $\widehat{\text{SNR}}$ is the unique UMVU estimator of SNR.

APPENDIX B: PROOF OF THEOREM 2

Next, we prove Theorem 2 and Corollary 1, which enable us to calculate our ROC confidence intervals. For this task, we need the following lemmas.

Lemma 1. Suppose that $T\sim t'_\nu(\delta)$. Then at arbitrary fixed values of t and ν, the cdf of T, $F_T(t;\,\nu,\delta)$, is a continuous, strictly decreasing function of δ.

Proof. Although this lemma seems like a property that should be well-known, we could not find any proof of it in the literature. One way to prove it is as follows.

From Ref. 15 (p. 514), the cdf for the noncentral t distribution may be written in the form

$$F_T(t;\,\nu,\delta)=\frac{1}{2^{(\nu/2)-1}\Gamma(\nu/2)}\times\int_0^\infty x^{\nu-1}e^{-x^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{(tx/\sqrt{\nu})-\delta}e^{-u^2/2}du\,dx. \tag{B1}$$

Suppose that t and ν are fixed quantities and define $h(\delta)=F_T(t;\,\nu,\delta)$. Making the change of variables $y=u-tx/\sqrt{\nu}$ in the inner integral of Eq. (B1) yields

$$h(\delta)=\frac{1}{\sqrt{\pi}\,2^{(\nu-1)/2}\Gamma(\nu/2)}\times\int_0^\infty\int_{-\infty}^{-\delta}x^{\nu-1}e^{-x^2/2}e^{-\frac{1}{2}(y+tx/\sqrt{\nu})^2}dy\,dx. \tag{B2}$$

By the combination of Tonelli's theorem and Fubini's theorem,[26(Chap. 8)] an interchange in the order of integration is justified, and the previous equation may be rewritten as

$$h(\delta)=\frac{1}{\sqrt{\pi}\,2^{(\nu-1)/2}\Gamma(\nu/2)}\int_{-\infty}^{-\delta}g(y)\,dy, \tag{B3}$$

where

$$g(y)=\int_0^\infty x^{\nu-1}e^{-x^2/2}e^{-\frac{1}{2}(y+tx/\sqrt{\nu})^2}dx \tag{B4}$$

is an integrable function of y. The theorem on absolute continuity for the Lebesgue integral[26(p. 141)] applied to Eq. (B3) implies that $h(\delta)$ is continuous. In addition, since $g(y)$ is strictly positive, Eq. (B3) indicates that $h(\delta)$ is a strictly decreasing function of δ.

Lemma 2. Let X be a continuous random variable with cdf, $F_X(x;\,\theta)$, that is a strictly decreasing function of the parameter θ for each x. Also, let $\alpha_1,\alpha_2\in(0,1)$ be such that $\alpha_1+\alpha_2=\alpha$ for some $\alpha\in(0,1)$. Suppose that, for each x in the sample space of X, the relations

$$F_X(x;\,\theta_L(x))=1-\alpha_1 \quad\text{and}\quad F_X(x;\,\theta_U(x))=\alpha_2$$

may be solved for $\theta_L(x)$ and $\theta_U(x)$. Then the functions $\theta_L(x)$ and $\theta_U(x)$ are uniquely defined and the random interval $[\theta_L(X),\ \theta_U(X)]$ is an exact $1-\alpha$ confidence interval for θ.

Proof. See Ref. 4 [Theorem 9.2.12(p. 432)] for a proof and Ref. 17 (Sec. 11.4) for a complementary discussion.

Finally, we state a lemma that facilitates construction of a confidence interval for any parameter that is related to another through a strictly increasing transformation. It is a well-known property of confidence intervals that, as observed in Ref. 5, is rarely formalized.

Lemma 3. Let $g(\theta)$ be a continuous, strictly increasing function of θ. If $[\theta_L,\ \theta_U]$ is a $1-\alpha$ confidence interval for θ, then $[g(\theta_L),\ g(\theta_U)]$ is a $1-\alpha$ confidence interval for $g(\theta)$.

Proof. The assumptions on g imply that g^{-1} exists and is strictly increasing. Because both g and g^{-1} are strictly increasing functions, it follows that $\theta\in[\theta_L,\ \theta_U]$ if and only if $g(\theta)\in[g(\theta_L),\ g(\theta_U)]$, i.e., the two events are equivalent. Hence, $P(g(\theta)\in[g(\theta_L),g(\theta_U)])=P(\theta\in[\theta_L,\theta_U])=1-\alpha$.

Combining the above results, it is straightforward to see that Theorem 2 follows from Theorem 1 and Lemmas 1, 2, and 3. Furthermore, it follows from our distributional assumptions that TPF, AUC, and pAUC are strictly increasing functions of SNR, and therefore, Theorem 2(ii) and Lemma 3 imply Corollary 1.

APPENDIX C: PROOF OF THEOREM 3

Here, we prove Theorem 3, which enables estimation of a confidence band for the entire ROC curve from a confidence interval for SNR.

For any fixed value of FPF $\in[0,1]$, define the function $g(\text{SNR})=\text{TPF}(\text{FPF};\ \text{SNR})$, where TPF is given by Eq. (3). Since $g(\text{SNR})$ is a continuous, strictly increasing function of SNR, its inverse, $g^{-1}(\text{TPF})$, exists and is a strictly increasing function of TPF. Therefore, SNR $\in[\text{SNR}_L,\ \text{SNR}_U]$ if and only if $g(\text{SNR})\in[g(\text{SNR}_L),\ g(\text{SNR}_U)]$. Because this is true for any FPF $\in[0,1]$, it follows that SNR $\in[\text{SNR}_L,\ \text{SNR}_U]$ if and only if $\Omega_{\text{ROC}}\subset\widehat{\Omega}_{\text{ROC}}$. Hence, $P(\Omega_{\text{ROC}}\subset\widehat{\Omega}_{\text{ROC}})=P(\text{SNR}\in[\text{SNR}_L,\ \text{SNR}_U])=1-\alpha$ for any value of SNR.

[a]Electronic mail: awunder@ucair.med.utah.edu
[b]Electronic mail: noo@ucair.med.utah.edu

[1] H. H. Barrett and K. J. Myers, *Foundations of Image Science* (John Wiley & Son, New York, 2004).

[2] S. Park, R. Jennings, H. Liu, A. Badano, and K. Myers, "A statistical, task-based evaluation method for three-dimensional x-ray breast imaging systems using variable-background phantoms," Med. Phys. **37**(12), 6253–6270 (2010).

[3] M. S. Pepe, The Statistical Evaluation of Medical Tests for Classification and Prediction (Oxford University Press, New York, 2003).

[4] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. (Duxbury, Belmont, CA, 2001).

[5] J. H. Steiger and R. T. Fouladi, "Noncentrality interval estimation and the evaluation of statistical models," in *What if There Were No Significance Tests?* edited by L. L. Harlow, S. A. Mulaik, and J. H. Steiger (Lawrence Erlbaum, Mahwah, NJ, 1997).

[6] C. E. Metz, "Quantification of failure to demonstrate statistical significance: The usefulness of confidence intervals," Invest. Radiol. **28**(1), 59–63 (1993).

[7] D. Bamber, "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," J. Math. Psychol. **12**, 387–415 (1975).

[8] N. A. Obuchowski and M. L. Lieber, "Confidence intervals for the receiver operating characteristic area in studies with small samples," Acad. Radiol. **5**(8), 561–571 (1998).

[9] R. G. Newcombe, "Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2: Asymptotic methods and evaluation," Stat. Med. **25**, 559–573 (2006).

[10] G. Ma and W. Hall, "Confidence bands for receiver operating characteristic curves," Med. Decis Making **13**(3), 191–197 (1993).

[11] S. A. Macskassy, F. Provost, and S. Rosset, "ROC confidence bands: An empirical evaluation," in *Proceedings of 22nd International Conference on Machine Learning* Bonn, Germany, 2005, pp. 537–544.

[12] H. Godwin and S. Zaremba, "A central limit theorem for partly dependent variables," Ann. Math. Stat. **32**(3), 677–686 (1961).

[13] J. Wang, H. Lu, Z. Liang, D. Eremina, G. Zhang, S. Wang, J. Chen, and J. Manzione, "An experimental study on the noise properties of x-ray CT sinogram data in Radon space," Phys. Med. Biol. **53**(12), 3327–3341 (2008).

[14] R. M. Gagne, B. D. Gallas, and K. J. Myers, "Toward objective and quantitative evaluation of imaging systems using images of phantoms," Med. Phys. **33**(1), 83–95 (2006).

[15] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, 2nd ed. (John Wiley & Son, New York, 1995), Vol. 2.

[16] W. C. Lee and C. K. Hsiao, "Alternative summary indices for the receiver operating characteristic curve," Epidemiology **7**(6), 605–611 (1996).

[17] L. J. Bain and M. Engelhardt, Introduction to Probability and Mathematical Statistics, 2nd ed. (Duxbury, 1992).

[18] A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging, Series. Classics in Applied Mathematics* (SIAM, Philadelphia, PA, 2001), Vol. **13**.

[19] A. Wunderlich and F. Noo, "Image covariance and lesion detectability in direct fan-beam x-ray computed tomography," Phys. Med. Biol **53**(10), 2471–2493 (2008).

[20] A. Agresti and B. A. Coull, "Approximate is better than "exact," for interval estimation of binomial proportions," Am. Stat. **52**(2), 119–126 (1998).

[21] QRM GmbH, http://www.qrm.de/. Last accessed May, 2011.

[22] H. Lilliefors, "On the Kolmogorov-Smirnov test for normality with mean and variance unknown," J. Am. Stat. Assoc. **62**(318), 399–402 (1967).

[23] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 5th ed. (Prentice-Hall, Englewood Cliffs, NJ, 2002).

[24] E. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. (Springer, New York, 1998).

[25] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. (Springer, New York, 1994).

[26] F. Jones, *Lebesgue Integration on Euclidean Space*, Revised ed. (Jones and Bartlett, Sudbury, MA, 2001).