HHS Public Access

Author manuscript

IEEE Trans Med Imaging. Author manuscript; available in PMC 2017 August 03.

Published in final edited form as:

IEEE Trans Med Imaging. 2015 February ; 34(2): 453–464. doi:10.1109/TMI.2014.2360496.

Exact Confidence Intervals for Channelized Hotelling Observer Performance in Image Quality Studies

Adam Wunderlich,
Department of Radiology, University of Utah, Salt Lake City, UT 84108 USA

Frédéric Noo [Member, IEEE],
Department of Radiology, University of Utah, Salt Lake City, UT 84108 USA

Brandon D. Gallas, and
Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, MD 20993 USA

Marta E. Heilbrun
Department of Radiology, University of Utah, Salt Lake City, UT 84108 USA

Abstract

Task-based assessments of image quality constitute a rigorous, principled approach to the evaluation of imaging system performance. To conduct such assessments, it has been recognized that mathematical model observers are very useful, particularly for purposes of imaging system development and optimization. One type of model observer that has been widely applied in the medical imaging community is the channelized Hotelling observer (CHO), which is well-suited to known-location discrimination tasks. In the present work, we address the need for reliable confidence interval estimators of CHO performance. Specifically, we show that the bias associated with point estimates of CHO performance can be overcome by using confidence intervals proposed by Reiser for the Mahalanobis distance. In addition, we find that these intervals are well-defined with theoretically-exact coverage probabilities, which is a new result not proved by Reiser. The confidence intervals are tested with Monte Carlo simulation and demonstrated with two examples comparing X-ray CT reconstruction strategies. Moreover, commonly-used training/testing approaches are discussed and compared to the exact confidence intervals. MATLAB software implementing the estimators described in this work is publicly available at http://code.google.com/p/iqmodelo/.

Index Terms

Image quality assessment; linear discriminant analysis (LDA); Mahalanobis distance; model observers; noncentral F-distribution; noncentrality parameter

Correspondence to: Adam Wunderlich.

A. Wunderlich is now with the Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, MD 20993 USA

I. Introduction

Objective, task-based image quality assessments with mathematical model observers are a valuable tool for medical imaging system development, optimization, and characterization [1]–[3]. One type of model observer that has drawn considerable interest in the medical imaging community is the channelized Hotelling observer (CHO) [1]–[3], which is well-suited to known-location binary discrimination tasks. The CHO is the population equivalent of the classical Fisher linear discriminant function used in multivariate statistics and pattern recognition [4], [5] applied to a vector of image features, called channel outputs in the medical imaging and vision literature. Through various choices of channels, which determine relevant features and are used for dimensionality reduction, CHOs have been shown to track both human [6]–[11] and ideal linear [12] observer performance for known-location discrimination tasks. As a consequence, CHO methodology has been widely employed in medical imaging research, e.g., [13]–[25].

In some settings, CHO performance can be calculated accurately from analytical models [13], [15], [21]. However, due to complexities in the physics of data acquisition and in image formation, this approach is usually not feasible, and most practical evaluations instead estimate CHO performance from a finite set of images. Furthermore, the number of images available is typically limited for both real and simulated data sets by the available resources (acquisition time, computational cost, etc.), so that statistical variability in performance estimates cannot be neglected when making inferences.

One way to assess statistical variability in experimental results is through the use of confidence intervals. In contrast to traditional hypothesis testing, which is limited to only testing statistical significance, confidence intervals communicate statistical precision and effect size, in addition to statistical significance. These virtues make confidence intervals an attractive option for the presentation of experimental findings.

Typically, confidence intervals are constructed by starting with a point estimate. Two general approaches can be used to obtain point estimates of CHO performance from a set of images. In the first approach, all images are used to directly estimate the figure of merit by substituting sample estimates for population parameters [1, p. 972]. In the second approach, the images are split into two subsets, where one subset is used to estimate the observer template (i.e., train the observer), and the second subset is used to estimate performance of the trained observer (i.e., test the observer) [1, p. 973]. For the goal of estimating the performance of a CHO defined by population parameters (i.e., infinitely-trained), both approaches result in biased estimates. In this work, we adopt the first (direct) approach, and demonstrate how the point estimate bias can be overcome to obtain accurate confidence intervals for CHO performance. Later, we present a comparison of our confidence intervals with estimators based on training and testing.

This paper addresses the need for reliable confidence interval estimators for CHO performance. Specifically, we find that an approach introduced by Reiser [26] for interval estimation of the Mahalanobis distance can be applied to obtain confidence intervals for CHO performance. In addition, we prove that these intervals have theoretically-exact

coverage probabilities, which is a new result not given by Reiser [26]. The exact confidence intervals are further evaluated with Monte Carlo simulation and are found to be superior to conventional Wald-style intervals [27, p. 499]. The application of the exact confidence intervals is demonstrated with examples comparing X-ray CT reconstruction strategies.

Our theoretical results rely on two assumptions: 1) the channel outputs follow a multivariate normal distribution for each image class, and 2) the covariance matrices for the channel outputs are the same for each class. These assumptions are typically well-justified for fixed-location discrimination tasks involving either a flat or a normally-distributed, variable background. Likewise, they can be employed when the measurement noise depends on the object and the mean difference between the two classes is small; in this case, the covariance matrices can be assumed to be the same.

More specifically, the first assumption is generally satisfied for reconstructed tomographic images, which are often approximately multivariate normal. Additionally, even for images that are not normally distributed, the central limit theorem implies that the channel outputs will tend to be normally distributed. For nuclear medicine applications, a strong argument justifying the normality assumption was provided by Khurd and Gindi [28]. In X-ray CT, the normality assumption was supported by Zeng *et al.* [29] with histogram plots, and by Wunderlich *et al.* in [30] using a univariate test and in [31] with a multivariate test. An argument in support of the second assumption for nuclear medicine was given by Barrett and Myers [1, p. 1209], and a quantitative analysis was presented by Wunderlich and Noo [30] in the context of X-ray CT.

In previous publications, we have presented related estimators for performance of linear observers [30], [32] and CHOs [33], [34]. To put the present work into proper perspective, it is helpful to clarify its relationship with these previous investigations. First, the estimators given in [30], [32] are designed for any linear observer defined by a fixed, known template, i.e., a classifier that is ready to be used in practice. Consequently, the estimators in [30], [32] are suitable for evaluations of finitely-trained observer performance when the template is determined by a fixed training set, or when the template is prespecified. By contrast, in this work and in [33], [34], the aim is to estimate performance of a CHO, which has a template that depends on unknown population parameters. Compared to [33], [34], in which the difference of class means is assumed to be known, the present work does not assume that any population parameters are known. Hence, the results given here can be seen as a complementary piece of a larger theory of model observer performance estimation. Further discussion of the relationships between the present work and other approaches is given at the end of the paper.

II. Background

In this section, we establish our notation by reviewing channelized Hotelling observers and associated performance measures.

One important component of a CHO is a set of channels. Channels are often used by model observers to reduce high-dimensional image data to a smaller number of relevant features.

Each channel is itself an image that corresponds to a feature, and the scalar product of a channel with an image yields a channel output. We write the image as a $q \times 1$ column vector, $\mathbf{g}$, and denote the number of channels by p, where p is typically much smaller than q. The weights defining each channel are collected into a column of a $q \times p$ channel matrix, U, which is applied to each image to obtain a $p \times 1$ channel output vector, $\mathbf{v}$, where $\mathbf{v} = U^T\mathbf{g}$. General considerations regarding the choice of channels are beyond the scope of this paper. The reader is referred to [1, pp. 936–937] and [3] for surveys of the literature on different channel models. In Section V, we present two CHO examples using Gabor channels, and provide an explanation of their essential characteristics. Note that if no channels are used, then U is equal to the identity matrix, i.e., $U = I$ with $p = q$.

A CHO is designed for a binary discrimination task in which the goal is to classify each image as belonging to one of two classes, denoted as class 1 and class 2. For example, in the case of lesion detection, the two classes correspond to lesion-absent and lesion-present, respectively. To classify an image with a corresponding channel output vector $\mathbf{v}$, a CHO starts by generating a rating statistic, $t = \mathbf{w}^T\mathbf{v}$, where $\mathbf{w}$ is the $p \times 1$ CHO template (defined below). Next, the rating statistic is compared to a threshold, c. If $t \geq c$, then the image is classified as belonging to class 2, otherwise, the image is classified as belonging to class 1 [1].

We denote the means of the channel output vector, $\mathbf{v}$, for classes 1 and 2 as $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively, and their difference as $\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. In addition, let the (nonsingular) covariance matrices of $\mathbf{v}$ for class 1 and class 2 be Σ_1 and Σ_2, respectively, and write their average as $\overline{\Sigma} = (\Sigma_1 + \Sigma_2)/2$. In this notation, the template for a CHO is defined as $\mathbf{w} = \overline{\Sigma}^{-1}\Delta\boldsymbol{\mu}$ [1]. Throughout this paper, it is assumed that $\Sigma_1 = \Sigma_2 = \Sigma$, so that the CHO template takes the simplified form $\mathbf{w} = \Sigma^{-1}\Delta\boldsymbol{\mu}$.

For a binary classification task, an observer's performance is completely characterized by its receiver operating characteristic (ROC) curve, which plots true positive fraction (TPF) versus false positive fraction (FPF) over all decision thresholds [1], [35]. (In the medical literature, TPF and 1 − FPF are known as sensitivity and specificity, respectively.) A commonly used summary figure of merit for observer performance is the area under the ROC curve, denoted as AUC. If the observer's rating statistic, t, is normally distributed for each image class, then AUC can be expressed as $\mathrm{AUC} = \Phi(\mathrm{SNR}/\sqrt{2})$, where $\Phi(x)$ is the cumulative distribution function (cdf) for the standard normal distribution, and SNR is the observer signal-to-noise ratio, defined as the difference of class means for t divided by the pooled standard deviation [1, p. 819], [35, pp. 83–84]. Above, because SNR and AUC are linked by a monotonic transformation, SNR is also useful as a figure of merit for observer performance [1, p. 819].

As discussed in the introduction, this work assumes that the channel output vector has a multivariate normal distribution for each class with a common covariance matrix. To express this assumption symbolically, we introduce the following notation. Let a $p \times 1$ random vector $\mathbf{x} \in \mathbb{R}^p$ that follows a nondegenerate, multivariate normal distribution with mean, $\boldsymbol{\mu}$, and positive-definite covariance matrix, Σ, be denoted as $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$. Writing the channel

output vectors for classes 1 and 2 as $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$, respectively, our distributional assumptions can be expressed compactly as $\mathbf{v}^{(1)} \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \Sigma)$ and $\mathbf{v}^{(2)} \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \Sigma)$.

The above distributional assumptions have three important implications. First, the CHO is optimal among all observers that operate on the channel output vector, in the sense that it maximizes SNR^2. In fact, if no channels are used, then the resulting observer coincides with the so-called "ideal observer" [1, p. 851]. Second, the observer rating statistic, t, is normally distributed under each class, so that $\text{AUC}=\Phi(\text{SNR}/\sqrt{2})$, i.e., SNR and AUC are linked by a monotonic transformation. Third, the expression for observer SNR^2 simplifies to $\text{SNR}^2 = \Delta\boldsymbol{\mu}^T\Sigma^{-1}\Delta\boldsymbol{\mu}$ [1, p. 967]. This last formula is our starting point in the next section. In the statistics literature, $\text{SNR}=\sqrt{\Delta\boldsymbol{\mu}^T\sum^{-1}\Delta\boldsymbol{\mu}}$ is also known as the Mahalanobis distance [4]. Note that since Σ is positive-definite, it follows that if $\Delta\boldsymbol{\mu}$ is nonzero, which is generally true for image-quality evaluations, then SNR > 0 and 0.5 < AUC < 1.

III. Exact Confidence Intervals

Consider a random variable, X, with a distribution depending on a nonrandom parameter, θ. A random interval estimate $[\theta_L(X), \theta_U(X)]$ for θ is said to be a $1-\alpha$ confidence interval [27] if for any θ in the parameter space, the probability that the interval covers θ is $1-\alpha$, i.e., $P(\theta \in [\theta_L(X), \theta_U(X)]) = 1-\alpha$. The value $1-\alpha$ is called the coverage probability for the confidence interval. A confidence interval is said to be exact if the coverage probability equation is exactly satisfied. Otherwise, a confidence interval is said to be approximate.

In this section, we construct exact SNR and AUC confidence intervals for a CHO. The construction is based on an approach first suggested by Reiser [26] for interval estimation of the Mahalanobis distance. Although Reiser stated that his intervals were approximate, we prove that they are, in fact, exact as long as SNR > 0, which is typically the case for image quality evaluations. We start by introducing an SNR^2 point estimator.

A. SNR^2 Point Estimation

Suppose that we are given m independent, identically distributed (i.i.d.) measurements of the class-1 channel output vector, $\mathbf{v}_1^{(1)}, \mathbf{v}_2^{(1)}, \ldots, \mathbf{v}_m^{(1)}$, and n i.i.d. measurements of the class-2 channel output vector, $\mathbf{v}_1^{(2)}, \mathbf{v}_2^{(2)}, \ldots, \mathbf{v}_n^{(2)}$. Write the sample means of the channel output vectors for classes 1 and 2 as $\overline{\mathbf{v}}_1=(1/m)\sum_{i=1}^{m}\mathbf{v}_i^{(1)}$ and $\overline{\mathbf{v}}_2=(1/n)\sum_{j=1}^{n}\mathbf{v}_j^{(2)}$, respectively, and their difference as $\Delta\overline{\mathbf{v}} = \overline{\mathbf{v}}_2 - \overline{\mathbf{v}}_1$. Also, define a pooled estimate of the channel output covariance matrix, Σ, as

$$S=\frac{1}{m+n-2}\left[\sum_{i=1}^{m}\left(\mathbf{v}_i^{(1)}-\overline{\mathbf{v}}_1\right)\left(\mathbf{v}_i^{(1)}-\overline{\mathbf{v}}_1\right)^T+\sum_{j=1}^{n}\left(\mathbf{v}_j^{(2)}-\overline{\mathbf{v}}_2\right)\left(\mathbf{v}_j^{(2)}-\overline{\mathbf{v}}_2\right)^T\right]. \quad (1)$$

Recall that under our distributional assumptions, $\text{SNR}^2 = \Delta\boldsymbol{\mu}^T\Sigma^{-1}\Delta\boldsymbol{\mu}$. Substituting the above sample estimates for $\Delta\boldsymbol{\mu}$ and Σ into this formula results in a direct, plug-in point estimate

given by $\hat{\theta}_p = \Delta\bar{\mathbf{v}}^T S^{-1}\Delta\bar{\mathbf{v}}$, where the subscript "p" stands for "plug-in." Since S is nonsingular with probability one if $m + n > p + 1$ [4, Th. 3.1.4, p. 82], it follows that $\hat{\theta}_p$ is well-defined as long as $m + n > p + 1$. Note that $\hat{\theta}_p$ is equivalent to the maximum likelihood estimator (MLE) when multiplied by $(m+n)/(m + n - 2)$.

It turns out that a multiplicative factor can be utilized to reduce the bias and variance of $\hat{\theta}_p$. Namely, let

$$\hat{\theta} = \gamma \Delta\bar{\mathbf{v}}^T S^{-1} \Delta\bar{\mathbf{v}} \quad (2)$$

where γ is a positive function of m, n, and p. Below, we derive an attractive choice for γ.

Under our distributional assumptions, the sampling distribution for $\hat{\theta}$ is closely related to the noncentral F-distribution, which arises as the ratio of a noncentral χ^2 random variable to an independent, central χ^2 random variable [4]. We denote a random variable, X, following a noncentral F-distribution with degrees of freedom ν_1 and ν_2, and noncentrality parameter, δ, with standard notation as $X \sim F'_{\nu_1,\nu_2}(\delta)$. The following theorem characterizes the sampling distribution of $\hat{\theta}$.

Theorem 1—Suppose that $\mathbf{v}_i^{(1)} \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \sum)$ for $i = 1, 2, \ldots, m$ and $\mathbf{v}_j^{(2)} \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \sum)$ for $j = 1, 2, \ldots, n$ are independent. If $m + n > p + 1$ and γ is positive, then

$$\frac{(m+n-p-1)(mn)}{p(m+n-2)(m+n)\gamma}\hat{\theta} \sim F'_{\nu_1,\nu_2}(\delta)$$

with $\nu_1 = p$, $\nu_2 = m+n-p-1$, and $\delta = \mathrm{SNR}^2(mn)/(m + n)$.

Proof: The result follows from the distribution of the two-sample Hotelling T^2 statistic [4, Th. 3.2.13, p. 98].

We can draw three interesting observations from the above theorem that have implications for the confidence intervals presented in the next subsection. First, note that Theorem 1 holds independent of the choice of γ, since this factor also appears in $\hat{\theta}$ and cancels out of the quantity on the left side. Second, observe that the distribution of $\hat{\theta}$ depends on four parameters: m, n, p and SNR. Since m, n, and p are fixed for a given experiment, the only unknown parameter is SNR. Third, for fixed m and n, the noncentrality parameter, δ, is proportional to SNR^2. In the next subsection, this fact is utilized to construct confidence intervals for SNR.

Applying classical formulae for the mean and variance of the noncentral F-distribution [36], it is straightforward to obtain expressions for the mean and variance of $\hat{\theta}$. These expressions are collected in the corollary below.

Corollary 1: Suppose that the hypotheses of Theorem 1 are satisfied. If $m + n > p + 5$, then

$$\mathrm{E}[\hat{\theta}]=\gamma\left(\frac{m+n-2}{m+n-p-3}\right)\left[\left(\frac{m+n}{\mathrm{mn}}\right)p+\mathrm{SNR}^2\right]$$

$$\begin{aligned}\mathrm{Var}[\hat{\theta}]=&2\gamma^2\left(\frac{m+n-2}{m+n-p-3}\right)^2\left(\frac{m+n}{\mathrm{mn}}\right)^2\\&\times\frac{1}{(m+n-p-5)}\left[\left(p+\frac{\mathrm{mn}}{m+n}\mathrm{SNR}^2\right)^2\right.\\&\left.+\left(p+2\frac{\mathrm{mn}}{m+n}\mathrm{SNR}^2\right)(m+n-p-3)\right].\end{aligned}$$

From the above expressions, it can be seen that by choosing

$$\gamma=\frac{m+n-p-3}{m+n-2}, \quad (3)$$

the multiplicative bias is removed, and the overall bias is reduced, as compared to $\gamma = 1$. Namely, the bias, $\mathrm{E}[\hat{\theta}] - \mathrm{SNR}^2$, becomes $(m + n)p/(mn)$, which increases with the number of channels, p, and decreases with the number of samples, $m + n$. Furthermore, the above choice of γ is seen to reduce the variance by a factor of $[(m + n - p - 3)/(m + n - 2)]^2$. Observe that since γ is strictly positive if $m + n > p + 3$, it follows that $\hat{\theta}$ is well-defined as long as $m + n > p + 3$. The multiplicative constant, γ, defined in (3) is an original contribution of this work; we use this definition throughout the remainder of the paper. However, the reader should note that since the distribution in Theorem 1 does not depend on γ, the exact confidence intervals defined in the next section are independent of γ.

Given the above observations, it is evident that an unbiased point estimate of SNR^2 can be obtained by subtracting the bias, $(m + n)p/(mn)$, from $\hat{\theta}$ with γ given by (3). However, such bias subtraction is problematic, since it shifts the distribution of point estimates, and leads to potentially negative estimates of SNR^2. Because conversion of SNR^2 to SNR and AUC requires a square root operation, a negative SNR^2 estimate cannot be converted to a real-valued estimate of SNR or AUC, and its interpretation as a figure of merit is unclear. Moreover, truncation of a negative estimate at zero comes with another complication, since this operation will alter the statistical properties of the estimator and introduce a bias. This situation is similar to the possibility of negative estimates of variance components in the analysis of variance (ANOVA) when using unbiased estimators [37]. For the reasons mentioned above, we do not pursue an additional bias reduction through subtraction in our point estimates of SNR^2.

B. Reiser Intervals and Main Result

Under our assumptions on the channel output vector

$$X=\frac{(m+n-p-1)(\mathrm{mn})}{p(m+n-2)(m+n)\gamma}\hat{\theta} \quad (4)$$

follows a noncentral F-distribution by Theorem 1. Moreover, the noncentrality parameter, δ, for this distribution is a strictly increasing function of SNR, and hence, AUC. As a result of these relationships, confidence intervals for SNR and AUC can be obtained from a confidence interval for the noncentrality parameter. To find a $1-\alpha$ confidence interval for δ, we use an approach suggested by Reiser [26], as described below.

Denote the cdf for X as $F_X(x;\ \nu_1,\ \nu_2,\ \delta)$ and suppose that $\alpha_1,\ \alpha_2 \in (0, 1)$ are fixed numbers such that $\alpha_1 + \alpha_2 = \alpha$ for some $\alpha \in (0, 1)$. (For typical two-sided intervals, $\alpha_1 = \alpha_2 = \alpha/2$.) Given an observation x of X, functions $\delta_L(x)$ and $\delta_U(x)$ are defined as follows, using the key property that the cdf for the noncentral F-distribution is a continuous, strictly decreasing function of δ; see Lemma 1 in the Appendix. First, if $F_X(x;\ \nu_1,\ \nu_2,\ 0) > 1-\alpha_1$, then $\delta_L(x)$ is defined as the implicit solution of the equation

$$F_X(x;\nu_1,\nu_2,\delta_L(x))=1-\alpha_1. \quad (5)$$

Otherwise, $\delta_L(x)$ is defined to be zero. Similarly, if $F_X(x;\ \nu_1,\ \nu_2,\ 0) > \alpha_2$, then $\delta_U(x)$ is defined as the implicit solution of

$$F_X(x;\nu_1,\nu_2,\delta_U(x))=\alpha_2. \quad (6)$$

Otherwise, $\delta_U(x)$ is defined to be zero.

From Lemma 1 in the Appendix, it follows that $\delta_L(x)$ and $\delta_U(x)$ are well-defined, which is a fact not proved by Reiser [26]. Because the noncentral F cdf is a built-in library function in many statistical software packages, $\delta_L(x)$ and $\delta_U(x)$ can be easily computed by iteratively solving the implicit equations in (5) and (6) with standard root-finding algorithms.

The functions $\delta_L(X)$ and $\delta_U(X)$ given above are the lower and upper endpoints, respectively, of a nominal $1-\alpha$ confidence interval for the noncentrality parameter, δ. From its definition it can be observed that the confidence interval for δ can potentially be [0, 0] when $\hat{\theta}$ is very small. When this happens, the result should be interpreted to mean that zero is the best estimate of δ.

The interval $[\delta_L(X),\ \delta_U(X)]$ for δ can be transformed into a confidence interval for SNR through the mapping given in Theorem 1. Namely, lower and upper endpoints for an SNR interval are given by $\mathrm{SNR}_L(\hat{\theta})=\sqrt{\delta_L(X)(m+n)/(\mathrm{mn})}$ and $\mathrm{SNR}_U(\hat{\theta})=\sqrt{\delta_U(X)(m+n)/(\mathrm{mn})}$, respectively. The functional dependence of SNR_L and SNR_U on $\hat{\theta}$ comes from the fact that X is a function of $\hat{\theta}$ (here, m, n, and p are fixed). Likewise, a confidence interval for AUC can be obtained by applying the transformation

$\Phi(\mathrm{SNR}/\sqrt{2})$ to the SNR interval endpoints. We call the above confidence intervals for δ, SNR, and AUC *Reiser intervals*, since they were first defined in [26].

From heuristic reasoning and limited numerical evidence, Reiser [26] concluded that the confidence interval $[\delta_L(X), \delta_U(X)]$ for δ has "essentially exact" coverage probability $1 - \alpha$. We have discovered that, in fact, this confidence interval has *exact* coverage probability $1 - \alpha$ as long as $\delta > 0$. Because the Reiser intervals for SNR and AUC are strictly increasing transformations of the interval for δ, they have the same coverage probability, and are thus exact under the condition that $\delta > 0$, i.e., SNR > 0. Our findings are summarized below; see the Appendix for a proof.

Theorem 2—Suppose that the conditions of Theorem 1 are satisfied, and let $\mathrm{SNR}_L(\hat{\theta})$ and $\mathrm{SNR}_U(\hat{\theta})$ be defined as above for some $\alpha \in (0, 1)$. If SNR > 0, then $[\mathrm{SNR}_L(\hat{\theta}), \mathrm{SNR}_U(\hat{\theta})]$ and $[\Phi(\mathrm{SNR}_L(\hat{\theta})/\sqrt{2}), \Phi(\mathrm{SNR}_U(\hat{\theta})/\sqrt{2})]$ are exact $1-\alpha$ confidence intervals for SNR and AUC, respectively.

The above theorem is the main theoretical result of this paper, and to our knowledge, is new. For readers interested in mathematical details, it is instructive to note that the proof given in the Appendix uses a different technique than the proofs in our previous papers on exact confidence intervals [30], [32], [34]. Specifically, the proofs in our previous papers relied on a standard theorem for pivoting a continuous cdf [27, Th. 9.2.12, p. 432], which does not apply in the present setting, because (5) and (6) do not have solutions for all observations x of X. To get around this difficulty, the proof of Theorem 2 (particularly Lemma 2), uses a slightly more general approach by returning to first principles.

As mentioned at the end of Section II, it is generally the case that for evaluations of image quality, which are the applications of interest in this paper, $\Delta\boldsymbol{\mu}$ is nonzero, and hence, SNR > 0. Nevertheless, it is useful to note that when SNR = 0, the Reiser intervals are conservative. This fact was observed by Reiser [26], who gave an argument demonstrating that the coverage probability is $1 - \alpha_1$ in this case.

IV. Monte Carlo Evaluation

A Monte Carlo simulation study was carried out to evaluate the coverage probability of the Reiser confidence intervals given in the previous section. For purposes of comparison, the coverage of approximate Wald-type intervals based on a normality approximation was also assessed. The Wald-type intervals are described below.

Let $\hat{V}$ be the point estimate of $\mathrm{Var}[\hat{\theta}]$ obtained by substituting the point estimate for $\hat{\theta}$ into the variance formula given in Corollary 1. Assuming that $(\hat{\theta} - \mathrm{SNR}^2)/\sqrt{\hat{V}}$ approximately follows a standard normal distribution, a classical Wald-type confidence interval [27, p. 499] for SNR^2 with approximate coverage probability $1 - \alpha$ is

$$\left[\hat{\theta} - z_{\alpha/2}\sqrt{\hat{V}}, \hat{\theta} + z_{\alpha/2}\sqrt{\hat{V}}\right] \quad (7)$$

where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ is the $1 - \alpha/2$ quantile for the standard normal distribution. Note that the lower bound of the Wald interval can possibly be negative. When the lower bound is positive, the above interval can be transformed into an SNR interval by taking the square root of the bounds, and into an AUC interval via the mapping $\Phi(\mathrm{SNR}/\sqrt{2})$.

For the simulation study, we used Monte Carlo simulation to test the coverage probabilities for the approximate Wald and exact Reiser confidence intervals. This was accomplished by simulating realizations of $\hat{\theta}$. For each estimate, we calculated the Reiser and Wald confidence intervals for SNR^2. Then we checked if the interval successfully covered the true value. The proportion of successes over many Monte Carlo trials then gave an estimate of the coverage probability, which ideally, should be $1-\alpha$. (Recall the definition of coverage probability stated at the beginning of Section III.)

For our evaluation, we took $\alpha = 0.05(\alpha_1 = \alpha_2 = 0.025)$, $p = 5, 20, 50$, AUC = 0.6, 0.75, 0.9, and $(m, n) = (150, 50), (100, 100)$. The sampling distribution in Theorem 1 was used to generate 100 000 random deviates of $\hat{\theta}$, from which SNR^2 confidence intervals were obtained, and the coverage probabilities were estimated. Regarding the exact Reiser confidence intervals, note that since SNR and AUC are linked to SNR^2 through a strictly increasing transformation, the coverage probabilities for SNR and AUC intervals are the same. However, for the approximate Wald confidence intervals, the link between SNR^2, SNR and AUC is broken when the lower interval bound is negative.

The estimated coverage probabilities are listed in Table I. Since each entry is an estimate of a binomial proportion, the corresponding standard deviation is $\sqrt{\mathrm{CP}(1-\mathrm{CP})/N}$ [27], where CP is the true coverage probability and N is the number of Monte Carlo trials. For example, since the coverage probability of the Reiser intervals is 0.95, the standard deviation for a coverage probability estimate based on $N = 100\,000$ trials is approximately 0.0007. For the Wald intervals, the true coverage probabilities are unknown, and the estimated coverage probability can be used in the above formula.

From Table I, it can be seen that the exact Reiser intervals have observed coverage probabilities very close to 95%, as expected. In contrast, the approximate Wald-style intervals are generally unreliable, especially for small AUC values and large numbers of channels. To better understand how the Wald intervals fail and the Reiser intervals succeed, plots of estimated probability density functions (pdfs) for the interval bounds (in the SNR^2 domain) are presented in Fig. 1 for the case $p = 50$, AUC = 0.75, and $m = 150$, $n = 50$, which is the fourth line from the bottom in Table I. The corresponding pdf for $\hat{\theta}$ is also plotted. In all plots, a dashed vertical line denoting the true SNR^2 value is given for reference. The estimated pdfs for the interval bounds are smoothed histogram estimates obtained with the kernel density method as offered in MATLAB, applied to the results of 50 000 Monte Carlo trials, whereas the pdf for $\hat{\theta}$ was obtained from the analytical expression in Theorem 1. From these plots, it can be seen that the distribution for $\hat{\theta}$ exhibits a positive bias, which strongly impacts the Wald interval bounds. On the other hand, the Reiser interval bounds are centered around the true SNR^2 value, and are not affected by the point estimation bias. Furthermore, it can be seen that the estimated pdfs for the lower and upper Reiser interval bounds both

have spikes at $SNR^2 = 0$. This observation is expected, since the definition of the Reiser intervals implies that both the lower bound and the upper bound can be zero. In summary, the inaccuracy of the Wald intervals is largely a result of the positive bias of $\hat{\theta}$, whereas the Reiser intervals overcome this problem by using knowledge of the complete distribution for $\hat{\theta}$.

V. Examples

To illustrate the exact CHO confidence intervals, we present two examples assessing image quality in X-ray computed tomography (CT). The first example uses simulated data of the XCAT phantom [38], and the second example is based on real CT scans of the QRM torso phantom.

Because noise in CT images is typically anisotropic due to its object-dependent nature, we implemented the CHOs in both examples with 18 Gabor channels centered on the location of interest. Gabor channels consist of a Gaussian modulated by a cosine function, and are circularly asymmetric and sensitive to image characteristics in different orientations. They have been used to model aspects of the human visual system [10], [39], and have been found to be yield higher detection performance than circularly symmetric channels in CT [40]. Mathematically, Gabor channels have the form [39]

$$G(x,y)=\exp[-4(\ln\ 2)((x-x_0)^2+(y-y_0)^2)/w_s^2]\ \times\cos[2\pi f_c((x-x_0)\ \cos\ \theta+(y-y_0)\ \sin\ \theta)+\xi] \quad (8)$$

where f_c denotes the center frequency of the channel, w_s is the spatial width of the channel, the point (x_0, y_0) is the center of the channel, θ is the channel orientation, and ξ is a phase factor. The spatial width of the channel, w_s, is related to w_f, the full-width-at-half-maximum of the Fourier transform of the channel, through $w_s = 4(\ln 2)/(\pi w_f)$. We used three channel passbands, given by [1/32, 1/16], [1/16, 1/8], and [1/8,1/4] cycles/pixel, respectively; this choice of passbands is similar to those in [21], and was motivated by previous studies involving model observers [6], [7], [9]. The center frequencies for the passbands are f_c = 3/64, 3/32, and 3/16 cycles/pixel, with spatial channel widths of w_s = 28.24, 14.12, and 7.06 pixels, respectively. In addition, we used three orientations: $\theta = 0, \pi/3$, and $2\pi/3$ radians, and two phases: $\xi = 0, \pi/2$. Hence, there were 3 frequency bands × 3 orientations × 2 phases for a total of 18 channels.

A. Example 1

This example compares image quality for three fan-beam CT image reconstruction methods by using a known-location, background-variable classification task in which a CHO is applied to discriminate between two types of kidney stones. According to Kambadakone *et al.* [41], knowledge of kidney stone composition is an important factor influencing patient management. In particular, CT attenuation measurements have been found to be valuable for differentiation of uric acid stones from other stone types [41]. Motivated by this knowledge,

we selected the task of discriminating between a uric acid stone (450 HU) and a struvite stone (600 HU) at a fixed location in the left kidney (recall that the left kidney is commonly displayed on the right side of transverse-plane CT images).

To generate images, we simulated CT data for the XCAT phantom [38] slice shown in Fig. 2. All images were reconstructed for a 96 × 96 region of interest (ROI) centered on the left kidney, with a pixel size of 1 mm × 1 mm. To simulate anatomical variations in the left kidney, we made two additions to the phantom, as follows. First, a circular disk of fat with a random radius between 5 mm and 10 mm was added to the kidney, near the ureter. Second, a random background was added to the kidney to simulate texture. This variable background was modeled as colored Gaussian noise following a power law model with an exponent of two [42], with an amplitude of ±6 HU. Last, a kidney stone of diameter 4 mm was added at a fixed location, with attenuation values of 450 HU for class 1 and 600 HU for class 2; see Fig. 2. Once a realization of the phantom was defined, fan-beam sinograms were computed, and Poisson noise with a photon level of 10 000 was added to the data with models for tube current modulation and a bowtie filter, as described in [21].

We compared three fan-beam filtered backprojection (FBP) image reconstruction algorithms: **A:** Noo *et al.* [43] for full-scan data, **B:** Dennerlein *et al.* [44] for full-scan data, and **C:** Noo *et al.* [43] for short-scan data (240°). All algorithms were implemented so that resolution was matched near the field of view center. Algorithms A and B were expected to yield comparable CHO performance (since the kidney is fairly close to the field-of-view center), whereas algorithm C was anticipated to be worse since it uses less CT data.

To carry out the evaluation, 250 class-1 sinograms and 200 class-2 sinograms were generated. Next, every sinogram was reconstructed with the three algorithms to yield three sets of 450 images, i.e., one set for each reconstruction method. Finally, exact AUC confidence intervals describing CHO performance were estimated for each reconstruction algorithm.

Multivariate normality of the channel outputs was validated at the 10% significance level with the Henze–Zirkler normality test [45]. Since the confidence intervals were correlated, the coverage probability for each confidence interval was selected to be 96.67% so that the joint coverage probability for the three intervals was at least 90% by the Bonferroni inequality.[1] The results, given in Table II, indicate that Algorithms A and B were both better than Algorithm C with statistical significance. However, there was no statistically significant difference between Algorithms A and B, which was consistent with our expectations.

B. Example 2

In this example, we evaluate image quality for two fan-beam CT image reconstruction methods applied to real CT data by using a CHO to detect a lesion at a known location in a torso phantom. Data was collected with a Siemens SOMATOM Sensation 64 CT scanner by

[1]For arbitrary events E_1, E_2, …, E_k, the Bonferroni inequality [27, (1.2. 10), p. 13] takes the form

$$P(\cap_{i=1}^{k} E_i) \geq \sum_{i=1}^{k} P(E_i) - (k-1).$$

repeatedly scanning a torso phantom 175 times over a circular source trajectory. The torso phantom, constructed by QRM (Möhrendorf, Germany), was used together with two water bottles placed on the sides to simulate arms. A mean image of the phantom estimated from all 175 scans is shown in Fig. 3. The scans were executed in thorax scan mode using a 2-slice acquisition with a slice thickness of 1 mm, a rotation speed of 3 rev/s, X-ray tube settings of 25 mAs and 120 kVp, and no tube current modulation. Fan-beam measurements for the first of the two slices over the 175 repeated scans were used for the evaluation.

The CT data was read with software provided by Siemens and then reconstructed using our implementation of the classical FBP algorithm for direct reconstruction from either short-scan or full-scan fan-beam data [21]. Short-scan reconstructions used data collected from 230° of the source trajectory. Images were reconstructed on a 550 × 550 grid centered on the heart insert, as shown in Fig. 3, with a pixel size of 0.2 mm × 0.2 mm.

For both short-scan and full-scan reconstructions, we considered a known-location lesion detection task in which the lesion was either absent or present at the center of a 64 × 64 region-of-interest (ROI). From each reconstructed image of the heart insert, we extracted three ROI images, labeled 1a, 1b, and 2 in Fig. 3. ROI-1a and ROI-1b were used for class-1 (lesion absent), and ROI-2 was used for class-2 (lesion-present). The lesion in ROI-2 had a diameter of 1 mm and a contrast of 210 HU with the background value of 40 HU. We assumed that the three ROIs obtained from a given CT image were statistically independent. This assumption was justified by a previous study of fan-beam FBP reconstruction from simulated CT data [21], which indicated that correlations between image pixels are negligible over the distance that separated the ROIs.

Since two class-1 ROIs and one class-2 ROI were extracted from each heart insert image, CHO performance for each reconstruction algorithm was estimated from a total of 350 class-1 and 175 class-2 ROI images. Multivariate normality of the channel outputs was checked with the Henze-Zirkler test [45]. As in Example 1, we found that the null hypothesis of multivariate normality was supported at the 10% significance level. Estimated 95% AUC confidence intervals for the short-scan and full-scan reconstructions are shown in Table III. By the Bonferroni inequality (see earlier footnote), the combined coverage probability of both intervals was at least 90%. The full-scan reconstruction was observed to be better than the short-scan reconstruction with statististical significance. This finding was consistent with the fact that the short-scan reconstruction used less data than the full-scan reconstruction.

VI. Comparison to Training/Testing Approaches

As mentioned in the introduction, CHO performance may also be estimated using a training/testing approach. In this approach, the CHO template is first estimated from a set of training images during the training phase. Second, during the testing phase, the estimated template is applied to a set of testing images to generate ratings, from which a performance estimate is obtained. Due to frequent use of training/testing in CHO performance evaluations, e.g., [11], [12], [14], [16]–[18], [20], [25], it is worthwhile to clarify the relationship between the approach of this paper and the training/testing paradigm.

The present work is concerned with performance estimation of a CHO that is defined by the template $\mathbf{w} = \Sigma^{-1}\Delta\boldsymbol{\mu}$. Generally, the population quantities $\Delta\boldsymbol{\mu}$ and Σ are unknown, and cannot be determined exactly from a finite sample of training images. Therefore, this work can be characterized in terms of training/testing terminology by saying that we aim to estimate *infinitely-trained* CHO performance, i.e., the performance that results from the true template. Alternatively, for a fixed number of training images, one can instead seek to estimate *finitely-trained* CHO performance, i.e., the performance that results from a template estimated with a finite training set. These contrasting goals are examined below.

Two strategies for training and testing are often employed in CHO evaluations. The first strategy, called the resubstitution method, uses the same images for training and testing. The second strategy, called the hold-out method, uses independent sets of images for training and testing. For the purpose of estimating the performance of an infinitely-trained CHO, it is well known from the pattern recognition literature that the resubstitution method yields point estimates with a positive bias, and that the hold-out method gives point estimates with a negative bias [46], [47].

Fig. 4 illustrates these biases for a CHO with $p = 18$ channels when the channel output vector is multivariate normal under each class, and the class covariance matrices are equal. In this example, Σ was taken to be the identity matrix and $\Delta\boldsymbol{\mu}$ was chosen so that AUC = 0.75. Specifically, the class means were taken to be $\boldsymbol{\mu}_1 = \mathbf{0}$, and $\boldsymbol{\mu}_2 = 0.831\boldsymbol{\nu}$, where $\boldsymbol{\nu}$ was a $p \times 1$ vector with entries increasing from 0.025 to 0.45 in increments of 0.025. For each point in Fig. 4, we generated 100 000 independent Monte Carlo trials with the number of training images equal to the number of testing images, $N_{\text{train}} = N_{\text{test}}$, and estimated AUC using the unbiased Mann-Whitney statistic [35]. The resulting mean AUC estimates are plotted as a function of N_{train}. As expected, the resubstitution and hold-out methods have positive and negative biases, respectively, which decrease with the number of training images.

Two aspects regarding the biases described above deserve comment. First, the point estimator presented in Section III-A, like the resubstitution approach, utilizes all available images. In fact, it is straightforward to show that the resubstitution estimate of SNR^2 is equivalent to the plug-in estimator, $\hat{\theta}_p$, given in Section III-A, which as discussed earlier, has a positive bias. This observation is in agreement with the results given in Fig. 4 for the resubstitution estimator.

Second, although the hold-out method yields a negatively-biased estimate of infinitely-trained CHO performance, it gives an unbiased estimate of finitely-trained performance (if an unbiased performance estimator is used on the testing results), since the testing set is independent from the training set. This is the performance that is expected for a classifier that is applied in practice. The dependence of classifier performance on training is often depicted with a curve like the hold-out curve in Fig. 4, called a “learning curve” ([48, p. 215]). Learning curves are vital to understanding the effectiveness of training/testing approaches, as we will see next.

In addition to point estimates for CHO performance, training/testing methods can also be used to estimate confidence intervals. To illustrate the accuracy of confidence intervals obtained-with training and testing, and to compare such approaches with the Reiser intervals introduced in Section III, we carried out a Monte Carlo evaluation in which the channel output vectors were assumed to satisfy our distributional assumptions, i.e., multivariate normal under each class with equal class covariance matrices. Specifically, four types of bootstrap AUC confidence intervals were evaluated: 1) resubstitution with bootstrapping over testing cases only, 2) resubstitution with bootstrapping over both training and testing cases, 3) hold-out with bootstrapping over testing cases only, and 4) hold-out with bootstrapping over both training and testing cases. Bootstrap samples were obtained by sampling with replacement, and each confidence interval was estimated from 1000 bootstrap samples with the percentile method [49]. As above, $p = 18$ channels were used, Σ was taken to be the identity matrix, and $\Delta\boldsymbol{\mu}$ was chosen so that AUC = 0.75. Coverage probabilities of confidence intervals were evaluated for $N_{\text{train}} = 50$, 100, 200, 500, and 1000 training images. The hold-out method was applied with an independent set of $N_{\text{test}} = 200$ testing images, while the resubstitution method utilized the same images for training and testing. In other words, letting N_{tot} denote the total number of images used, the Reiser and resubstitution methods utilized $N_{\text{tot}} = N_{\text{train}}$ images, while the hold-out method was applied with $N_{\text{tot}} = N_{\text{train}} + N_{\text{test}} = N_{\text{train}} + 200$ images. Coverage probability estimates were obtained from 100 000 independent Monte Carlo trials, in which the channel output vectors were generated as random deviates from multivariate normal distributions with the previously specified parameters.

Estimated coverage probabilities of 95% AUC confidence intervals for infinitely-trained and finitely-trained CHO performance are listed in Tables IV and V, respectively. Here, the results for infinitely-trained performance assessed coverage of AUC = 0.75, whereas the results for finitely-trained performance assessed coverage of the mean AUC value obtained with the hold-out method, listed as the “target AUC” in the table, which was estimated from 100 000 Monte Carlo trials. From Table IV, we see that the Reiser intervals resulted in consistently accurate coverage for infinitely-trained performance. By contrast, the resubstitution method was not reliable, and the hold-out method was accurate only for a large number of training images. Given the learning curve in Fig. 4, this phenomenon is not surprising, since the biases of resubstitution and hold-out methods decrease with increasing numbers of training images. On the other hand, for the goal of estimating finitely-trained CHO performance, i.e., the value on the hold-out (learning) curve in Fig. 4, Table V illustrates that the hold-out method with bootstrapping over both training and testing sets gives conservative coverage, and that the hold-out method with bootstrapping over testing only is reliable for large numbers of training images. This latter result is consistent with the expectation that training variability becomes smaller as the number of training images increases. For both types of bootstrapping, the resubstitution method is seen to be unreliable, which is likely due to the positive bias of resubstitution (relative to the learning curve) as seen in Fig. 4. Last, as expected, the Reiser intervals were not as reliable for estimating finitely-trained performance, since they are not designed for this purpose.

In summary, the Reiser intervals were found to be well-suited for estimating infinitely-trained performance for all sample sizes. By contrast, the hold-out method, which uses independent training and testing sets, was observed to be best-suited for estimating finitely-trained performance, although it was also effective for estimation of infinitely-trained performance when the size of the training set was large. Lastly, the resubstitution method was not generally reliable for either purpose, although its accuracy improved as the number of training images increased.

It should be emphasized that the results presented above are for a limited number of examples, and that the channel output vector was taken to be multivariate normal under each class with equal class covariance matrices. A more thorough comparison of the Reiser intervals with methods based on training and testing is beyond the scope of the present work.

VII. DISCUSSION AND CONCLUSION

A widely-used approach for the characterization of medical image quality involves estimating the performance of a CHO applied to a known-location discrimination task. For the problem of estimating confidence intervals for CHO performance measures, we have shown that a method proposed by Reiser [26] can be used to obtain confidence intervals for SNR and AUC. In addition, we rigorously proved that these intervals are well-defined and exactly achieve the desired coverage probability, as long as SNR > 0, which is a new result not given in [26]. The coverage probability of the Reiser confidence intervals was verified with Monte Carlo simulation, and was found to be more robust than that for conventional Wald-style intervals based on a normality assumption. Application of the Reiser confidence intervals was demonstrated with two examples comparing reconstruction algorithms for fan-beam CT. Last, the relationship between the Reiser intervals and training/testing approaches was discussed and illustrated with examples based on Monte Carlo simulation.

To our knowledge, the Reiser intervals investigated here are presently the only known exact confidence intervals for CHO performance when both Σ and $\Delta\boldsymbol{\mu}$ are unknown. As clarified in Section VI, these intervals are designed for estimation of infinitely-trained CHO performance, which is typically the goal of image quality studies aimed at imaging system optimization. On the other hand, if the goal is to estimate finitely-trained CHO performance, then approaches based on training and testing with independent data sets are preferable.

Our theoretical results assume that the channel output vector for a given class follows a multivariate normal distribution with equal covariance matrices for each class. As discussed in the introduction, these assumptions are often well-justified for image discrimination tasks involving a small lesion at a known location with either a flat or normally-distributed, variable background. Indeed, these assumptions are typically well-justified for tomographic imaging applications. To be sure of the applicability of the Reiser intervals in practice, it is generally desirable to check the aforementioned distributional assumptions; recall from the image quality evaluation examples in Section V, that multivariate normality was checked using the Henze-Zirkler normality test. We did not investigate robustness of the Reiser intervals to violations of the distributional assumptions. However, since the Reiser intervals are based on the two-sample Hotelling T^2 test statistic, which is invariant under nonsingular

affine transformations of the channel output vector, it can be conjectured that the Reiser intervals possess robustness properties similar to those of the T^2 test [4]. Further research into this issue, both on theoretical grounds and with Monte Carlo studies, is a topic of high interest for future work.

It is useful to note that although our presentation concentrated on exact confidence intervals for SNR and AUC, exact confidence intervals can also be obtained for other figures of merit. In particular, because our distributional assumptions imply that the ROC curve is parameterized by only SNR, it follows that TPF at fixed FPF and partial AUC (pAUC) are strictly increasing functions of SNR. Consequently, Lemma 3 implies that exact confidence intervals for these figures of merit can be obtained through suitable transformations of an exact SNR interval; see [30], [34] for further details. Moreover, it turns out that the union of exact $1 - \alpha$ TPF intervals over all FPF values is an exact, simultaneous $1 - \alpha$ confidence band for the entire ROC curve in this setting [30].

As one might intuitively expect, smaller confidence intervals can be obtained when prior knowledge is available. For example, in a recent paper [34], we have introduced exact intervals for situations when $\Delta\boldsymbol{\mu}$ is known. To illustrate the advantage offered by prior knowledge of $\Delta\boldsymbol{\mu}$, Fig. 5 contains plots of mean confidence interval length (MCIL) for exact 95% AUC confidence intervals versus sample size. These plots compare the MCIL for the unknown-$\Delta\boldsymbol{\mu}$ Reiser confidence intervals presented here to the known-$\Delta\boldsymbol{\mu}$ confidence intervals given in [34]. Each point on these curves was calculated through a numerical integration as described in [34]. Note that in both cases, for given values of m, n, and p, the underlying statistical distributions depend only on AUC, and that the plots are valid for any choices of $\Delta\boldsymbol{\mu}$ and Σ that yield the stated AUC value. From the plots in Fig. 5, it is evident that a large decrease in confidence interval length can be gained when $\Delta\boldsymbol{\mu}$ is known, and that this decrease is larger as the nominal AUC value becomes smaller.

MATLAB software implementing the confidence interval estimators described in this work and in previous related publications [30], [32], [34] is publicly available at http://code.google.com/p/iqmodelo/.

Acknowledgments

This work was supported in part by the National Institutes of Health under Grant R01 EB007236 and Grant R21 EB009168, and in part by a grant from the Ben B. and Iris M. Margolis Foundation.

Appendix

We prove Theorem 2 with the aid of three lemmas.

Lemma 1

The cdf of the noncentral F-distribution is a strictly-decreasing, continuous function of the noncentrality parameter.

Proof

Let $X \sim F'_{\nu_1,\nu_2}(\delta)$. By the definition of the non-central-F distribution, $X = U/V$, where $U \sim \chi'^2_{\nu_1}(\delta)$ and $V \sim \chi^2_{\nu_2}$ are independent random variables, i.e., U is a noncentral χ^2 random variable with ν_1 degrees of freedom and noncentrality parameter δ, and V is a central χ^2 random variable with ν_2 degrees of freedom. Denote the pdf's for U and V as $f_U(u; \nu_1, \delta)$ and $f_V(v; \nu_2)$, respectively, and denote the cdf of U as $F_U(u; \nu_1, \delta)$. Since $X = U/V$, the cdf for X is $F_X(x; \nu_1, \nu_2, \delta) = P(X \leq x) = P(U \leq xV)$, and hence

$$F_X(x;\nu_1,\nu_2,\delta) = \int_0^\infty f_V(v;\nu_2)dv \int_0^{xv} f_U(u;\nu_1,\delta)\mathrm{du} = \int_0^\infty F_U(xv;\nu_1,\delta) f_V(v;\nu_2)dv.$$

Since F_U is differentiable with respect to δ [36, p. 442], we can differentiate under the integral sign to obtain

$$\frac{\partial F_X}{\partial \delta} = \int_0^\infty \frac{\partial F_U}{\partial \delta}(xv;\nu_1,\delta) f_V(v;\nu_2)dv. \quad (9)$$

From (29.33e) in [36, p. 443], for any and u, ν_1 and δ, $\partial F_U(u; \nu_1, \delta)/\partial\delta < 0$. Combining this fact with (9), we see that $\partial F_X(x; \nu_1, \nu_2, \delta)/\partial\delta < 0$.

Lemma 2

Let $\alpha_1, \alpha_2 \in (0, 1)$ be fixed numbers such that $\alpha_1 + \alpha_2 = \alpha$ for some $\alpha \in (0, 1)$ and let $X \sim F'_{\nu_1,\nu_2}(\delta)$, where ν_1, ν_2, and δ are fixed. Further, for each observation x of X, define the functions $\delta_L(x)$ and $\delta_U(x)$ as in Section III-B. If $\delta > 0$, then the random interval $[\delta_L(X), \delta_U(X)]$ is an exact $1 - \alpha$ confidence interval for δ.

Proof

Denote the cdf of X by $F_X(x; \nu_1, \nu_2, \delta)$. We partition the sample space with the following three disjoint events:

$$A = \{X : F_X(X;\nu_1,\nu_2,\delta) > 1 - \alpha_1\}$$

$$B = \{X : \alpha_2 \leq F_X(X;\nu_1,\nu_2,\delta) \leq 1 - \alpha_1\}$$

$$C = \{X : F_X(X;\nu_1,\nu_2,\delta) < \alpha_2\}$$

By a standard theorem [27, Th. 2.1.10, p. 54], $F_X(X; \nu_1, \nu_2, \delta)$ is uniformly distributed. It then follows that $P(A) = \alpha_1$, $P(B) = 1 - \alpha$ and $P(C) = \alpha_2$. We will consider each of these events separately to determine their contribution to the overall coverage probability. Below,

we repeatedly utilize the fact that $F_X(x;\ \nu_1,\ \nu_2,\ \delta)$ is a strictly decreasing function of δ, as stated by Lemma 1.

For $X \in A$, the definition of $\delta_L(X)$ and Lemma 1 imply that $\delta_L(X) > \delta$. Hence, $P(\delta \in [\delta_L(X),\ \delta_U(X)]|A) = 0$. Now, consider $X \in B$. The definition of $\delta_U(X)$ and Lemma 1 imply that $\delta \leq \delta_U(X)$. For $\delta_L(X)$, there are two cases: either (a) $F_X(X;\ \nu_1,\ \nu_2,\ 0) \geq 1 - \alpha_1$ or (b) $F_X(X;\ \nu_1,\ \nu_2,\ 0) < 1 - \alpha_1$. In case (a), the definition of $\delta_L(X)$ and Lemma 1 imply that $0 < \delta_L(X) \leq \delta$. In case (b), the definition of $\delta_L(X)$ and Lemma 1 imply that $\delta_L(X) = 0 < \delta$. Either way, we have $\delta_L(X) \leq \delta$, and thus, $P(\delta \in [\delta_L(X),\ \delta_U(X)]|B) = 1$. Finally, consider $X \in C$. If $F_X(X;\ \nu_1,\ \nu_2,\ 0) \geq \alpha_2$ then the definition of $\delta_U(X)$ and Lemma 1 imply that $\delta_U(X) < \delta$. Otherwise, if $F_X(X;\ \nu_1,\ \nu_2,\ 0) < \alpha_2$, then the definition of $\delta_U(X)$ and Lemma 1 imply that $\delta_U(X) = 0 < \delta$. Thus, $P(\delta \in [\delta_L(X),\ \delta_U(X)]|C) = 0$.

Therefore, the coverage probability is

$$\begin{aligned} &P(\delta \in [\delta_L(X), \delta_U(X)]) \\ &\quad = P(\delta \in [\delta_L(X), \delta_U(X)]|A)P(A) \\ &\quad + P(\delta \in [\delta_L(X), \delta_U(X)]|B)P(B) \\ &\quad + P(\delta \in [\delta_L(X), \delta_U(X)]|C)P(C) \\ &\quad = (0)(\alpha_1) + (1)(1-\alpha) + (0)(\alpha_2) = 1 - \alpha. \end{aligned}$$

Lemma 3

Let $g(\delta)$ be a continuous, strictly increasing function of δ. If $[\delta_L,\ \delta_U]$ is a $1 - \alpha$ confidence interval for δ, then $[g(\delta_L),\ g(\delta_U)]$ is a $1 - \alpha$ confidence interval for $g(\theta)$.

Proof

See Lemma 3 in [30].

Theorem 1 and Lemma 2 imply that $[\delta_L(X),\ \delta_U(X)]$ is an $1 - \alpha$ exact confidence interval for δ. Finally, since SNR and AUC are strictly increasing functions of δ, Lemma 3 yields the stated result for Theorem 2.

References

1. Barrett, HH., Myers, KJ. Foundations of Image Science. Hoboken, NJ: Wiley; 2004.
2. Barrett HH, Yao J, Rolland JP, Myers KJ. Model observers for assessment of image quality. Proc. Nat. Acad. Sci. USA. Nov; 1993 90(21):9758–9765. [PubMed: 8234311]
3. He X, Park S. Model observers in medical imaging research. Theranostics. 2013; 3(10):774–786. [PubMed: 24312150]
4. Muirhead, RJ. Aspects of Muiltivariate Statistical Theory. Hoboken, NJ: Wiley; 2005.
5. Fukunaga, K. Statistical Pattern Recognition. 2. San Diego, CA: Academic; 1990.
6. Myers KJ, Barrett HH. Addition of a channel mechanism to the ideal-observer model. J Opt. Soc. Am. A. Dec; 1987 4(12):2447–2457. [PubMed: 3430229]
7. Wollenweber SD, Tsui BMW, Lalush DS, Frey EC, LaCroix KJ, Gullberg GT. Comparison of Hotelling observer models and human observers in defect detection from myocardial SPECT imaging. IEEE Trans. Nucl. Sci. Dec; 1999 46(6):2098–2103.

8. Gifford HC, King MA, de Vries DJ, Soares EJ. Channelized Hotelling and human observer correlation for lesion detection in hepatic SPECT imaging. J Nucl. Med. Mar; 2000 41(3):514–521. [PubMed: 10716327]
9. Abbey CK, Barrett HH. Human-and model-observer performance in ramp-spectrum noise: Effects of regularization and object variability. J Opt. Soc. Am. A. 2001; 18(3):473–488.
10. Zhang Y, Pham BT, Eckstein MP. The effect of nonlinear human visual system components on performance of a channelized Hotelling observer in structured backgrounds. IEEE Trans. Med. Imag. Oct; 2006 25(10):1348–1362.
11. Park S, Badano A, Gallas BD, Myers KJ. Incorporating human contrast sensitivity in model observers for detection tasks. IEEE Trans. Med. Imag. Mar; 2009 28(3):339–347.
12. Gallas BD, Barrett HH. Validating the use of channels to estimate the ideal linear observer. J Opt. Soc. Am. A. Sep; 2003 20(9):1725–1738.
13. Bonetto P, Qi J, Leahy RM. Covariance approximation for fast and accurate computation of channelized Hotelling observer statistics. IEEE Trans. Nucl. Sci. Aug; 2000 47(4):1567–1572.
14. Frey EC, Gilland KL, Tsui BMW. Application of task-based measures of image quality to optimization and evaluation of three-dimensional reconstruction-based compensation methods in myocardial perfusion SPECT. IEEE Trans. Med. Imag. Sep; 2002 21(9):1040–1050.
15. Qi J. Analysis of lesion detectability in Bayesian emission reconstruction with nonstationary object variability. IEEE Trans. Med. Imag. Mar; 2004 23(3):321–329.
16. Gagne RM, Gallas BD, Myers KJ. Toward objective and quantitative evaluation of imaging systems using images of phantoms. Med. Phys. Jan; 2006 33(1):83–95. [PubMed: 16485413]
17. Son I-Y, Yazici B, Xu XG. X-ray imaging optimization using virtual phantoms and computerized observer modelling. Phys. Med. Biol. Sep; 2006 51(17):4289–4310. [PubMed: 16912382]
18. Tisdall MD, Atkins MS. Using human and model performance to compare MRI reconstructions. IEEE Trans. Med. Imag. Nov; 2006 25(11):1510–1517.
19. El Fakhri G, Santos PA, Badawi RD, Holdsworth CH, Van Den Abbeele AD, Kijewski MF. Impact of acquisition geometry, image processing, and patient size on lesion detection in whole-body 18F-FDG PET. J Nucl. Med. Dec; 2007 48(12):1951–1960. [PubMed: 18006613]
20. Marchessoux C, Kimpe T, Bert T. A virtual imaging chain for perceived and clinical image quality of medical display. J Disp. Technol. Dec; 2008 4(4):356–368.
21. Wunderlich A, Noo F. Image covariance and lesion detectability in direct fan-beam x-ray computed tomography. Phys. Med. Biol. May; 2008 53(10):2471–2493. [PubMed: 18424878]
22. Tang J, Rahmim A, Lautamäki R, Lodge MA, Bengel FM, Tsui BMW. Optimization of Rb-82 PET acquisition and reconstruction protocols for myocardial perfusion defect detection. Phys. Med. Biol. 2009; 54:3161–3171. [PubMed: 19420417]
23. Park S, Jennings R, Liu H, Badano A, Myers K. A statistical, task-based evaluation method for three-dimensional x-ray breast imaging systems using variable-background phantoms. Med. Phys. 2010; 37:6253–6270. [PubMed: 21302782]
24. Cao N, Huesman RH, Moses WW, Qi J. Detection performance analysis for time-of-flight PET. Phys. Med. Biol. 2010; 55:6931–6950. [PubMed: 21048292]
25. He X, Links JM, Frey EC. An investigation of the trade-off between the count level and image quality in myocardial perfusion SPECT using simulated images: The effects of statistical noise and object variability on defect detectability. Phys. Med. Biol. 2010; 55:4949–4961. [PubMed: 20693615]
26. Reiser B. Confidence intervals for the Mahalanobis distance. Commun. Stat. Simulat. 2001; 30(1): 37–45.
27. Casella, G., Berger, RL. Statistical Inference. 2. Independence, KY: Duxbury; 2001.
28. Khurd P, Gindi G. Fast LROC analysis of Baysian reconstructed tomographic images using model observers. Phys. Med. Biol. 2005; 50:1519–1532. [PubMed: 15798341]
29. Zeng R, Petrick N, Gavrielides MA, Myers KJ. Approximations of noise covariance in multi-slice helical CT scans: Impact on lung-nodule size estimation. Phys. Med. Biol. 2011; 56:6223–6242. [PubMed: 21896963]

30. Wunderlich A, Noo F. Confidence intervals for performance assessment of linear observers. Med. Phys. Jul; 2011 38(S1):S57–S68.
31. Wunderlich, A., Noo, F., Heilbrun, M. New results for efficient estimation of CHO performance; Proc. 2nd Int. Conf. Image Format. X-ray CT; Jun. 2012 p. 153-156.
32. Wunderlich A, Noo F. On efficient assessment of image-quality metrics based on linear model observers. IEEE Trans. Nucl. Sci. Jun; 2012 59(3):568–578. [PubMed: 23335815]
33. Wunderlich A, Noo F. Estimation of channelized Hotelling observer performance with known class means or known difference of class means. IEEE Trans. Med. Imag. Aug; 2009 28(8):1198–1207.
34. Wunderlich A, Noo F. New theoretical results on channelized Hotelling observer performance estimation with known difference of class means. IEEE Trans. Nucl. Sci. Feb; 2013 60(1):182–193. [PubMed: 24436497]
35. Pepe, MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford, U.K.: Oxford Univ. Press; 2003.
36. Johnson, NL., Kotz, S., Balakrishnan, N. Continuous Univariate Distributions. 2. Vol. 2. Hoboken, NJ: Wiley; 1995.
37. Sahai, H., Ageel, MI. The Analysis of Variance: Fixed, Random and Mixed Models. Boston, MA: Birkhäuser; 2000.
38. Segars WP, Mahesh M, Beck TJ, Frey EC, Tsui BMW. Realistic CT simulation using the 4D XCAT phantom. Med. Phys. Aug; 2008 35(8):3800–3808.
39. Eckstein MP, Bartroff JL, Abbey CK, Whiting JS, Bochud FO. Automated computer evaluation and optimization of image compression of x-ray coronary angiograms for signal known exactly detection tasks. Opt. Exp. Mar; 2003 11(5):460–475.
40. Wunderlich, A., Noo, F. Evaluation of the impact of tube current modulation on lesion detectability using model observers; Proc. Int. Conf. IEEE Eng. Med. Biol. Soc; Aug. 2008M p. 2705-2708.
41. Kambadakone AR, Eisner BH, Catalano OA, Sahani DV. New and evolving concepts in the imaging and management of urolithiasis: Urologists perspective. Radiographics. 2010; 30(3):603–623. [PubMed: 20462984]
42. Metheany KG, Abbey CK, Packard N, Boone JM. Characterizing anatomical variability in breast CT images. Med. Phys. Oct; 2008 35(10):4685–4694. [PubMed: 18975714]
43. Noo F, Defrise M, Clackdoyle R, Kudo H. Image reconstruction from fan-beam projections on less than a short scan. Phys. Med. Biol. 2002; 47(14):2525–2546. [PubMed: 12171338]
44. Dennerlein F, Noo F, Hornegger J, Lauritsch G. Fan-beam filtered-backprojection reconstruction without backprojection weight. Phys. Med. Biol. 2007; 52(11):3227–3240. [PubMed: 17505099]
45. Henze N, Zirkler B. A class of invariant consistent tests for multivariate normality. Commun. Stat. Theory. 1990; 19(10):3595–3617.
46. Fukunaga K, Hayes RR. Estimation of classifier performance. IEEE Trans. Pattern Anal. Mach. Intell. Oct; 1989 11(10):1087–1101.
47. Chan H-P, Sahiner B, Wagner RF, Petrick N. Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers. Med. Phys. Dec; 1999 26(12):2654–2668. [PubMed: 10619251]
48. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning. New York: Springer; 2001.
49. Efron, B., Tibshirani, RJ. An Introduction to the Bootstrap. Boca Raton, FL: Chapman Hall/CRC; 1993.

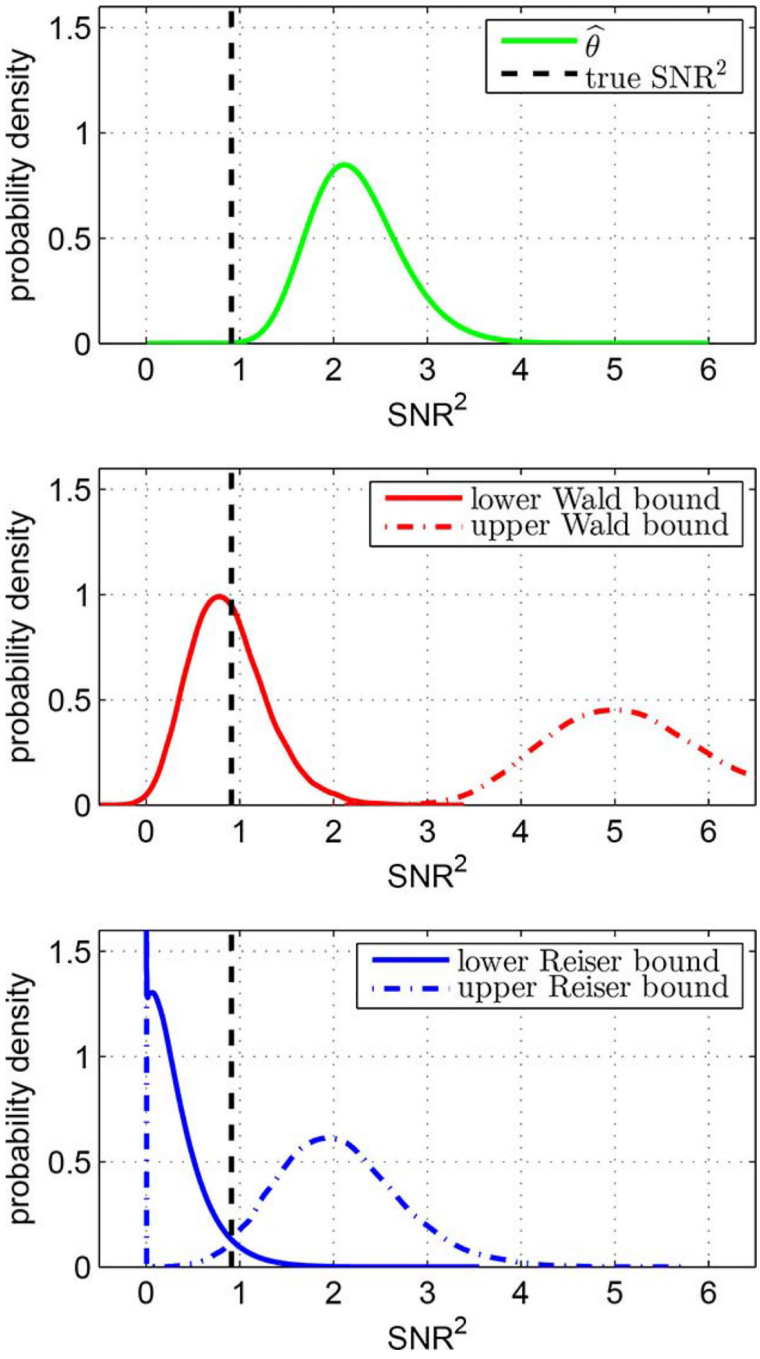

Fig. 1.

Estimated marginal distributions for Wald and Reiser SNR^2 confidence interval bounds with $p = 50$, AUC = 0.75(SNR^2 = 0.91), $m = 150$, and $n = 50$. (Top) pdf of $\hat{\theta}$, (Middle) estimated pdfs of upper and lower Wald bounds, (Bottom) estimated pdfs of upper and lower Reiser bounds. The true SNR^2 is denoted by the dashed vertical line.

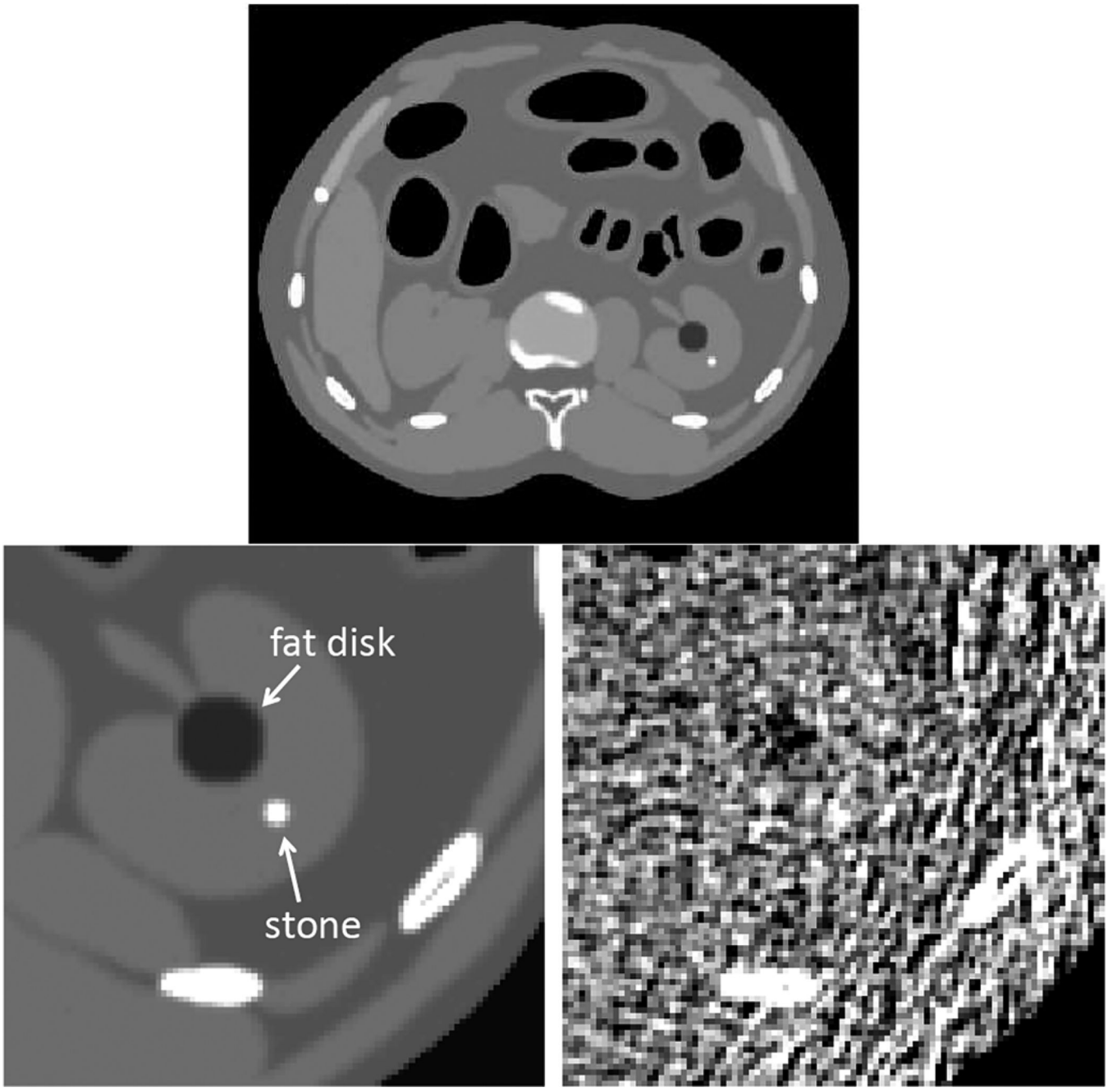

Fig. 2.
(Top) Slice of the XCAT phantom. (Lower Left) ROI centered on left kidney. (Lower Right) Noisy reconstruction. Grayscale: [−150, 250] HU.

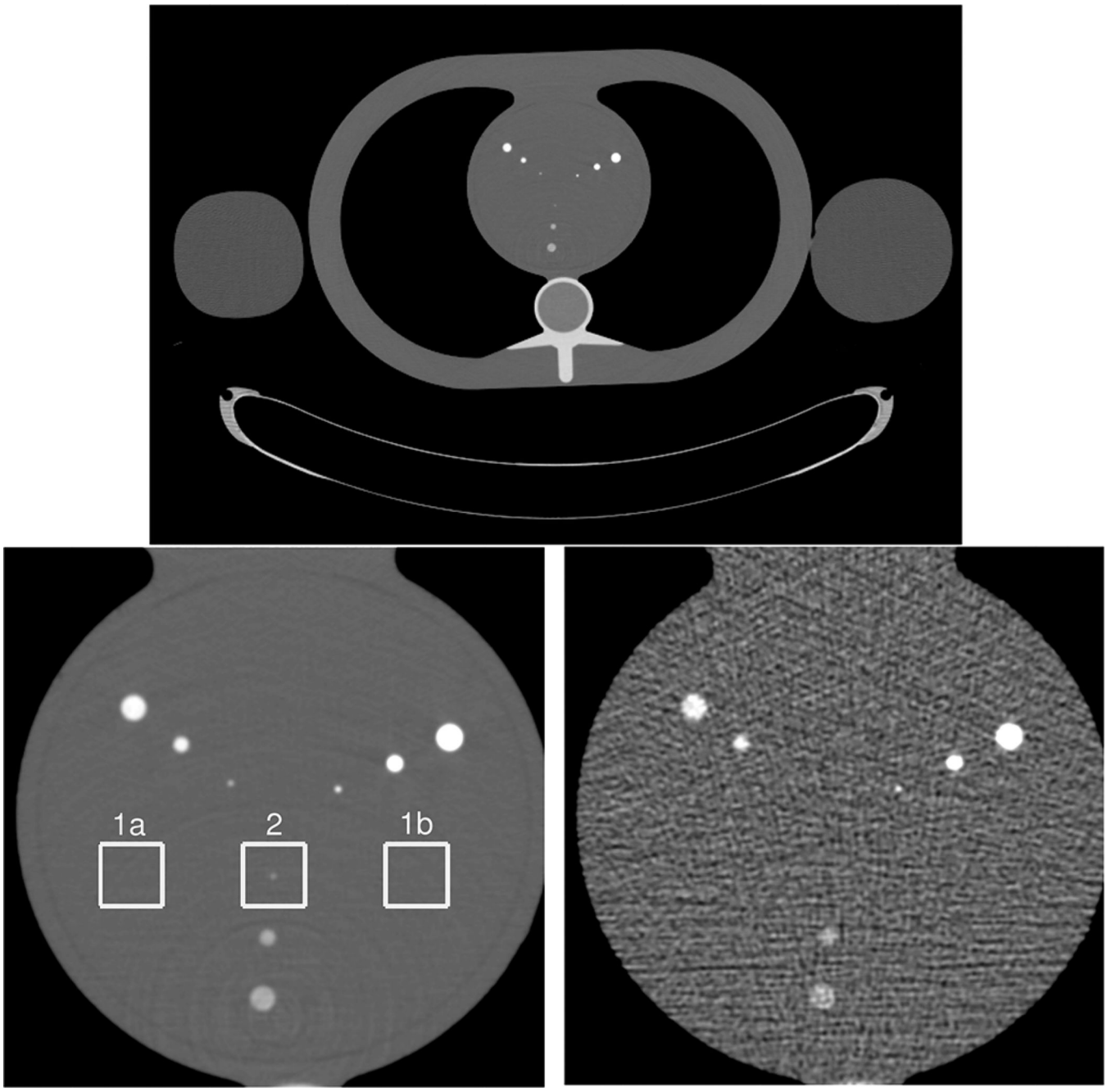

Fig. 3.

(Top) Mean image of QRM torso phantom. (Lower Left) Mean image focused on the heart insert with ROIs marked by white boxes. (Lower Right) A noisy reconstruction of the heart insert. Grayscale: [−250, 500] HU.

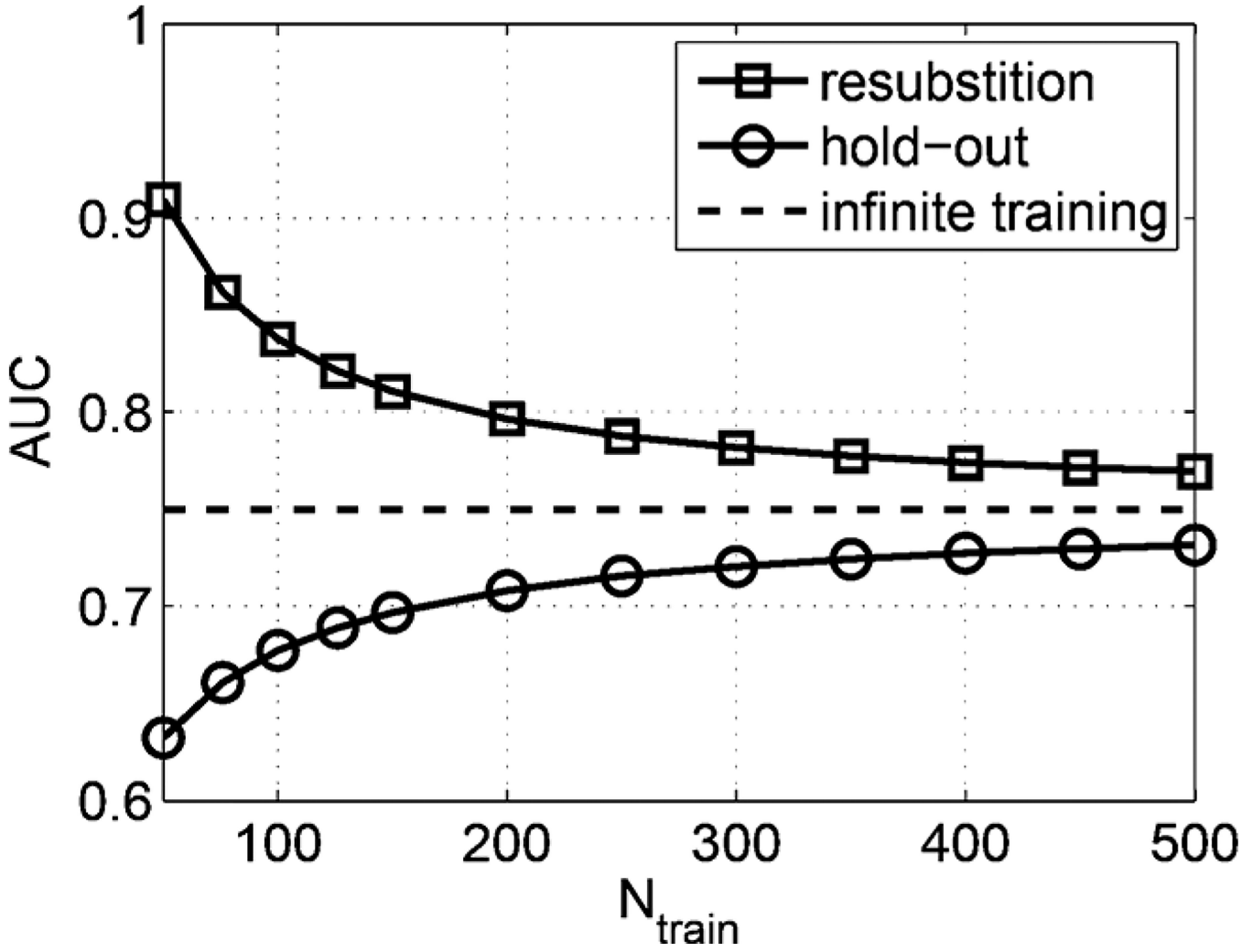

Fig. 4.

Mean estimated AUC obtained from the resubstitution and hold-out methods as a function of the number of training images, for a CHO with AUC = 0.75 and $p = 18$ channels.

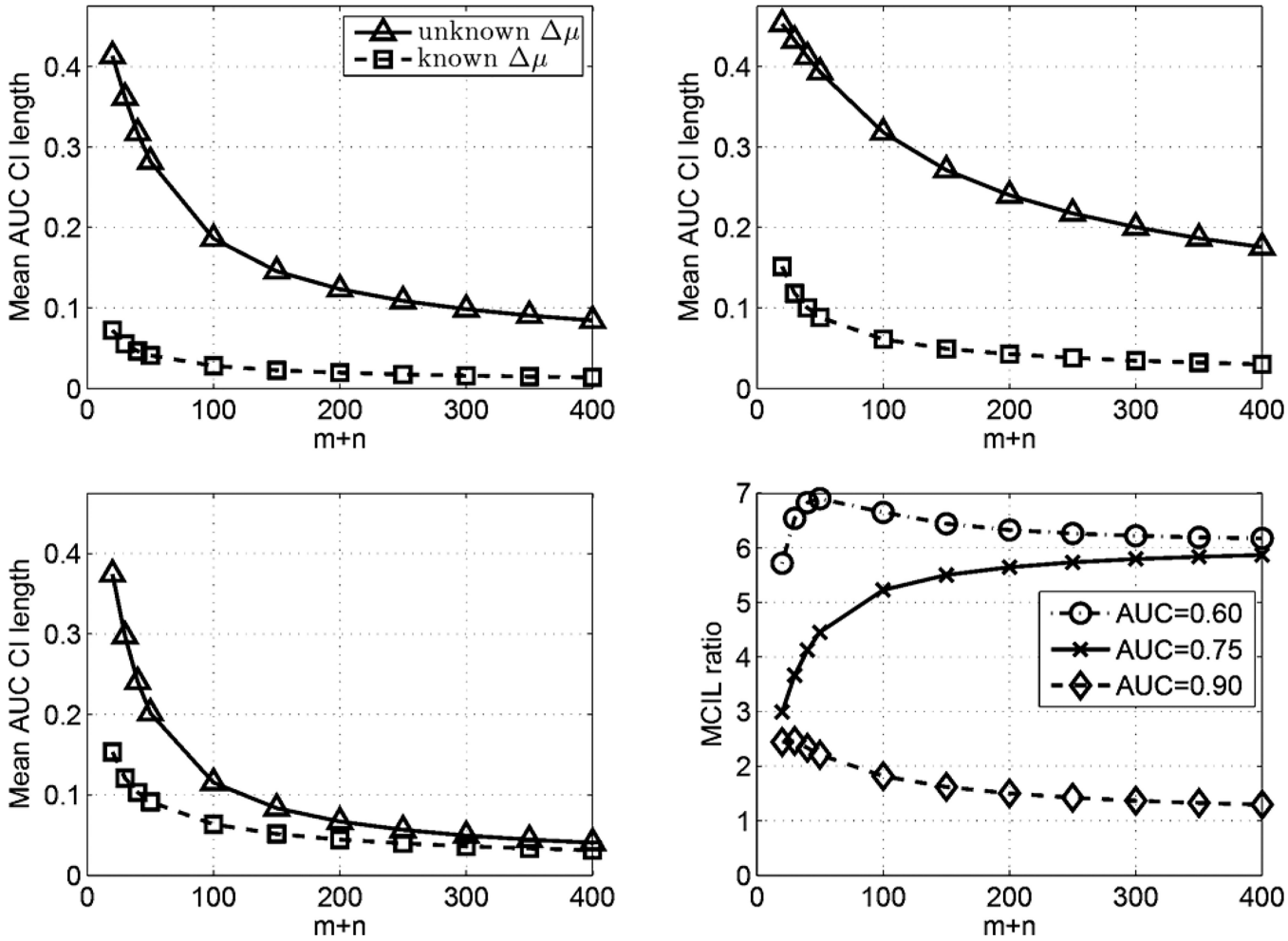

Fig. 5.

Mean confidence interval length (MCIL) for exact 95% AUC confidence intervals plotted versus sample size in the cases of unknown and known difference of class means, $\Delta\boldsymbol{\mu}$, for a CHO with $p = 5$ channels and $m = n$. Top Left: AUC = 0.6, Top Right: AUC = 0.75, Bottom Left: AUC = 0.9, Bottom Right: Ratio of MCIL for unknown $\Delta\boldsymbol{\mu}$ to MCIL for known $\Delta\boldsymbol{\mu}$ plotted versus sample size for AUC = 0.6, 0.75, and 0.9.

TABLE I

Estimated Coverage Probabilities for 95% SNR^2 Confidence Intervals

p	AUC	m	n	CP_{Wald}	CP_{Reiser}
5	0.60	150	50	0.991	0.950
5	0.60	100	100	0.991	0.952
5	0.75	150	50	0.968	0.950
5	0.75	100	100	0.965	0.950
5	0.90	150	50	0.959	0.950
5	0.90	100	100	0.957	0.950
20	0.60	150	50	0.843	0.950
20	0.60	100	100	0.854	0.950
20	0.75	150	50	0.922	0.951
20	0.75	100	100	0.932	0.951
20	0.90	150	50	0.956	0.951
20	0.90	100	100	0.959	0.951
50	0.60	150	50	0.204	0.950
50	0.60	100	100	0.233	0.951
50	0.75	150	50	0.584	0.950
50	0.75	100	100	0.660	0.952
50	0.90	150	50	0.860	0.950
50	0.90	100	100	0.895	0.950

IEEE Trans Med Imaging. Author manuscript; available in PMC 2017 August 03.

TABLE II

Estimated 96.67% AUC Confidence Intervals for the Three Reconstruction Methods in Example 1

A:	[0.8044	0.8828]
B:	[0.7947	0.8754]
C:	[0.6892	0.7917]

TABLE III

Estimated 95% AUC Confidence Intervals for the Two Reconstruction Methods in Example 2

short-scan	[0.8736	0.9267]
full-scan	[0.9282	0.9635]

TABLE IV

Estimated Coverage Probabilities of 95% AUC Confidence Intervals for an Infinitely-Trained CHO With $p = 18$ and AUC = 0.75

N_{train}	target AUC	Reiser	resubstitution test bootstrap	resubstitution train/test bootstrap	hold-out test bootstrap	hold-out train/test bootstrap
50	0.75	0.951	0.153	0	0.213	0.124
100	0.75	0.951	0.453	0.007	0.510	0.419
200	0.75	0.951	0.683	0.172	0.781	0.725
500	0.75	0.950	0.842	0.560	0.921	0.912
1000	0.75	0.949	0.897	0.754	0.942	0.950

TABLE V

Estimated Coverage Probabilities of 95% AUC Confidence Intervals for Mean Performance of a Finitely-Trained CHO With $p = 18$

N_{rain}	target AUC	Reiser	resubstitution test bootstrap	resubstitution train/test bootstrap	hold-out test bootstrap	hold-out train/test bootstrap
50	0.632	0.873	0.004	0	0.827	0.981
100	0.677	0.816	0.056	0	0.890	0.969
200	0.708	0.821	0.236	0.006	0.927	0.967
500	0.732	0.875	0.567	0.229	0.943	0.969
1000	0.741	0.908	0.748	0.518	0.946	0.966