Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Nonparametric estimation receiver operating characteristic analysis for performance evaluation on combined detection and estimation tasks

Adam Wunderlich
Bart Goossens

Nonparametric estimation receiver operating characteristic analysis for performance evaluation on combined detection and estimation tasks

Adam Wunderlich^{a,*} and Bart Goossens^b

^aU.S. Food and Drug Administration, Center for Devices and Radiological Health, Silver Spring, Maryland, United States

^bGhent University, Department of Telecommunications and Information Processing, TELIN-IPI-iMinds, Ghent, Belgium

Abstract. In an effort to generalize task-based assessment beyond traditional signal detection, there is a growing interest in performance evaluation for combined detection and estimation tasks, in which signal parameters, such as size, orientation, and contrast are unknown and must be estimated. One motivation for studying such tasks is their rich complexity, which offers potential advantages for imaging system optimization. To evaluate observer performance on combined detection and estimation tasks, Clarkson introduced the estimation receiver operating characteristic (EROC) curve and the area under the EROC curve as a summary figure of merit. This work provides practical tools for EROC analysis of experimental data. In particular, we propose nonparametric estimators for the EROC curve, the area under the EROC curve, and for the variance/covariance matrix of a vector of correlated EROC area estimates. In addition, we show that reliable confidence intervals can be obtained for EROC area, and we validate these intervals with Monte Carlo simulation. Application of our methodology is illustrated with an example comparing magnetic resonance imaging k -space sampling trajectories. MATLAB® software implementing the EROC analysis estimators described in this work is publicly available at <http://code.google.com/p/iqmodel/>. © 2014 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: 10.1117/1.JMI.1.3.031002]

Keywords: image quality; receiver operating characteristic; U -statistics; confidence intervals.

Paper 14037SSPR received Apr. 2, 2014; revised manuscript received Jul. 2, 2014; accepted for publication Jul. 7, 2014; published online Aug. 26, 2014.

1 Introduction

Task-based image quality assessment is a principled and objective approach to the assessment of imaging systems. In a task-based approach, the figure of merit measures the ability of an observer to perform an image-evaluation task.¹ Most of the literature on task-based image assessment has focused on detection and discrimination tasks, using the receiver operating characteristic (ROC) curve, or related variants, to measure performance.² Among the most popular alternatives to ROC curves are localization ROC (LROC)³ and free-response ROC (FROC)⁴ curves, which assess combined signal detection and localization performance. Although these alternatives offer more richness than traditional ROC analysis, they do not characterize performance on general combined detection and estimation tasks in which signal parameters, such as size, orientation, or contrast, are unknown and must be estimated.

Recently, there has been an increased interest in general combined detection and estimation tasks^{5–7} because of their greater complexity compared to traditional signal detection tasks. Indeed, this additional complexity offers potential advantages for imaging system optimization through the concurrent optimization of detection and estimation performance. To assess observer performance on combined detection and estimation tasks, Clarkson⁵ proposed the estimation ROC (EROC) curve, which plots the expected utility of a true-positive (U_{TP}) (combining both estimation accuracy and detection performance) versus the false-positive fraction (FPF). Also, as a summary

figure of merit, Clarkson⁵ proposed the area under the EROC curve, which we denote as $\mathcal{A}_E$.

In order for EROC curves to be useful in experimental settings, it is necessary to have practical estimators for the EROC curve and its area. Moreover, because statistical variability arises from typical experimental sample sizes, variance estimators, and procedures for statistical inference are also needed. However, such estimators were not given by Clarkson,⁵ and are not presently available to the researchers that would like to apply EROC analysis to experimental data. Here, we address this need by building on existing nonparametric theory for ROC analysis.^{8–10} In particular, we introduce nonparametric estimators for the EROC curve, the area under the EROC curve, and for the variances and covariances of correlated $\mathcal{A}_E$ estimates. Moreover, we show how confidence intervals can be obtained for $\mathcal{A}_E$, and validate these intervals with Monte Carlo simulations. Our methodology is demonstrated with an example in which the channelized joint observer (CJO) of Goossens et al.^{6,7} is used to compare different magnetic resonance imaging (MRI) k -space trajectories.

2 EROC Curves

Consider a combined detection and estimation task in which each image either contains no signals or contains one signal that is described by a (possibly random) parameter vector, θ . In this task, the observer is asked to decide whether or not a signal is present and to estimate θ . According to the usual decision model, it is assumed that the observer's decision is made by

*Address all correspondence to: Adam Wunderlich, E-mail: adam.wunderlich@fda.hhs.gov

thresholding a confidence score that is higher when the signal is believed to be present and lower otherwise. Let X and Y be continuous random variables denoting the confidence rating for signal-absent and signal-present images, respectively. Further, let $\hat{\theta}$ be the observer's estimate of θ .

Before we introduce the definition of the EROC curve, it is helpful to first review the definition of the ROC curve.¹ Recall that the ROC curve plots the true-positive fraction (TPF), the probability of a true-positive given that the signal is present, versus the FPF, the probability of a false-positive given that the signal is absent, over all decision thresholds, c . In terms of the confidence ratings X and Y , TPF and FPF can be expressed as $\text{TPF}(c) = P(Y \geq c) = E[\mathcal{I}(Y \geq c)]$ and $\text{FPF}(c) = P(X \geq c) = E[\mathcal{I}(X \geq c)]$, where $P(\cdot)$ and $E[\cdot]$ denote probability and expectation, respectively; the indicator function, $\mathcal{I}(\cdot)$, is defined as $\mathcal{I}(S) = 1$ if the proposition S is true and $\mathcal{I}(S) = 0$ otherwise.

Clarkson⁵ defined the EROC curve as the plot of the expected utility of a true-positive given that the signal is present, denoted U_{TP} , versus the FPF, over all decision thresholds. The expected utility of a true-positive is defined in terms of a utility function, $u(\hat{\theta}, \theta)$ that is higher for an accurate parameter estimate and lower for an inaccurate estimate. Specifically, for a given threshold c , the expected utility of a true-positive is defined to be $U_{\text{TP}}(c) = E[u(\hat{\theta}, \theta)\mathcal{I}(Y \geq c)]$, where the expectation is taken over the random variables Y , $\hat{\theta}$, and θ (if applicable). Comparing this definition to the expression for TPF given in the previous paragraph, note that U_{TP} is more general due to the inclusion of the utility function, $u(\hat{\theta}, \theta)$. A useful property of the EROC curve is that its value at $\text{FPF} = 1$ is equal to the mean utility of the parameter estimate, since $U_{\text{TP}}(\text{FPF} = 1) = U_{\text{TP}}(c = -\infty) = E[u(\hat{\theta}, \theta)]$. An example of an EROC curve is shown in Fig. 1.

An important aspect of the EROC curve concerns the choice of the utility function. For ease of interpretation, it is desirable to limit the EROC ordinate to the interval $[0, 1]$, which is guaranteed if $0 \leq u(\hat{\theta}, \theta) \leq 1$. Later, in Secs. 4 and 5, we will use the utility function

$$u(\hat{\theta}, \theta) = \begin{cases} 1 - \left(\frac{\|\hat{\theta} - \theta\|}{\|\theta\|}\right)^2 & \text{if } \|\hat{\theta} - \theta\| \leq \|\theta\| \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

Fig. 1 Example of an estimation receiver operating characteristic (EROC) curve with the utility function limited to the interval $[0, 1]$.

which can be interpreted as one minus the normalized squared error, truncated to be non-negative. Of course, the above utility function is merely one possible option among many. Further considerations regarding appropriate utility functions for specific applications are beyond the scope of this work.

It is interesting to note that both the ROC curve and the LROC curve are special cases of the EROC curve. Namely, if the utility function is identically equal to one, i.e., $u(\hat{\theta}, \theta) \equiv 1$, then the EROC ordinate is equal to the TPF, and the EROC curve is identical to the ROC curve. To recover the LROC curve, let θ be a vector of signal coordinates, let $\mathcal{L}$ be the event that a signal is correctly localized within an acceptable distance, R , and define the utility function as the indicator function for correct signal localization, i.e., $u(\hat{\theta}, \theta) = \mathcal{I}(\mathcal{L}) = \mathcal{I}(\|\hat{\theta} - \theta\| \leq R)$. In this case, the EROC ordinate becomes $U_{\text{TP}}(c) = P(Y \geq c \cap \mathcal{L})$, which is the ordinate of the LROC curve,¹⁰ and the EROC curve is the same as the LROC curve.

As a summary figure of merit for the EROC curve, Clarkson⁵ proposed the area under the EROC curve, $\mathcal{A}_E$, and showed that it can be expressed as

$$\mathcal{A}_E = E[u(\hat{\theta}, \theta)\mathcal{I}(Y > X)], \quad (2)$$

where the expectation is taken over X , Y , $\hat{\theta}$, and θ (if applicable). Just as the EROC curve generalizes the ROC curve, the area under the EROC curve is a generalization of the area under the ROC curve. This can be seen from the above equation, since in the special case that the utility function is identically equal to one, $\mathcal{A}_E = E[\mathcal{I}(Y > X)] = P(Y > X)$, which is the area under the ROC curve.^{8,11}

3 Nonparametric Estimators

3.1 Empirical EROC Curve and its Area

Suppose that the observer generates m independent, identically distributed (i.i.d.) ratings for signal-absent images, $X_1, X_2, \dots, X_m$, and n i.i.d. ratings and parameter estimates for signal-present images, denoted as $Y_1, Y_2, \dots, Y_n$ and $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_n$, respectively. In addition, suppose that each signal-present image has a parameter vector θ_j for $j = 1, 2, \dots, n$. For a given utility function $u(\hat{\theta}, \theta)$, the EROC coordinates $U_{\text{TP}}(c)$ and $\text{FPF}(c)$ can be estimated with the unbiased estimators

$$\hat{U}_{\text{TP}}(c) = \frac{1}{n} \sum_{j=1}^n u(\hat{\theta}_j, \theta_j) \mathcal{I}(Y_j \geq c) \quad (3)$$

and

$$\widehat{\text{FPF}}(c) = \frac{1}{m} \sum_{i=1}^m \mathcal{I}(X_i \geq c), \quad (4)$$

respectively. Plotting $\hat{U}_{\text{TP}}(c)$ versus $\widehat{\text{FPF}}(c)$ over all thresholds, c , we obtain an unbiased estimate of the EROC curve, which we call the empirical EROC curve, by analogy to the empirical ROC curve.¹¹

The area under the EROC curve, $\mathcal{A}_E$, can be estimated with the nonparametric estimator

$$\hat{\mathcal{A}}_E = \frac{1}{mn} \sum_{j=1}^n \sum_{i=1}^m u(\hat{\theta}_j, \theta_j) \mathcal{I}(Y_j > X_i). \quad (5)$$

From Eq. (2), it follows that the above estimator is unbiased. In fact, this estimator is a two-sample U -statistics with a kernel, $u(\hat{\theta}_j, \theta_j) \mathcal{I}(Y_j > X_i)$, that is a function of X_i and the vector $[Y_j, \hat{\theta}_j, \theta_j]$ for $i \in \{1, 2, \dots, m\}$, $j \in \{1, 2, \dots, n\}$. Moreover, by a classical theorem for U -statistics,¹² $\hat{\mathcal{A}}_E$ is the minimum variance unbiased estimator over a large class of distributions. Observe that if $u(\hat{\theta}_j, \theta_j) \equiv 1$, then the estimator in Eq. (5) is equivalent to the classical Mann–Whitney U estimate for the area under the ROC curve.^{8,11}

As noted by Clarkson,⁵ $\mathcal{A}_E$ can also be estimated using a two-alternative forced choice (2AFC) experiment. In such an experiment, the observer is shown random pairs of images, where each pair consists of one signal-absent image and one signal-present image. For each pair of images, the observer is asked to (i) choose the signal-present image and (ii) provide an estimate, $\hat{\theta}$, of the signal parameter vector, θ . An unbiased estimate of $\mathcal{A}_E$ can then be obtained by summing the utility of each parameter estimate over the set of image pairs for which the signal-present image was correctly chosen, and then dividing by the total number of image pairs. Mathematically, for n independent image pairs, the 2AFC estimator for $\mathcal{A}_E$, denoted $\widehat{\text{FC}}_E$ takes the form

$$\widehat{\text{FC}}_E = \frac{1}{n} \sum_{j=1}^n u(\hat{\theta}_j, \theta_j) \mathcal{C}_j, \quad (6)$$

where $\mathcal{C}_j = 1$ if the signal-present image is correctly selected in the j 'th image pair and $\mathcal{C}_j = 0$ otherwise. The tradeoffs between Eqs. (5) and (6) are analogous to those between the Mann–Whitney U and the 2AFC estimators for the area under the ROC curve.¹³ Namely, the 2AFC estimator for the area under the ROC curve,¹ $\widehat{\text{FC}}_E$, has potential advantages for human-reader studies in which the number of images is not constrained and the limiting factor is reading time, since the reader does not have to report image ratings. On the other hand, for studies in which the total number of images is the constraining factor, $\hat{\mathcal{A}}_E$ is preferable because it includes more image comparisons, and, consequently, has lower variability. For this reason, the remainder of the paper focuses on EROC analysis with $\hat{\mathcal{A}}_E$. Nonetheless, it is useful to note that statistical analysis of the 2AFC approach is straightforward and can be carried out with classical methods, since $\widehat{\text{FC}}_E$ is simply a sample mean.

3.2 Nonparametric Covariance Estimation

Now, suppose that we seek to compare the correlated EROC area estimates corresponding to q fixed imaging scenarios. Here, a “fixed” imaging scenario is a nonrandom imaging condition, in which the observer is also nonrandom (fixed). Examples of fixed imaging scenarios include different image reconstruction algorithms, different scanner settings, or different fixed observers. We denote the ratings and parameter estimates corresponding to these scenarios as $X_i^{(k)}$, $Y_j^{(k)}$, and $\hat{\theta}_j^{(k)}$ for $i = 1, 2, \dots, m$, $j = 1, 2, \dots, n$, and $k = 1, 2, \dots, q$, respectively. To simplify our notation, let $\psi_{ij}^{(k)} = u(\hat{\theta}_j^{(k)}, \theta_j) \mathcal{I}(Y_j^{(k)} > X_i^{(k)})$, so that the estimated area under the EROC curve for the k 'th scenario is

$$\hat{\mathcal{A}}_E^{(k)} = \frac{1}{mn} \sum_{j=1}^n \sum_{i=1}^m \psi_{ij}^{(k)}. \quad (7)$$

Similar to the method of DeLong et al.⁹ for ROC analysis, we can estimate the covariance matrix of the vector

$[\hat{\mathcal{A}}_E^{(1)}, \hat{\mathcal{A}}_E^{(2)}, \dots, \hat{\mathcal{A}}_E^{(q)}]$ by applying the general approach of Sen,^{14,15} as described below.

First, define the “structural components,”

$$V_{10}^{(k)}(i) = \frac{1}{n} \sum_{j=1}^n \psi_{ij}^{(k)}, \quad i = 1, 2, \dots, m \quad (8)$$

and

$$V_{01}^{(k)}(j) = \frac{1}{m} \sum_{i=1}^m \psi_{ij}^{(k)}, \quad j = 1, 2, \dots, n. \quad (9)$$

Note that each structural component has mean $\mathcal{A}_E^{(k)}$. Moreover, it can be shown that the structural components are asymptotically uncorrelated.¹⁰ Next, denote the sample covariances of the structural components as

$$s_{10}^{uv} = \frac{1}{m-1} \sum_{i=1}^m [V_{10}^{(u)}(i) - \hat{\mathcal{A}}_E^{(u)}][V_{10}^{(v)}(i) - \hat{\mathcal{A}}_E^{(v)}] \quad (10)$$

$$s_{01}^{uv} = \frac{1}{n-1} \sum_{j=1}^n [V_{01}^{(u)}(j) - \hat{\mathcal{A}}_E^{(u)}][V_{01}^{(v)}(j) - \hat{\mathcal{A}}_E^{(v)}]. \quad (11)$$

Finally, following Sen,¹⁵ the desired covariance between $\hat{\mathcal{A}}_E^{(u)}$ and $\hat{\mathcal{A}}_E^{(v)}$ can be estimated as

$$s^{uv} = \frac{s_{10}^{uv}}{m} + \frac{s_{01}^{uv}}{n}. \quad (12)$$

Let $\mathbf{S}$, $\mathbf{S}_{10}$, and $\mathbf{S}_{01}$ be matrices with (u, v) entries given by s^{uv} , s_{10}^{uv} , and s_{01}^{uv} , respectively, for $u, v \in \{1, 2, \dots, q\}$. With this notation, the nonparametric Sen estimator for the $q \times q$ covariance matrix of the vector $[\hat{\mathcal{A}}_E^{(1)}, \hat{\mathcal{A}}_E^{(2)}, \dots, \hat{\mathcal{A}}_E^{(q)}]^T$ is expressed as $\mathbf{S} = \mathbf{S}_{10}/m + \mathbf{S}_{01}/n$. Note that since the matrices $\mathbf{S}_{10}$ and $\mathbf{S}_{01}$ can be written as sums of vector outer products, they are semipositive definite. Consequently, the covariance matrix estimator, $\mathbf{S}$, is also semipositive definite.

The following theorem characterizes the asymptotic behavior of the above covariance estimator. The proof is omitted here, since it is nearly identical to the proof of a similar result in Wunderlich and Noo,¹⁰ with the only change being the different U -statistic kernel.

Theorem 1. Suppose that $m/(m+n) = a > 0$ and $n/(m+n) = b > 0$, where a and b are fixed constants that are independent of $m+n$. Also, suppose that $E[(\psi_{ij}^{(k)})^2] < \infty$ for all i, j , and k . Then as $m+n \rightarrow \infty$, the estimator s^{uv} converges with probability one to $\text{Cov}[\hat{\mathcal{A}}_E^{(u)}, \hat{\mathcal{A}}_E^{(v)}]$, i.e.,

$$P\left(\lim_{m+n \rightarrow \infty} s^{uv} = \text{Cov}[\hat{\mathcal{A}}_E^{(u)}, \hat{\mathcal{A}}_E^{(v)}]\right) = 1.$$

In the terminology of estimation theory, the above theorem shows that the Sen estimator for covariance is strongly consistent.¹⁶ Combining this result with a standard theorem on asymptotic normality for a vector of generalized U -statistics yields the next theorem, which justifies the construction of confidence intervals or confidence regions using the Sen covariance estimator and asymptotic normality. Again, the proof is omitted, since

it is essentially identical to that for a similar result in Wunderlich and Noo.¹⁰

Theorem 2. Let $\hat{\mathcal{A}}_E = [\hat{\mathcal{A}}_E^{(1)}, \hat{\mathcal{A}}_E^{(2)}, \dots, \hat{\mathcal{A}}_E^{(q)}]^T$ and $\mathcal{A}_E = [\mathcal{A}_E^{(1)}, \mathcal{A}_E^{(2)}, \dots, \mathcal{A}_E^{(q)}]^T$, and suppose that $E[(\psi_{ij}^{(k)})^2] < \infty$ for all i, j , and k . Then as $m + n \rightarrow \infty$, the random vector $S^{-1/2}(\hat{\mathcal{A}}_E - \mathcal{A}_E)$ converges in distribution to a multivariate normal vector with a mean zero and identity covariance matrix.

4 Monte Carlo Evaluation

The theoretical results presented in the previous section indicate that the estimators are reliable in an asymptotic sense, i.e., if the sample size is large enough. However, as with all asymptotic results, it is not clear if the estimators are dependable for typical experimental sample sizes. Therefore, to provide further evidence for the effectiveness of the estimators presented in the previous section, a Monte Carlo simulation study was carried out to evaluate the coverage probability of confidence intervals for a difference of $\mathcal{A}_E$ values. We begin with a description of the simulation model, followed by the evaluation details.

4.1 Simulation Model

For a typical EROC data set, in which the same set of test cases is used for each imaging scenario, the confidence scores and parameter estimates will be correlated across scenarios. Moreover, the signal parameter estimates will generally be correlated with the signal-present confidence scores. Therefore, our simulation model is designed to include these correlations. As in the previous section, we assume that the only source of variability is due to the case set and that the imaging scenario is fixed.

For the i 'th signal-absent image under the k 'th imaging scenario, we model the confidence rating as

$$X_i^{(k)} = g(c_i + e_{ik}), \quad (13)$$

where $g(x)$ is a transformation function to be specified and c_i and e_{ik} are independent, normally distributed random variables with zero means and variances, σ_c^2 and σ_e^2 , respectively. Here, c_i is a random factor for the case effect and e_{ik} is an error term. Observe that since c_i is the same across scenarios, it induces a correlation between scenarios for each case.

We simulate signal-present ratings and parameter estimates with a similar model. However, since the parameter estimates are generally correlated with the signal-present ratings for the same case, a vector-valued model is used. For simplicity, we take the signal parameter vector, θ , to be a one-dimensional, nonrandom scalar, written as θ . However, it is straightforward to generalize our model to multidimensional parameter vectors. The vector $\mathbf{Z}_j^{(k)} = [Y_j^{(k)}, \hat{\theta}_j^{(k)}]$ for the j 'th signal-present image under the k 'th imaging scenario is modeled as

$$\mathbf{Z}_j^{(k)} = \mathbf{h}(\boldsymbol{\mu}_k + \mathbf{b}_j + \mathbf{c}_j + \mathbf{e}_{jk}), \quad (14)$$

where $\mathbf{h}$ is a vector-valued function of a vector that is computed by applying the transformation function, g , from Eq. (13) to each component of the argument, i.e., $\mathbf{h}(\mathbf{x}) = [g(x), g(y)]$ for $\mathbf{x} = [x, y]$. Above, $\boldsymbol{\mu}_k = [\lambda_k, 0]$ is a fixed factor for the imaging scenario effect in $Y^{(k)}$, $\mathbf{b}_j = [0, \nu_j]$ is a measurement modeling term, which can be either fixed or random, that connects the distribution of $\hat{\theta}$ to the distribution of θ (specified below), $\mathbf{c}_j$ is a random factor for the case effect, and $\mathbf{e}_{jk}$ is a random error term. The random factors $\mathbf{c}_j$ and $\mathbf{e}_{jk}$ are taken to be distributed as independent bivariate normal distributions with zero means and covariance matrices

$$\Sigma_c = \begin{bmatrix} \sigma_c^2 & \rho_c \sigma_c \sigma_{\hat{\theta}_c} \\ \rho_c \sigma_c \sigma_{\hat{\theta}_c} & \sigma_{\hat{\theta}_c}^2 \end{bmatrix} \quad \text{and} \quad \Sigma_e = \begin{bmatrix} \sigma_e^2 & \rho_e \sigma_e \sigma_{\hat{\theta}_e} \\ \rho_e \sigma_e \sigma_{\hat{\theta}_e} & \sigma_{\hat{\theta}_e}^2 \end{bmatrix}, \quad (15)$$

respectively. As in Eq. (13), since $\mathbf{c}_i$ is the same across scenarios, it induces a correlation between scenarios for each case. Furthermore, correlations between Y and $\hat{\theta}$ are determined by the correlation coefficients ρ_c and ρ_e .

Our simulations used two variations of the simulation model that we call model I and model II. Each of these models specifies three ingredients: the functional form for $g(x)$, the model for ν , and the model for θ . Below, we describe these models, as summarized in Table 1. To keep the number of parameters small, both models use a fixed constant, ω , to parameterize the model for ν .

For model I, we set $\nu = \omega$ and take $g(x)$ to be the identity function, so that the scores and parameter estimates are normally distributed. In addition, we define $\theta = \nu = \omega$, i.e., θ is a fixed constant equal to the mean of $\hat{\theta}$. For model II, ν is defined as a random variable following a uniform distribution on the interval $[\omega - 0.1, \omega + 0.1]$ denoted as $\nu \sim \mathcal{U}[\omega - 0.1, \omega + 0.1]$. Moreover, we take $g(x) = \exp(x)$, so that the scores and parameter estimates follow a log-normal distribution when conditioned on ν , and we define $\theta = \exp[\nu + (\sigma_{\hat{\theta}_c}^2 + \sigma_{\hat{\theta}_e}^2)/2]$. Note that since ν is a random variable under model II, θ is also random. Moreover, θ is defined so that $\hat{\theta}^{(k)}$ and θ have the same mean. Namely, from the equation for the mean of a log-normal distribution, $E[\hat{\theta}^{(k)}|\nu] = \exp[\nu + (\sigma_{\hat{\theta}_c}^2 + \sigma_{\hat{\theta}_e}^2)/2]$, and hence, $E[\hat{\theta}^{(k)}] = E[E[\hat{\theta}^{(k)}|\nu]] = E[\exp[\nu + (\sigma_{\hat{\theta}_c}^2 + \sigma_{\hat{\theta}_e}^2)/2]] = E[\theta]$.

For a given sample size, determined by m and n , both of the model variations described above require specification of nine parameters to model $q = 2$ imaging scenarios that we denote with the letters A and B . Namely, the model for X requires two parameters: σ_c^2 and σ_e^2 . Furthermore, the model for $\mathbf{Z}$ requires two scenario effect parameters: λ_A and λ_B , one parameter for the ν model: ω , plus four additional parameters for the

Table 1 Summary of simulation models. Above, ω is a fixed constant. The signal parameter, θ , is fixed in model I and random in model II.

Model	$g(x)$	Distribution	ν model	θ model
I	Identity	Normal	$\nu = \omega$	$\theta = \omega$
II	$\exp(x)$	Log-normal	$\nu \sim \mathcal{U}[\omega - 0.1, \omega + 0.1]$	$\theta = \exp[\nu + (\sigma_{\hat{\theta}_c}^2 + \sigma_{\hat{\theta}_e}^2)/2]$

covariance matrices: ρ_c , $\sigma_{\hat{\theta}_c}$, ρ_e , and $\sigma_{\hat{\theta}_e}$. Thus, for two imaging scenarios and a given sample size, a total of $2 + 2 + 1 + 4 = 9$ parameters are required.

4.2 Confidence Interval Validation

A Monte Carlo evaluation was carried out in which random EROC ratings and signal parameter estimates were generated with both variations of the simulation model described in the previous section, using the utility function defined in Eq. (1) for two imaging scenarios denoted A and B . Writing the difference of EROC areas as $\Delta\mathcal{A}_E = \mathcal{A}_E^B - \mathcal{A}_E^A$, and the estimate of this difference as $\Delta\hat{\mathcal{A}}_E = \hat{\mathcal{A}}_E^B - \hat{\mathcal{A}}_E^A$, a 95% confidence interval for $\Delta\mathcal{A}_E$ was estimated using a normal approximation as

$$\Delta\hat{\mathcal{A}}_E \pm 1.96\sqrt{\widehat{\text{Var}}(\Delta\hat{\mathcal{A}}_E)}. \quad (16)$$

Above, $\widehat{\text{Var}}(\Delta\hat{\mathcal{A}}_E)$ was an estimate of $\text{Var}(\Delta\hat{\mathcal{A}}_E)$, obtained by estimating each term in

$$\text{Var}(\Delta\hat{\mathcal{A}}_E) = \text{Var}(\hat{\mathcal{A}}_E^A) + \text{Var}(\hat{\mathcal{A}}_E^B) - 2\text{Cov}(\hat{\mathcal{A}}_E^A, \hat{\mathcal{A}}_E^B) \quad (17)$$

with the covariance estimator described in Sec. 3.2.

For the evaluation, EROC data sets were generated for a large number of parameter combinations. Specifically, for sample sizes of $(m, n) = (25, 50)$, $(50, 50)$, $(75, 75)$, and $(50, 150)$, we took $\lambda_A, \lambda_B, \omega \in \{.25, 1.5\}$, $\sigma_c, \sigma_{\hat{\theta}_c}, \sigma_e, \sigma_{\hat{\theta}_e} \in \{.25, .5\}$, and $\rho_c, \rho_e \in \{.2, .8\}$. Therefore, for each sample size, the nine model parameters each took one of two values, for a total of $2^9 = 512$ parameter combinations. Under model I, this set of parameter combinations yielded $\Delta\mathcal{A}_E$ values ranging from -0.33 to 0.33 , with $\mathcal{A}_E$ for each scenario ranging from 0.11 to 0.94 . By contrast, under model II, $\Delta\mathcal{A}_E$ ranged from -0.31 to 0.31 with $\mathcal{A}_E$ for each scenario ranging from 0.41 to 0.87 .

For each parameter combination, the coverage probability of the approximate 95% confidence interval in Eq. (16) was estimated as the proportion out of 10,000 Monte Carlo trials for which the interval covered the true $\Delta\mathcal{A}_E$, which was estimated as the mean of $\Delta\hat{\mathcal{A}}_E$ over all trials. Since each coverage probability estimate was an estimate of a binomial proportion,¹⁷ the

standard deviation was $\sqrt{\text{CP}(1 - \text{CP})/N}$, where CP is the true coverage probability and N is the number of Monte Carlo trials. Hence, in this evaluation, the standard deviation for each coverage probability estimate was roughly $\sqrt{(0.95)(0.05)/(10000)} \approx 0.002$.

Figure 2 contains box plots summarizing the simulation results for the 512 model parameter combinations at each sample size for models I and II. These plots were created with the MATLAB® command “box-plot.” To interpret the plots, note that the lower and upper edges of each box are the 25% and 75% percentiles and that the horizontal line inside the box is the median. The length of each whisker is at most 1.5 times the distance between the 75% and the 25% percentiles, and the data points outside this range are plotted individually with crosses. From these plots, we see that for most of the 512 parameter combinations considered, with the exception of a few outliers, the observed coverage probabilities were within ± 0.01 of the desired 0.95. Moreover, as expected, the accuracy of the confidence intervals improved with increasing sample size. Therefore, these validation experiments provide evidence that the estimated confidence intervals for $\Delta\mathcal{A}_E$ are generally reliable for typical experimental samples sizes, complementing the asymptotic theory in Sec. 3.2.

5 Example

We demonstrate our EROC estimators and the $\mathcal{A}_E$ figure of merit with an example comparing the quality of images reconstructed from different MRI k -space sampling trajectories. First, six T1-weighted head images of size $256 \times 256 \times 256$ were acquired from repeated scans of a single test subject with a 3T SIEMENS Magnetom Tim Trio MRI. From each image volume, two adjacent, central sagittal slices of size 256×256 were selected, and signals were simulated in one of these slices by inserting a rotationally symmetric, Gaussian signal with a scale parameter (standard deviation) equal to 1 pixel at 13 specified locations; see Fig. 3. After these insertions, the slices were resampled by applying a forward nonuniform Fourier transform¹⁸ to the reconstructed fully sampled k -space data to obtain k -space samples along an Archimedean spiral, with coordinates given by

Fig. 2 Box plots for model I (a) and model II (b) summarizing the estimated coverage probability of approximate 95% confidence intervals for $\mathcal{A}_E^B - \mathcal{A}_E^A$. Each box plot depicts the results for 512 combinations of the simulation parameters for a given sample size (m, n) .

Fig. 3 Examples of magnetic resonance imaging (MRI) images for the experiment: (a) original, fully sampled image, (b) fully sampled image with 13 signals inserted, (c) reconstructed image for acquisition A, and (d) reconstructed image for acquisition B.

$$\mathbf{p}_i = \frac{r\phi_i}{2\pi} \left\{ \cos \left[\phi_i + \pi \left(\frac{1+r}{r} \right) \right], \quad \sin \left[\phi_i + \pi \left(\frac{1+r}{r} \right) \right] \right\}, \quad (18)$$

with $\phi_i = \frac{\pi}{r(N-1)}i$ for $i = 0, \dots, N-1$, where N is the number of samples, which was set to $N = 256^2/4$ (25% of the Nyquist rate), and r is the radius parameter of the spiral, which was set to either $r = 1.2/256$ or $r = 1.21/256$. Note that due to the k -space resampling and the reconstruction algorithm, the signal was altered in an unpredictable way, so that it could be considered to be random. In other words, randomness in the images resulted not only from measurement noise in k -space, but also from unpredictable aspects of the k -space resampling and a reconstruction algorithm. We call the two acquisitions corresponding to the different choices of the radius parameter acquisition A and acquisition B, respectively. After resampling in k -space, images were subsequently reconstructed using the method of Aelterman et al.¹⁸ Observe that since the two k -space acquisitions were obtained by resampling the same set of images, the reconstructions for acquisitions A and B were statistically correlated.

Nonoverlapping patches of size, 32×32 centered around the inserted signals, were extracted from the reconstructed images. For a given k -space trajectory, this yielded a total of $13 \times 6 \times 2 = 156$ image patches, of which 78 were signal-present and 78 were signal-absent. For the purpose of

illustrating our statistical methodology, the image patches were assumed to be statistically independent. (Since it is possible that neighboring image patches were weakly correlated, this assumption represents a potential limitation of the example.) Given an image patch, the observer was tasked with deciding whether or not the signal was present and with estimating the signal size (θ), as described by a Gaussian scale parameter.

For the evaluation, a numerical observer was used. Namely, each image patch was processed with the CJO described in Goossens et al.⁶ and Zhang et al.,⁷ yielding a confidence score and an estimated signal size. The CJO observer model^{6,7} assumes that a known parametric signal is degraded by the image formation process in an unknown way. In this example, the degraded Gaussian signal was modeled as a Gaussian function with an increased, but unknown signal size. The CJO was implemented with five scale-shiftable channels and computed the confidence score with a simple matched filter¹ in the channel space, i.e., each score was calculated as the inner product of the estimated degraded signal and each individual image patch after projecting onto the scale-shiftable channels. The obtained confidence scores, together with the estimated signal sizes, were used to estimate the area under the EROC curve with Eq. (5). Note that because the background noise statistics were not needed in this approach, a training step was not required for the observer. For further details on the CJO model, see Ref. 6.

Fig. 4 Empirical EROC curves (a) and ROC curves (b) for two MRI acquisitions with different k -space trajectories, estimated from a sample of 156 image patches.

Empirical EROC and ROC curves for the two acquisitions are plotted in Fig. 4, where the EROC curve used the utility function in Eq. (1). From the plots, we see that the empirical EROC curve for acquisition B is uniformly higher than the EROC curve for acquisition A . By contrast, the empirical ROC curves for both acquisitions are very close. Table 2 lists the estimated values for the area under the EROC curve, $\mathcal{A}_E$, the area under the ROC curve, $\mathcal{A}$, and the mean utility of the signal size estimates for each acquisition. Using the covariance estimator introduced in Sec. 3, a 95% confidence interval for the difference of EROC areas, $\mathcal{A}_E^{(B)} - \mathcal{A}_E^{(A)}$, was found to be $[0.097, 0.202]$. Similarly, using the method of DeLong et al.,⁹ a 95% confidence interval for the difference of ROC areas, $\mathcal{A}^{(B)} - \mathcal{A}^{(A)}$, was found to be $[-0.031, 0.056]$. Thus, a statistically significant difference was observed for $\mathcal{A}_E$, but not for $\mathcal{A}$.

Recall that acquisition A used a k -space sampling trajectory with a smaller radius parameter than acquisition B . Consequently, acquisition A sampled lower frequencies more densely, whereas acquisition B placed more samples at higher frequencies. As a result of this sampling, acquisition A was expected to result in greater image blur compared to acquisition B . Indeed, this blur is evident in the mean utility of the estimated signal size, which was observed to be smaller for acquisition A than for acquisition B . The blurring introduced by the denser sampling of lower frequencies by acquisition A can also be seen by comparing the bottom left and bottom right images of Fig. 3. In terms of signal detection, the additional blurring of acquisition A had little effect, since the estimated ROC curves were similar for both acquisitions. By contrast, the additional blurring of acquisition A was reflected in the estimated

Table 2 For the two acquisitions, point estimates for the area under the EROC curve, the area under the ROC curve, and the mean utility of the signal size estimates. Estimated standard deviations are shown to the right of each point estimate and 95% confidence intervals for the interacquisition difference are given in the last row. All estimates are based on a sample of 156 image patches.

	EROC area	ROC area	Mean utility
Acquisition A	0.215 ± 0.031	0.757 ± 0.040	0.364 ± 0.046
Acquisition B	0.364 ± 0.042	0.770 ± 0.039	0.517 ± 0.051
95% CI for $B - A$	$[0.097, 0.202]$	$[-0.031, 0.056]$	$[0.095, 0.211]$

Fig. 5 Empirical EROC curves (a) and ROC curves (b) estimated from a sample of 520 image patches.

EROC curves, which indicate better observer performance for acquisition B .

Since the results for this example are based on a relatively small sample of only 156 image patches, the confidence interval for the difference in EROC areas is fairly large, i.e., the precision of the estimated difference is poor. However, because the six scans used in this example were actually a subset of 40 repeated scans of the subject, we were able to estimate the difference in EROC areas with greater precision using all 40 scans. Namely, using the central sagittal slices of the 40 image volumes, which resulted in a total of $13 \times 40 \times 2 = 520$ image patches, a 95% confidence interval for $\mathcal{A}_E^{(B)} - \mathcal{A}_E^{(A)}$ was observed to be $[0.080, 0.114]$; see Fig. 5 for plots of the empirical EROC and ROC curves. These results are consistent with the conclusions drawn above with the smaller sample.

6 Discussion and Conclusions

EROC curves are a generalization of ROC curves designed to summarize observer performance for combined detection and estimation tasks. Building on parallels with classical ROC analysis, we introduced practical, nonparametric estimators for the EROC curve, the area under the EROC curve, and the covariance matrix for a vector of area estimates. In addition, we gave theoretical results showing the asymptotic consistency of the covariance estimator, and justifying the construction of confidence intervals based on asymptotic normality for EROC area.

The effectiveness of the confidence intervals for the area under the EROC curve was further supported with a Monte Carlo simulation study. For this study, we used a simulation model that included correlations both across imaging scenarios and between signal-present ratings and parameter estimates. Moreover, the simulation model was implemented for conditions including normal and log-normal distributions for the ratings and parameter estimates, with both deterministic and random lesion parameters. For several typical experimental sample sizes and a large number of simulation model parameters, the coverage probability of the confidence intervals was found to be reliable. A more comprehensive simulation study of the effectiveness of the confidence intervals is beyond the scope of the present work and is a worthy topic for future investigation; the simulation model given here is a potential starting point.

The application of our estimators was illustrated with an example comparing MRI images reconstructed from different k -space sampling trajectories. This example showed that EROC curves can be used to summarize differences in observer performance that are not evident from ROC curves alone, and that EROC curves can be utilized to measure combined

detection and estimation performance in a meaningful manner. Namely, we found that the results of EROC analysis were consistent with the fact that one k -space trajectory resulted in greater image blur. Also, the example demonstrated how the tools presented here for EROC analysis can be applied for rigorous statistical inference in image quality evaluations. As mentioned in Sec. 5, a potential limitation of the example stems from the assumption of statistical independence between extracted image patches. Therefore, when applying such an approach in practice, the independence assumption should be justified appropriately.

The nonparametric covariance estimator for a vector of estimated EROC areas can be viewed as a generalization of the method of DeLong et al.⁹ for ROC analysis. Like DeLong's method, this estimator treats readers as fixed, and only assesses variability due to the case set. Consequently, it is best suited to image quality evaluations involving either a fixed model observer (such as in our example) or early stage evaluations involving a small pool of readers. For studies that aim to draw inferences for a large population of readers, it is preferable to model readers as a random effect. In this case, the covariance estimator given here can be combined with the method of Obuchowski, Rockette, and Hillis,^{19–22} which requires estimates of fixed-reader covariances. Alternatively, it is straightforward to adapt the U -statistics method of Gallas et al.^{23–25} Validation of these possible approaches for random-reader EROC analysis is an important topic for future work.

For readers interested in applying EROC analysis methodology, MATLAB® code implementing the analysis estimators described in this work is publicly available.²⁶

Acknowledgments

We would like to thank the Professor Karel Deblaere (Department of Radiology and Nuclear Medicine, Ghent University Hospital, Ghent, Belgium) for providing the MRI data sets that were used in the example. Bart Goossens acknowledges support from a postdoctoral fellowship of the Research Foundation—Flanders (FWO, Belgium).

References

1. H. H. Barrett and K. J. Myers, *Foundations of Image Science*, John Wiley & Sons, Hoboken, NJ (2004).
2. R. F. Wagner, C. E. Metz, and G. Campbell, "Assessment of medical imaging systems and computer aids: a tutorial review," *Acad. Radiol.* **14**(6), 723–748 (2007).
3. S. J. Starr et al., "Visual detection and localization of radiographic images," *Radiology* **116**(3), 533–538 (1975).
4. P. Bunch et al., "A Free-response approach to the measurement and characterization of radiographic observer performance," *J. Appl. Photogr. Eng.* **4**(4), 166–171 (1978).
5. E. Clarkson, "Estimation receiver operating characteristic curve and ideal observers for combined detection/estimation tasks," *J. Opt. Soc. Am. A* **24**(12), B91–B98 (2007).
6. B. Goossens et al., "Objectively measuring signal detectability, contrast, blur and noise in medical images using channelized joint observers," *Proc. SPIE* **8673**, 86730J (2013).
7. L. Zhang et al., "Channelized model observer for the detection and estimation of signals with unknown amplitude, orientation, and size," *J. Opt. Soc. Am. A* **30**(11), 2422–2432 (2013).
8. D. Bamber, "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *J. Math. Psychol.* **12**(4), 387–415 (1975).
9. E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics* **44**(3), 837–845 (1988).
10. A. Wunderlich and F. Noo, "A nonparametric procedure for comparing the areas under correlated LROC curves," *IEEE Trans. Med. Imaging* **31**(11), 2050–2061 (2012).
11. M. S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, Oxford, UK (2003).
12. D. A. S. Fraser, *Nonparametric Methods in Statistics*, p. 142, John Wiley & Sons, New York (1957).
13. A. E. Burgess, "Comparison of receiver operating characteristic and forced choice observer performance measurement methods," *Med. Phys.* **22**(5), 643–655 (1995).
14. P. K. Sen, "On some convergence properties of U -statistics," *Calcutta Stat. Assoc.* **10**(37), 1–18 (1960).
15. P. K. Sen, *Sequential Nonparametrics*, p. 82, John Wiley & Sons, New York (1981).
16. R. J. Serfling, *Approximation Theorems of Mathematical Statistics*, p. 48, John Wiley & Sons, New York (1980).
17. J. S. Milton and J. C. Arnold, *Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences*, 3rd ed., McGraw-Hill, Boston, Massachusetts (1995).
18. J. Aelterman et al., "Augmented Lagrangian based reconstruction of non-uniformly sub-Nyquist sampled MRI data," *Signal Process.* **91**(12), 2731–2742 (2011).
19. N. A. Obuchowski and H. E. Rockette, "Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations," *Commun. Stat. Simulat.* **24**(2), 285–308 (1995).
20. S. L. Hillis et al., "A comparison of the Dorfman–Berbaum–Metz and Obuchowski–Rockette methods for receiver operating characteristic (ROC) data," *Stat. Med.* **24**(10), 1579–1607 (2005).
21. S. L. Hillis, "A comparison of denominator degrees of freedom methods for multiple observer ROC analysis," *Stat. Med.* **26**(6), 596–619 (2007).
22. S. L. Hillis, "A marginal-mean ANOVA approach for analyzing multi-reader multicase radiological imaging data," *Stat. Med.* **33**(2), 330–360 (2014).
23. B. D. Gallas, "One-shot estimate of MRMC variance: AUC," *Acad. Radiol.* **13**(3), 353–362 (2006).
24. B. D. Gallas et al., "A framework for random-effects ROC analysis: biases with the bootstrap and other variance estimators," *Commun. Stat. A* **38**(15), 2586–2603 (2009).
25. W. Chen, B. D. Gallas, and W. A. Yousef, "Classifier variability: accounting for training and testing," *Pattern Recognit.* **45**(7), 2661–2671 (2012).
26. A. Wunderlich, "IQmodelo: statistical software for image quality assessment with model observers," <http://code.google.com/p/iqmodelo/> (2014).

Adam Wunderlich received a PhD degree in electrical and computer engineering from the University of Utah (Salt Lake City, Utah) in 2009. Subsequently, from 2009 to 2012, he was a postdoctoral fellow in the Department of Radiology at the University of Utah. Since 2012, Wunderlich has been with the U.S. Food and Drug Administration (Silver Spring, Maryland). His research interests include task-based image quality assessment, statistical methodology, and CT imaging.

Bart Goossens earned his master's degree in computer science engineering from Ghent University, Belgium, in 2006 and a PhD degree from the same university in 2010. Since October 2010, he has been a postdoctoral fellow of FWO Flanders and since October 2013, he is also a part-time lecturer at Ghent University. For his research, he received the Barco/FWO prize for master's theses in 2006 and the scientific prize IBM Belgium for Informatics in 2011.