Published in final edited form as:

Acad Radiol. 2013 July ; 20(7): 908–914. doi:10.1016/j.acra.2013.04.002.

Validation of Monte Carlo estimates of three-class ideal observer operating points for normal data

Darrin C. Edwards

The Department of Physiology, Northwestern University, Chicago, IL, USA

Abstract

Rationale and Objectives—Traditional two-class receiver operating characteristic (ROC) analysis is inadequate for the complete evaluation of observer performance in tasks with more than two classes.

Materials and Method—Here a Monte Carlo estimation method for operating point coordinates on a three-class ROC surface is developed and compared with analytically calculated coordinates in two special cases: 1) univariate and 2) restricted bivariate trinormal underlying data.

Results—In both cases the statistical estimates were found to be good in the sense that the analytical values lay within the 95% confidence interval of the estimated values about 95% of the time.

Conclusions—The statistical estimation method should be key in the development of a pragmatic performance metric for evaluation of observers in classification tasks with three or more classes.

Keywords

ROC analysis; three-class classification; ideal observer decision rules

1. Introduction

Receiver operating characteristic (ROC) analysis has, for many years, been the standard for evaluating observer performance in a medical decision task with two classes to which observations belong [1]. A particularly familiar example is the canonical radiological task of identifying whether an abnormality, such as a fracture or lesion, is present in an image or not.

Not all medical, or even radiological, tasks are so readily restricted to two outcomes, however. A particular task might require distinguishing among multiple types of abnormality [2]; distinguishing normal tissue from abnormalities of different types [3]; or, in the computer-aided diagnosis (CAD) task which originally motivated much of the work in this area, one might need to distinguish malignant and benign actual lesions from the false-positive detections produced by an automated scheme [4, 5, 6]. Traditional two-class ROC analysis is inadequate for the complete evaluation of observer performance in such tasks.

© 2013 The Association of University Radiologists. Published by Elsevier Inc. All rights reserved.

darrin.c.edwards@gmail.com (Darrin C. Edwards) .

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Unfortunately, although the broader theoretical characteristics of observer behavior in a three-class classification task were outlined many decades ago [7], the extension of this knowledge to a complete understanding, and the implementation of such knowledge into practical tools for addressing real clinical problems, have remained elusive.

Why should this be the case? After all, two-class ROC analysis is successful not just because of its practical usefulness, but also because of the theoretical simplicity and elegance it possesses. One measures some quality of the objects being classified, and compares this decision variable with a critical threshold. If the decision variable is above the critical threshold, the object is classified as positive; the fraction of actually positive objects called positive is the true-positive fraction (TPF), while the fraction of actually negative objects (incorrectly) called positive is the false-positive fraction (FPF). The FPF and TPF values give the observer's ROC operating point at that critical threshold; if a different critical threshold is chosen, a different operating point is obtained, and the collection of all (FPF, TPF) pairs for all values of the critical threshold is the ROC curve [7, 8, 1]. The observer which achieves the best possible ROC performance, given the variability of the decision variables across the actually positive and negative populations, is called the ideal observer [7, 8]. Objectively comparing two different observers based on their ROC curves is simple if one curve is always above the other, but problematic if the two curves cross; however, the area under the ROC curve (AUC) is a readily calculated quantity that allows such comparison, and furthermore is theoretically justifiable since the AUC can be shown to be the chance of the observer making a correct decision in a two-alternative forced-choice task (that is, when presented with a pair of observations, one actually positive and one actually negative, the chance of correctly deciding which is which) [8]. Given a simple model for the underlying data, the binormal model, ROC curves can be computed under both assumptions in which the observer's decision variable is directly related to the underlying data (the conventional binormal model [9]), and in which the decision variable is an ideal observer decision function of the underlying data (the proper binormal model [10]).

Turning to the three-class task, difficulties appear at the outset and accumulate at a discouraging rate. The performance of an observer is characterized by six out of the nine possible conditional classification probabilities given a particular decision rule (the probabilities of deciding an observation belongs to one class when it is actually drawn from that or another class, analogous to TPF and FPF), not three as one might naively expect [11, 12, 13, 14, 15]. (The nine classification probabilities are related by three equations, one for each class, because they are conditional probabilities [16], just as $TPF + FNF = 1$ where FNF is the false negative fraction; thus three of the classification probabilities may be eliminated, leaving six.) A number of researchers have proposed applying restrictions to the form of the observer, to reduce the dimensionality of the performance description from six down to three in a more principled fashion [17, 18, 19, 20, 21]; however, despite the resulting observer behavior being consistent with that of the ideal observer given the restrictions imposed [22, 23], these descriptions of observers and their performance are nevertheless not fully general. In the general case, six conditional classification probabilities (the most principled choice of which is the set of six misclassification probabilities) form the coordinates of a six-dimensional ROC space, and the observer's ROC "surface" has as many as five degrees of freedom. For the ideal observer, these coordinates are the probabilities of a random decision variable vector lying in one of a set of three wedge-shaped regions [7, 24]; this is analogous to the two-class case, but the integrals involved in their calculation are far less tractable. (As for the behavior of non-ideal observers, it is not even completely clear what the general form of the decision function should be.) To evaluate the performance of such an observer, one might naively hope to calculate the volume under this (five-dimensional) ROC surface, and use that for comparison with other observers in a manner analogous to the two-class AUC. Unfortunately, this quantity was shown not to correlate in

any useful way with intuitive concepts of “performance” due to degeneracy issues in ideal observer ROC surfaces [25].

This result implies that a performance metric is required that is not just a simple generalization of AUC. Nevertheless, since the six misclassification probabilities that form the coordinates of the observer’s ROC surface remain the fundamental descriptors of the observer’s performance, it is to be expected that a valid performance metric will still be derived in some more sophisticated fashion from the ROC surface. Conceptual characterization of such a performance metric is not impossible [26], but regardless of the theoretical details, it is clear that any such performance metric will require a detailed description of the observer’s ROC surface. Lacking fully general analytic methods for calculating three-class ideal observer operating points, we can safely assume that numerical or statistical methods will be required—essentially Monte Carlo methods, since the operating point coordinates are conditional probabilities, i. e., integrals of particular probability density functions (PDFs) over particular regions. Of course, such a statistical operating point estimation method would need to be validated in practice; this requires, in turn, an analytic method for calculating the “true” value of the operating point.

Our most recently published work finally gave Charles Metz and me hope of breaking this vicious circle; in it, we fully and analytically characterized the ROC operating point behavior of a three-class ideal observer acting on *univariate* normal underlying data [27]. However, unlike the two-class task for which the sufficient statistic used by the ideal observer is always univariate, the three-class ideal observer makes use of a pair of decision variables. To consider the fully general case, therefore, it is necessary to be able to take into account *bivariate* data. That is, while it is perhaps imaginable that a particular medical imaging task could be found such that the highly multivariate image data could be classified by an ideal observer acting on univariate decision variable data, this is highly unlikely to be true in general. In any case, a fully general approach would still be needed in order to establish that the univariate method was adequate and valid for that particular task.

A brief review of the univariate work is given in Sec. 2, along with a more detailed description of a second special case in which three-class ideal observer operating points may be calculated analytically, in the case of bivariate normal data under certain restrictions on both the data distributions and the form of the decision rule. The basis of the statistical method for estimating ideal observer operating points is also described. In Sec. 3, a set of simulation studies is described in which analytically calculated and statistically estimated operating points are compared for a variety of underlying data distributions and particular choices of ideal observer decision criteria. The results of these simulations studies are presented in Sec. 4, and their implications are discussed in Sec. 5. The conclusions that can be drawn from them are finally given in Sec. 6. The goal of the present work is to show that the general validity of the statistical estimation method can be adopted as a working hypothesis, and that this validity can continue to be evaluated as new analytic methods (beyond the two considered here) become available.

2. Theory

In a two-class classification task, with underlying data x , the ideal observer makes decisions by comparing a function of the data, called the likelihood ratio (LR), with a critical threshold γ . The LR is the ratio of the PDFs of the underlying data, conditional on the classes from which the data are drawn; that is, $LR(\vec{x}) \equiv p(\vec{x}|\text{class 1})/p(\vec{x}|\text{class 2})$. If $LR(x) > \gamma$, the observation is called positive, otherwise it is called negative. (A bold typeface is used to denote random variables.) The point $LR = \gamma$ is the boundary between the region where observations are called positive and the region where observations are called negative.

In the three-class case, there are three distributions from which the data are drawn. Thus, there are two likelihood ratio functions of the underlying data:

$$LR_1(\vec{x}) \equiv \frac{p(\vec{x}|\text{class 1})}{p(\vec{x}|\text{class 3})} \quad (1)$$

$$LR_2(\vec{x}) \equiv \frac{p(\vec{x}|\text{class 2})}{p(\vec{x}|\text{class 3})}; \quad (2)$$

and instead of a simple point boundary between regions where the various classifications are made, we have a set of three intersecting decision boundary line segments

$$\gamma_{121}LR_1 - \gamma_{121}LR_2 = \gamma_{313} - \gamma_{323} \quad (3)$$

$$\gamma_{131}LR_1 + (\gamma_{232} - \gamma_{212})LR_2 = \gamma_{313} \quad (4)$$

$$(\gamma_{131} - \gamma_{121})LR_1 + \gamma_{232}LR_2 = \gamma_{323}, \quad (5)$$

which we call, respectively, the “1-vs.-2” boundary, the “1-vs.-3” boundary, and the “2-vs.-3” boundary [28]. The γ_{ij} are the decision criteria used by the ideal observer, only five of which are independent (the three equations above can be multiplied by any number, but the lines described will be left unchanged). For brevity this collection of values can be written as a vector $\vec{\gamma}$. The conditional probability of the observer assigning an observation to class i , given that it is actually drawn from class j , is the integral of the PDF of the likelihood ratios (conditional on membership in class j) over the region where class i is decided; this misclassification probability is denoted by P_{ij} , and forms one of the six coordinates of the ROC space [25].

Consider the special case in which the underlying data $\vec{x}$ are univariate (as was the assumption for the two-class case); that is,

$$p(x|\text{class 1}) = N(x; \mu_1, \sigma_1^2), \quad (6)$$

$$p(x|\text{class 2}) = N(x; \mu_2, \sigma_2^2), \quad (7)$$

$$p(x|\text{class 3}) = N(x; 0, 1), \quad (8)$$

where $N(x; \mu, \sigma^2)$ is a normal PDF of x with mean μ and variance σ^2 . This means LR_1 and LR_2 are each functions of the single variable x , and so we can express LR_2 as a relation of LR_1 . Using the properties of normal functions, these relations can be “cataloged”, and the results used to analytically calculate operating points of the ideal observer given particular values of the decision criteria $\vec{\gamma}$ [27]. An example of such a likelihood ratio curve for a particular set of data parameters μ and σ^2 , and decision criteria $\vec{\gamma}$, is shown in Fig 1.

Now consider the case of bivariate underlying data, that is, $\vec{x} = (x, y)$. In the unrestricted case, limited insights can be gained at the price of much algebraic drudgery [29]. A more elegant insight can be obtained from applying a restriction inspired by a well-known result

in the two-class case: if the data distributions have equal variance, then the logarithm of the likelihood ratio is linear in the underlying data [30]. That is, the underlying data are now drawn from three bivariate distributions:

$$p(\vec{x}|\text{class 1}) = N(\vec{x}; \vec{\mu}_1, I), \quad (9)$$

$$p(\vec{x}|\text{class 2}) = N(\vec{x}; \vec{\mu}_2, I), \quad (10)$$

$$p(\vec{x}|\text{class 3}) = N(\vec{x}; \vec{0}, I), \quad (11)$$

where I is the 2×2 identity matrix. It can be shown, generalizing the approach of Barrett et al. for the two-class case [30], that

$$\lambda_1(\vec{x}) = \vec{x} \cdot \vec{\mu}_1 - \frac{|\vec{\mu}_1|^2}{2} \quad (12)$$

$$\lambda_2(\vec{x}) = \vec{x} \cdot \vec{\mu}_2 - \frac{|\vec{\mu}_2|^2}{2} \quad (13)$$

where $\lambda_1(\vec{x}) \equiv \log \text{LR}_1(\vec{x})$. From this it follows that the random variables λ_1 and λ_2 also follow bivariate normal distributions [16] for each of the three classes.

One can then apply the so-called “equal error utility” restriction made by He et al. [19] on the form of the ideal observer’s decision rule, which results in a Scurfield-type decision rule [17] applied to log-likelihood ratio decision variables. With these two restrictions, it can readily be shown that the observer’s operating points are just values of the bivariate normal cumulative distribution function (CDF), which are straightforward to obtain. For example, for a given pair of decision criteria (a, b) , such that class 3 is decided if $\text{LR}_1 > a$ and $\text{LR}_2 > b$, or equivalently $\lambda_1 > \log a$ and $\lambda_2 > \log b$, it can be shown that $P_{31} = F(\vec{v}; \mu_{\lambda_1}, \Sigma_{\lambda})$ and $P_{32} = F(\vec{v}; \mu_{\lambda_2}, \Sigma_{\lambda})$, where $\vec{v}$ is the vector $(\log a, \log b)$, μ_{λ_i} is the mean of (λ_1, λ_2) for class i , Σ_{λ} is their covariance matrix (the same for all three classes), and $F(\cdot)$ is the (bivariate) cumulative normal distribution function. (The other four misclassification probabilities may also be expressed in terms of $F(\cdot)$ through transformations of variables [16] that are fairly straightforward but of little direct interest here.)

The final question to be addressed is the statistical estimation of the ideal observer’s operating point coordinates. Suppose we generate a set of n observations $\vec{x}$ from a particular normal distribution, say the one for class 1 (with parameters $\vec{\mu}_1$ and Σ_1). We can calculate $\text{LR}_1(\vec{x})$ and $\text{LR}_2(\vec{x})$ for each of these observations. Next, given a particular choice of decision criteria $\vec{\gamma}$, we count the number of observations such that $(\text{LR}_1, \text{LR}_2)$ falls in the region where observations are classified as belonging to class 2; call this number k_{21} (the number “called” class 2 that are “actually” class 1). Assuming the observations $\vec{x}$ are generated independently, the random variables k_{21} and k_{31} will follow a multinomial distribution (with $k_{11} \equiv n - k_{21} - k_{31}$). The variable k_{21}/n is then an unbiased estimate of the probability of an observation, which actually belongs to class 1, being assigned to class 2

(the coordinate P_{21} of the ROC operating point for $\vec{\gamma}$); and $\frac{k_{21}(n - k_{21})}{(n - 1)n^2}$ is an unbiased estimate of the variance of k_{21}/n [16]. The 95% confidence interval for this operating point

coordinate can be shown to be well approximated by $k_{21} \pm 2 \sqrt{\frac{k_{21}(n - k_{21})}{(n - 1)n^2}}$ [31]. Generalizing to the other coordinates, it is clear that estimates of all six operating points, and 95% confidence intervals on those estimates, are readily obtained.

3. Materials and Method

As a preliminary test of the validity of the concepts outlined in the preceding section, a single value of the decision criteria $\vec{\gamma}$, and of the parameters μ_1, σ_1^2, μ_2 , and σ_2^2 of a univariate data distribution, were randomly selected. A set of ten simulation studies was then run, in each of which the statistically estimated operating point was compared with the analytically calculated operating point. The number of samples n for statistically estimating each coordinate of the ROC operating point was varied across the simulation studies, from $n = 10^6$ to $n = 10^8$. The 95% confidence interval for the statistical estimate of each operating

point coordinate was calculated using the expression $k_{ij} \pm 2 \sqrt{\frac{k_{ij}(n - k_{ij})}{(n - 1)n^2}}$ as described in the preceding section. The sample size n of the statistical estimation method can be considered adequate if the “true” (analytically calculated) value of the operating point coordinate lies inside the confidence interval (at least 95% of the time).

Based on this preliminary work, a value of $n = 10^7$ was chosen for the remaining simulation studies. For the univariate simulation studies, ten values of the decision criteria $\vec{\gamma}$ were randomly selected; in addition, ten sets of parameters of the underlying data distribution $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ were randomly generated (for brevity, this collection of data parameters is referred to as $\vec{\beta}$). For each of the 100 combinations of decision criteria and underlying data distributional parameters, the observer operating point was analytically calculated, and this value was compared with the statistical estimate of the operating point and the 95% confidence interval of that estimate.

For the bivariate simulation studies, the restriction placed on the form of the decision rule (described in Sec. 2) means that the slopes of the decision boundaries are completely determined by their intersection point (a, b) ; this (a, b) pair was thus extracted from each of the ten decision criteria $\vec{\gamma}$ from the univariate simulation studies, and used to define the decision rules for the bivariate simulation studies. The means of the bivariate underlying observational data were taken to be $\vec{\mu}_1 = (\mu_1, 0)$ and $\vec{\mu}_2 = (0, \mu_2)$, with μ_1 and μ_2 being the means from the univariate simulation studies; $\vec{\mu}_3$ is $(0, 0)$ by analogy with Eq. 8. Since the assumption that the distributions have equal variance was also imposed, the covariance matrices Σ_1, Σ_2 , and Σ_3 were all 2×2 identity matrices. As in the univariate simulation studies, for each of the 100 combinations of decision criteria and underlying data distributional parameters, the observer operating point was analytically calculated, and this value was compared with the statistical estimate of the operating point and the 95% confidence interval of that estimate.

4. Results

Figures 2 through 7 show, respectively, the statistically estimated operating point coordinates P_{12} , P_{13} , P_{21} , P_{23} , P_{31} , and P_{32} , for a randomly selected decision criterion $\vec{\gamma}$ and randomly selected univariate normal underlying data parameters as described in Sec. 3.

Table 1 lists, for the univariate underlying data, the fraction of all conditions (decision criteria $\vec{\gamma}$ and data parameters $\vec{\beta}$) for which the analytically calculated operation point coordinate lay within the 95% confidence interval of the estimated operating point. Table 2 lists the equivalently defined fraction for the bivariate underlying data.

5. Discussion

The fractions given in the tables in the preceding section show that the 95% confidence intervals of the statistical estimates of the three-class ROC operating points contain the analytically calculated operating points (the “true” values) 95% of the time. For the two special cases considered, then, the statistical estimation method can be considered a reliable means of obtaining accurate measures of the ROC operating points of the three-class observers modeled by those special cases.

Of course, the statistical estimation method makes no assumptions about the form of the decision strategy used by the three-class observer being evaluated. The only assumption made is that the observations being classified be sampled independently of one another, i. e., that they follow a multinomial distribution. Provided this is the case, the set of probabilities governing that multinomial distribution, which are just the coordinates of an ROC operating point, should be just as readily estimated regardless of the underlying data model (univariate trinormal; bivariate trinormal under an “equal error utility” observer restriction; or a fully general three-class ideal observer model).

The objection can immediately be raised that such a claim has not been proven outside the two special cases considered here. However, given the success of the method in those two particular cases, combined with the obvious generality of the principles underlying this method, it seems eminently reasonable to adopt the validity of the statistical estimation method as a working hypothesis; this validity can continue to be evaluated as new analytic methods become available.

In the meantime, the statistical estimation method has the obvious benefit that accurate measurements of ROC operating points for arbitrary underlying data models and decision strategies can be made, subject only to limitations of time and computational resources. During the many years that Charles Metz and I worked on the three-class classification problem, we had fairly early on resigned ourselves to the fact that analytical methods for observer performance evaluation, analogous to the extraordinarily successful two-class methods that had been developed [9, 10], were highly unlikely to be achievable. What proved more frustrating was that even determining what could serve as a suitable performance metric should present so many unexpected difficulties. Although we had begun to work out theoretical descriptions of what such a performance metric might entail [26, 32], those descriptions demanded detailed knowledge of the ROC surface for their explicit evaluation.

The statistical estimation method enables one to obtain such detailed descriptions of an ROC surface without the need to deal with analytically intractable expressions for the operating points on such a surface. It is hoped that this methodology will prove crucial in developing a

practical and general ROC performance metric that would finally allow the evaluation of observer performance in classification tasks with more than two classes.

6. Conclusion

A Monte-Carlo-based statistical estimation method for operating point coordinates on a three-class ROC surface was compared with analytically calculated coordinates in two special cases: that of univariate trinormal underlying data, and that of bivariate trinormal underlying data with the ideal observer restricted by the so-called “equal error utility” assumption. In both cases the statistical estimates were found to be good in the sense that the analytical values lay within the 95% confidence interval of the estimated values about 95% of the time. Provided this validity can be shown to hold for more general data and observer models, the statistical estimation method should be key in the development of a pragmatic performance metric for evaluation of observers in classification tasks with three or more classes.

Acknowledgments

This work was supported by grant R01 MH090937-02 from the National Institutes of Health (Nancy J. Cox, principal investigator). I thank Dr. Maryellen Giger and Chun-Wai Chan for generously continuing to provide me with access to the SIRAF computational cluster after funding for my position at the University of Chicago was no longer available.

I wish to express my gratitude and appreciation for the life and work of Charles E. Metz. His encouragement of my efforts and enthusiasm for our collaboration were truly inspirational. That Dr. Metz chose to spend the last years of his career and of his life working with me in this difficult but fascinating area of investigation is a source to me of great pride and of hope that I will at some point be permitted the resources to continue in this work.

References

- [1]. Metz CE. Basic principles of ROC analysis. *Seminars in Nuclear Medicine* VIII. 1978;283–298.
- [2]. Ashizawa K, MacMahon H, Ishida T, Nakamura K, Vyborny CJ, Katsuragawa S, Doi K. Effect of an artificial neural network on radiologists' performance in the differential diagnosis of interstitial lung disease using chest radiographs. *AJR*. 1999; 172:1311–1315. [PubMed: 10227508]
- [3]. He X, Song X, Frey EC. Application of three-class ROC analysis to task-based image quality assessment of simultaneous dual-isotope myocardial perfusion SPECT (MPS). *IEEE Trans. Med. Imag.* 2008; 27:1556–1567.
- [4]. Chan, H-P.; Sahiner, B.; Hadjiiski, LM.; Petrick, N.; Zhou, C. Design of three-class classifiers in computer-aided diagnosis: Monte Carlo simulation study. In: Milan, Sonka; Fitzpatrick, J. Michael, editors. *Proc. SPIE Vol. 5032 Medical Imaging 2003: Image Processing*. SPIE; Bellingham, WA: p. 567-578.
- [5]. Edwards DC, Lan L, Metz CE, Giger ML, Nishikawa RM. Estimating three-class ideal observer decision variables for computerized detection and classification of mammographic mass lesions. *Med. Phys.* 2004; 31:81–90. [PubMed: 14761024]
- [6]. Drukker K, Edwards DC, Giger ML, Metz CE. Computerized 3-way classification of lesions using breast ultrasound and mammography. *IWDM 2004: International Workshop on Digital Mammography*. :540–545.
- [7]. Van Trees, HL. *Detection, Estimation and Modulation Theory: Part I*. John Wiley & Sons; New York: 1968.
- [8]. Egan, JP. *Signal Detection Theory and ROC Analysis*. Academic Press; New York: 1975.
- [9]. Metz CE, Herman BA, Shen J-H. Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statist. Med.* 1998; 17:1033–1053.

- [10]. Metz CE, Pan X. 'Proper' binormal ROC curves: Theory and maximum-likelihood estimation. *J. Math. Psychol.* 1999; 43:1–33. [PubMed: 10069933]
- [11]. Mossman D. Three-way ROCs. *Med. Decis. Making.* 1999; 19:78–89. [PubMed: 9917023]
- [12]. Dreiseitl S, Ohno-Machado L, Binder M. Comparing three-class diagnostic tests by three-way ROC analysis. *Med. Decis. Making.* 2000; 20:323–331. [PubMed: 10929855]
- [13]. Nakas CT, Yiannoutsos CT. Ordered multiple-class ROC analysis with continuous measurements. *Statist. Med.* 2004; 23:3437–3449.
- [14]. Li J, Fine JP. ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. *Biostatistics.* 2008; 9:566–576. [PubMed: 18304996]
- [15]. Inácio V, Turkman AA, Alonzo TA. Nonparametric Bayesian estimation of the three-way receiver operating characteristic surface. *Biometrical Journal.* 2011; 53:1011–1024. [PubMed: 22069202]
- [16]. Papoulis, A. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, Inc.; New York: 1991.
- [17]. Scurfield BK. Generalization of the theory of signal detectability to n -event m -dimensional forced-choice tasks. *J. Math. Psychol.* 1998; 42:5–31. [PubMed: 9606159]
- [18]. He X, Metz CE, Tsui BMW, Links JM, Frey EC. Three-class ROC analysis — A decision theoretic approach under the ideal observer framework. *IEEE Trans. Med. Imag.* 2006; 25:571–581.
- [19]. He X, Frey EC. Three-class ROC analysis — The equal error utility assumption and the optimality of three-class ROC surface using the ideal observer. *IEEE Trans. Med. Imag.* 2006; 25:979–986.
- [20]. Sahiner B, Chan H-P, Hadjiiski LM. Performance analysis of three-class classifiers: Properties of a 3-D ROC surface and the normalized volume under the surface for the ideal observer. *IEEE Trans. Med. Imag.* 2008; 27:215–227.
- [21]. He X, Gallas BD, Frey EC. Three-class ROC analysis — toward a general decision theoretic solution. *IEEE Trans. Med. Imag.* 2010; 29:206–215. PMID: PMC2821068.
- [22]. Edwards DC, Metz CE. Analysis of proposed three-class classification decision rules in terms of the ideal observer decision rule. *J. Math. Psychol.* 2006; 50:478–487.
- [23]. Edwards DC, Metz CE. Optimization of restricted ROC surfaces in three-class classification tasks. *IEEE Trans. Med. Imag.* 2007; 26:1345–1356.
- [24]. Edwards DC, Metz CE. Restrictions on the three-class ideal observer's decision boundary lines. *IEEE Trans. Med. Imag.* 2005; 24:1566–1573.
- [25]. Edwards DC, Metz CE, Nishikawa RM. The hypervolume under the ROC hypersurface of 'near-guessing' and 'near-perfect' observers in N -class classification tasks. *IEEE Trans. Med. Imag.* 2005; 24:293–299.
- [26]. Edwards, DC.; Metz, CE. A utility-based performance metric for ROC analysis of N -class classification tasks. In: Yulei, Jiang; Berkman, Sahiner, editors. *Proc. SPIE Vol. 6515 Medical Imaging 2007: Image Perception, Observer Performance, and Technology Assessment.* SPIE; Bellingham, WA: p. 6515031-65150310.
- [27]. Edwards DC, Metz CE. The three-class ideal observer for univariate normal data: Decision variable and ROC surface properties. *J. Math. Psychol.* 2012; 56:256–273. [PubMed: 23162165]
- [28]. Edwards DC, Metz CE, Kupinski MA. Ideal observers and optimal ROC hypersurfaces in N -class classification. *IEEE Trans. Med. Imag.* 2004; 23:891–895.
- [29]. Edwards, DC. Support of the decision variable densities of the three-class ideal observer for bivariate trinormal data. In: Manning, David J.; Abbey, Craig K., editors. *Proc. SPIE Vol. 7966 Medical Imaging 2011: Image Perception, Observer Performance, and Technology Assessment.* SPIE; Bellingham, WA: p. 7966011-79660110.
- [30]. Barrett HH, Abbey CK, Clarkson E. Objective assessment of image quality. III. ROC metrics, ideal observers, and likelihood-generating functions. *J. Opt. Soc. Am. A.* 1998; 15:1520–1535.
- [31]. Metz CE. Quantification of failure to demonstrate statistical significance: The usefulness of confidence intervals. *Invest. Radiol.* 1993; 28:59–63. [PubMed: 8425854]

- [32]. Edwards, DC.; Metz, CE. Optimality of a utility-based performance metric for ROC analysis, in: Berkman Sahiner. In: Manning, David J., editor. Proc. SPIE Vol. 6917 Medical Imaging 2008: Image Perception, Observer Performance, and Technology Assessment. SPIE; Bellingham, WA: p. 69170F1-68170F6.

Figure 1.

The likelihood ratio curve when the underlying data are drawn from normal distributions with parameters $\mu_1 = 1/8$, $\sigma_1^2 = 1/8$, $\mu_2 = -2/3$, $\sigma_2^2 = 1/2$, $\mu_3 = 0$, and $\sigma_3^2 = 1$; and where the decision boundary line segments are shown with dashed lines. An expression such as " π_1 " indicates a region where a particular decision is made.

Figure 2.

Statistically estimated values of the ROC operating point coordinate P_{12} (the probability of classifying an observation as class 1 when it is actually class 2), with error bars showing the 95% confidence intervals on the estimates, plotted versus the number of samples used to obtain the estimates. The horizontal line is the analytically calculated operating point coordinate.

Figure 3.

Statistically estimated values of the ROC operating point coordinate P_{13} (the probability of classifying an observation as class 1 when it is actually class 3), with error bars showing the 95% confidence intervals on the estimates, plotted versus the number of samples used to obtain the estimates. The horizontal line is the analytically calculated operating point coordinate.

Figure 4.

Statistically estimated values of the ROC operating point coordinate P_{21} (the probability of classifying an observation as class 2 when it is actually class 1), with error bars showing the 95% confidence intervals on the estimates, plotted versus the number of samples used to obtain the estimates. The horizontal line is the analytically calculated operating point coordinate.

Figure 5.

Statistically estimated values of the ROC operating point coordinate P_{23} (the probability of classifying an observation as class 2 when it is actually class 3), with error bars showing the 95% confidence intervals on the estimates, plotted versus the number of samples used to obtain the estimates. The horizontal line is the analytically calculated operating point coordinate.

Figure 6.

Statistically estimated values of the ROC operating point coordinate P_{31} (the probability of classifying an observation as class 3 when it is actually class 1), with error bars showing the 95% confidence intervals on the estimates, plotted versus the number of samples used to obtain the estimates. The horizontal line is the analytically calculated operating point coordinate.

Figure 7.

Statistically estimated values of the ROC operating point coordinate P_{32} (the probability of classifying an observation as class 3 when it is actually class 2), with error bars showing the 95% confidence intervals on the estimates, plotted versus the number of samples used to obtain the estimates. The horizontal line is the analytically calculated operating point coordinate.

Table 1**Univariate data**

The fraction of all conditions (decision criteria $\vec{\gamma}$ and univariate data parameters $\vec{\beta}$) for which the analytically calculated operation point coordinate lay within the 95% confidence interval of the estimated operating point.

$f_{CI}(P_{12})$	$f_{CI}(P_{13})$	$f_{CI}(P_{21})$	$f_{CI}(P_{23})$	$f_{CI}(P_{31})$	$f_{CI}(P_{32})$
0.96	0.97	0.92	0.99	0.89	0.91

Table 2**Bivariate data**

The fraction of all conditions (decision criteria $\vec{\gamma}$ and bivariate data parameters $\vec{\beta}$) for which the analytically calculated operation point coordinate lay within the 95% confidence interval of the estimated operating point.

$f_{CI}(P_{12})$	$f_{CI}(P_{13})$	$f_{CI}(P_{21})$	$f_{CI}(P_{23})$	$f_{CI}(P_{31})$	$f_{CI}(P_{32})$
0.94	0.96	0.94	0.93	0.93	0.9