Published in final edited form as:
IEEE Trans Med Imaging. 2008 May ; 27(5): 577–588. doi:10.1109/TMI.2007.908687.

The Meaning and Use of the Volume Under a Three-Class ROC Surface (VUS)

Xin He, IEEE[*] [Member] and
Department of Radiology, Johns Hopkins School of Medicine, 601 N. Caroline Street, Baltimore, MD 21287 USA.

Eric. C. Frey, IEEE [Senior Member]
Department of Radiology, Johns Hopkins School of Medicine, Baltimore, MD 21287 USA (e-mail: efrey@jhmi.edu).

Abstract

Previously, we have proposed a method for three-class receiver operating characteristic (ROC) analysis based on decision theory. In this method, the volume under a three-class ROC surface (VUS) serves as a figure-of-merit (FOM) and measures three-class task performance. The proposed three-class ROC analysis method was demonstrated to be optimal under decision theory according to several decision criteria. Further, an optimal three-class linear observer was proposed to simultaneously maximize the signal-to-noise ratio (SNR) between the test statistics of each pair of the classes provided certain data linearity condition. Applicability of this three-class ROC analysis method would be further enhanced by the development of an intuitive meaning of the VUS and a more general method to calculate the VUS that provides an estimate of its standard error. In this paper, we investigated the general meaning and usage of VUS as a FOM for three-class classification task performance. We showed that the VUS value, which is obtained from a rating procedure, equals the percent correct in a corresponding categorization procedure for continuous rating data. The significance of this relationship goes beyond providing another theoretical basis for three-class ROC analysis—it enables statistical analysis of the VUS value. Based on this relationship, we developed and tested algorithms for calculating the VUS and its variance. Finally, we reviewed the current status of the proposed three-class ROC analysis methodology, and concluded that it extends and unifies decision theoretic, linear discriminant analysis, and psychophysical foundations of binary ROC analysis in a three-class paradigm.

Keywords

Ideal observer; receiver operating characteristic (ROC) analysis; three-class classification

I. Introduction

PREVIOUSLY, we have developed a three-class decision model which produces a 2-D receiver operating characteristic (ROC) surface in a 3-D ROC space. The volume under the three-class ROC surface (VUS) was proved to be a figure-of-merit (FOM) for three-class task performance [1]. Having explored the decision theoretic and linear discriminant analysis (LDA) foundations of three-class ROC analysis [2], [3], this paper aims at exploring the psychophysical foundation of this proposed three-class ROC analysis, which is

[*](e-mail: xinhe@jhmi.edu).

inspired by the two-alternative forced-choice (2AFC) procedure and its relationship to binary ROC analysis [4], [5]. This relationship provides the psychophysical foundation for binary ROC analysis, and will be extended in a three-class paradigm in the present paper. The significance of the extension of this relationship goes beyond providing another theoretical basis for three-class ROC analysis—it enables statistical analysis of VUS value. We present and test algorithms for calculating VUS and its variance. At the end of this paper, we provide a short discussion of the current status of the proposed three-class ROC methodology.

II. Background

As a method to evaluate binary classification task performance, conventional ROC analysis has been extensively studied [4]-[11]. A binary classification task can be performed using two procedures: a rating procedure, whose performance is described by area under an ROC curve (AUC), or a 2AFC procedure, whose performance is described by the percent correct. We briefly introduce these two procedures and their relationship.

A medical diagnostic task is often modeled as a classification task using a rating procedure. In such a procedure, the observer is presented with one of two mutually exclusive alternatives (e.g., signal-present image versus signal-absent image) at one time. In other words, the rating procedure has one observation interval [5]. The observer is then asked to rate his confidence level of which alternative is presented. Any number of responses may be used to rate the confidence level. For example, in a human observer signal detection task, a set of five confidence level responses is often used. Alternatively, an observer might be asked to use a continuous rating scale. An ROC curve is then traced out by calculating the sensitivity/specificity (TPF/TNF) pair for each confidence level. The area under this ROC curve gives the AUC value, which serves as a figure-of-merit for describing the task performance using a rating procedure.

A forced-choice design is a psychophysical procedure that can be used to avoid the problem of determining the observer's criterion (or confidence level) [5]. In a 2AFC procedure, two observation intervals are provided, i.e., two images, one from each alternative, are shown at the same time, e.g., a signal-absent image and a signal present image. The observer is instructed to categorize one of the images as signal-present and the other as signal absent. Note that since there are only two classes, this is equivalent to selecting which of the images has the signal present. The probability of correctly identifying which of the two stimuli is “signal-present” and which is “signal-absent” is defined as the percent correct.

In a binary classification task, Green and Swets showed the percent correct of a forced-choice procedure equals the AUC value in a rating procedure [5], [6]. This equivalence is of particular importance in binary ROC analysis [4]. It indicates that, for a given data source, the performance of a binary classification task could be identically determined using either a rating procedure or a 2AFC procedure. Further, both procedures result in the same scalar value that summarizes classification performance. In particular, when an investigator calculates the AUC value from a rating procedure, “he is in fact, or at least in mathematical fact, reconstructing random pairs of images, one from a diseased subject and one from a normal subject, and using the reader's separate ratings of these two images to simulate what the reader would have decided if theses two images had in fact been presented together as a pair in a 2AFC experiment” [4]. Bamber showed that this “probability of correctly ranking a (normal, abnormal)” pair is connected with the quantity calculated in the Wilcoxon or Mann–Whitney (M-W) statistical test [6]. As a result, the extensively-studied properties of M-W test can be used to predict the statistical properties of the area under a ROC curve.

Furthermore, this relationship enables us to study the AUC value and its properties without assumptions either on the distributions of the data or on the properties of the decision variable. Therefore, the AUC value obtained from a nonideal and non-Hotelling observer, e.g., a human observer, is interpretable in the sense of 2AFC.

III. Theory

Scurfield has previously investigated "*n*-event, *m*-dimensional" forced-choice tasks [12]. In that work, Scurfield first reformulated the two-class decision rules by introducing, for mathematical convenience, two dummy parameters which do not play a role in the observer's decision. As a natural extension of the reformulated two-class decision rules, a three-class decision rule was introduced which also added dummy parameters that do not play a role in the observer's decision. The resulting decision space is 2-D, and the decision structure has the same shape as the decision structure we derived under the ideal observer framework [1]. Scurfield proved that the volume under a 123-ROC surface (i.e., a surface in the 3-D space with axes T1F, T2F, and T3F, where TiF is the probability of correctly classifying the *i*th class, or the sensitivity of the *i*th class) equals the percent correct of an I_3A_6 (three interval, six-alternative) task. Since Scurfield's decision structure is identical to the decision-theory based one we have previously proposed, Scurfield's proof can be applied readily to this work. However, Scurfield's proof is very hard to follow, and it is not couched in terms that are familiar to the imaging community.

In the following, we first define a forced-choice procedure in a three-class paradigm; we then introduce the three-class decision model that has been proved to extend and unify the decision theoretic and LDA foundations of binary ROC analysis [1]-[3]. Next, we elaborate on the underlying connections between the proposed decision model and a three-class categorization procedure, and prove the equivalence of the percent correct and VUS in a different, and, we believe, more easily understood way. Based on this relationship, we propose methods and algorithms for calculating the VUS value and its standard deviation.

In the following, scalar variables are denoted with italic fonts, and functions are in regular fonts. For example, in $T1F = \mathrm{T1F}(x, y)$, $T1F$ is a variable, and T1F is a function.

A. Definition of the Three-Class Categorization Procedure

We now define a three-class categorization procedure that is analogous to the 2AFC procedure. In this procedure, three randomly sampled objects, one from each of the three distinct classes, are presented to the observer simultaneously. The observer's task is to categorize the three objects into each of the three hypotheses. The observer is said to make a correct decision when, and only when, all three objects are correctly categorized.

B. Decision Model for Three-Class ROC Analysis

We have derived the optimal decision variables and decision rules for practical three-class ROC analysis using a rating procedure [1]. For a given data vector $\vec{g}$, two decision variables (rating values), LR_{13}, and LR_{23}, are computed, i.e.,

$$\log LR_{13} = \log \frac{\mathrm{f}\left(\vec{g} \mid H_1\right)}{\mathrm{f}\left(\vec{g} \mid H_3\right)}$$

and

$$\log LR_{23} = \log \frac{f\left(\vec{g} \mid H_2\right)}{f\left(\vec{g} \mid H_3\right)} \quad (1)$$

where H_i (i = 1, 2, 3) denotes the ith hypothesis, $\vec{g}$ is the data vector, and $f\left(\vec{g} \mid H_i\right)$ is the likelihood of the data vector $\vec{g}$ under the ith hypothesis, H_i. To make a decision, a pair of ratings are calculated and compared to a decision structure centered on a critical point, which is determined by prior information (i.e., the decision utilities and prior probabilities of the classes, or predetermined sensitivity pairs). Fig. 1(a) shows the three-class decision plane and decision structure that was proved to be optimal under certain decision criteria in two previous papers [1], [2].

In order to relate the decision model to a corresponding categorization procedure, where any pairs of decision variables might be used, we provide the following mathematical treatment for the decision model. Since the result to be presented in this paper does not depend on the use of the decision variables, we replace log LR_{13} and log LR_{23} with a pair of general decision variables, x and y, as shown in Fig. 1(b). The general decision variables, (x, y), might be any pair of possible decision variables representing a pair of ratings assigned to each object. Note that for the same classification problem, if the decision variables that span the decision plane are different, the distributions on the decision plane would also be different.

The rating pair distributions of the three classes on this general decision plane can thus be represented by $f_1(x, y)$, $f_2(x, y)$, and $f_3(x, y)$, respectively. Fig. 1(c) shows a typical ROC surface. The volume under the ROC surface is given by

$$VUS = \int_0^1 \int_0^1 T1F\, dT2F\, dT3F \quad (2)$$

where TiF(i = 1, 2, 3) is the probability that class i can be correctly classified. It can be seen that TiF is a function of x and y, i.e.,

$$TiF = \mathrm{TiF}(x, y). \quad (3)$$

C. Mathematical Treatment of the Three-Class Categorization Procedure

In order to relate the three-class categorization procedure to the rating procedure described above, we now analyze the categorization procedure mathematically as a three-step process.

Step 1) Present a triplet of randomly grouped class 1, class 2, and class 3 objects to the observer.

Step 2) Rate each object independently as in a rating procedure, resulting in a rating pair, (x, y), for each object as in the rating procedure.

Step 3) Test the three rating pairs associated with each object to see if the differences among the rating pairs lead to them being correctly categorized according to the decision rules suggested by the decision structure in Fig. 1. That is to say, a triplet can be correctly classified if there exists a decision structure position (defined by a critical point) such that all three rating pairs are correctly classified. Note here that the outcome of this test depends only on the relative positions of the rating pairs on the decision plane.

The described procedure is illustrated graphically in Fig. 2, where the triangle, square, and disk represent the randomly sampled objects, with one from each rating pair distributions $f_1(x, y)$, $f_2(x, y)$, and $f_3(x, y)$, respectively. Fig. 2(a) illustrates a case when at least one decision structure position exists that results in correctly categorizing these three rating pairs, and Fig. 2(b) illustrates a case when no such position exists. Note that the underlying assumption here is that the rating procedure and the categorization procedure use the same decision structure and decision variables. By repeating this procedure and computing the ratio of correct to total trials, we can thus estimate the percent correct for this categorization procedure from the rating data produced by a rating procedure.

D. Equivalence of VUS in a Rating Procedure and Percent Correct in a Categorization Procedure

Given the above mathematical treatment, we can compute the percent correct (PC) as the integration of the product of the probability distributions of the rating pairs for the three classes over all the combinations that can give correct decisions. The formulation of PC is thus obtained by finding a strategy such that a complete set of correctly classified triplets is obtained without counting any triplet more than once, and then integrating the probabilities of all possible correct classifications. In the Appendix, we prove that such a formulation of percent correct leads to the same expression for VUS, i.e.,

$$VUS = PC. \quad (4)$$

Note that the proof is done for continuous rating data. In other words, samples of rating values from the continuous distributions cannot give rise to rating triplets at identical positions in the decision space.

IV. Statistical Analysis

Given the equivalence of VUS and percent correct, we now propose methods for statistical analysis of VUS value.

A. Calculation of the VUS Value

Since, as described above and proved in the Appendix, the VUS equals the percent correct in the corresponding three-class categorization task, we can thus estimate PC as a substitute for estimating the VUS. With sample sizes of n_1, n_2, and n_3 from continuous class 1, class 2, and class 3 distributions, respectively, the rating procedure will result in n_k rating pairs for class k (where $k = 1, 2, 3$). We denote each rating pair as $\vec{p}_{ki} = (x_{ki}, y_{ki})$, where $i = 1 \ldots n_i$ refers to the ith sample in the kth class. The corresponding categorization procedure, at least conceptually, consists of making all possible $n_1 \cdot n_2 \cdot n_3$ comparisons among the ratings from the three classes and summing the score, $\mathrm{U}\left(\vec{p}_{1l}, \vec{p}_{2m}, \vec{p}_{3n}\right)$, for each comparison according to the rule

$$\mathrm{U}\left(\vec{p}_{1l}, \vec{p}_{2m}, \vec{p}_{3n}\right) = \begin{cases} 1, & \text{correct categorization} \\ 0, & \text{otherwise} \end{cases} . \quad (5)$$

where l, m, and n are the lth, mth, and nth samples from classes 1, 2, and 3, respectively. In (5), correct categorization is said to occur when there exists a critical point (this is a position for the decision structure) such that $\vec{p}_{1l}$ falls into the area for a class 1 decision, $\vec{p}_{2m}$ falls into the area for a class 2 decision, and $\vec{p}_{3n}$ falls into the area for a class 3 decision. The percent correct, PC, is then estimated by averaging the $\{U\}$ over all $n_1 \cdot n_2 \cdot n_3$ comparisons, i.e.,

$$\widehat{PC}=\frac{1}{n_1\cdot n_2\cdot n_3}\sum_{1}^{n_1}\sum_{1}^{n_2}\sum_{1}^{n_3}\mathrm{U}\left(\vec{p}_{1l},\vec{p}_{2m},\vec{p}_{3n}\right) \quad (6)$$

where the caret above *PC* indicates that it is an estimate. Given the equivalence between the VUS and PC, VUS can be evaluated using (6).

Equation (5) is very similar to its counterpart in the 2AFC procedure, as described in [4]. In the binary case, when discrete rating data are used, it is possible for two samples from different classes to be identical, and the decision rule will assign a value of 0.5 for percent correct calculation [4]. One may note (5) does not include a contingency for the case where some of the ratings are the same. This is because, in this paper we only proved $VUS = PC$ for a continuous rating scale, where we do not need to consider the case where random samples of the ratings are identical. The possible relationship between VUS and PC for three-class discrete rating scale is more complicated than in the binary case and is beyond the scope of this paper.

B. Estimation of the Variance of the VUS Value

We have implemented two methods for variance estimation. The first method is based on Dreiseitl's extension [13] of Lehmann's nonparametric approach [14]. In Dreiseitl's work, algorithms for estimating the volume under the surface value and its variance were proposed for Mossman's three-class decision model. Dreiseitl's algorithms for volume-under-the-surface estimation are very similar to (5) and (6). The only difference lies in the definition of correct categorization in (5). This is because Mossman's decision model uses different decision structures, requiring a different correct categorization test. However, this difference does not affect the formulation for VUS and its variance, and Dreiseitl's variance calculation algorithm can be readily applied to this work as explained below.

In this method, the variance is given by [13]

$$\begin{aligned}\mathrm{var}(VUS) &= \mathrm{cov}(VUS, VUS)\\ &= \frac{1}{n_1^2 n_2^2 n_3^2}\\ &\times\sum_i\sum_j\sum_k\sum_I\sum_J\sum_K \mathrm{cov}\left(U_{ijk}, U_{IJK}\right).\end{aligned} \quad (7)$$

In (7), $U_{ijk}=\mathrm{U}\left(\vec{p}_{1i},\vec{p}_{2j},\vec{p}_{3k}\right)$, $\mathrm{cov}(U_{ijk}, U_{mnl}) = \mathrm{E}(U_{ijk}U_{mnl}) - \theta^2$, where E() is the expectation operation, $U_{ijk}=\mathrm{U}\left(\vec{p}_{1i},\vec{p}_{2j},\vec{p}_{3k}\right)$, and $\theta = \mathrm{E}(U_{ijk})$, which is the true percent correct. Note that $\mathrm{E}(U_{ijk}U_{mnl})$ is the probability that both U_{ijk} and U_{mnl} are 1. Expanding (7), as described in [13] results in the following formula for variance of the VUS

$$\begin{aligned}&\mathrm{var}(VUS)\\ &= \frac{1}{n_1 n_2 n_3}\left\{\theta(1-\theta)+(n_3-1)\left[\mathrm{E}\left(U_{ijk}U_{ijK}\right)-\theta^2\right]\right.\\ &\quad +(n_1-1)(n_2-1)\left[\mathrm{E}\left(U_{ijk}U_{IJk}\right)-\theta^2\right]\}\\ &\quad +\frac{1}{n_1 n_2 n_3}\left\{(n_2-1)\left[\mathrm{E}\left(U_{ijk}U_{iJk}\right)-\theta^2\right]\right.\\ &\quad +(n_1-1)\left[\mathrm{E}\left(U_{ijk}U_{Ijk}\right)-\theta^2\right]\}\\ &\quad +\frac{1}{n_1 n_2 n_3}\left\{(n_2-1)(n_3-1)\left[\mathrm{E}\left(U_{ijk}U_{iJK}\right)-\theta^2\right]\right.\\ &\quad +(n_1-1)(n_3-1)\left[\mathrm{E}\left(U_{ijk}U_{IjK}\right)-\theta^2\right]\}.\end{aligned} \quad (8)$$

Note that the elements in each expectation in (8) have at least one identical subscript, e.g., the two elements in E($U_{ijk}U_{IJk}$) have subscripts *ijk* and *IJk*, respectively. E($U_{ijk}U_{IJk}$) is de fined as

$$\mathrm{E}\left(U_{ijk}U_{IJk}\right)=\sum_{i}\sum_{j}\sum_{k}\sum_{I\neq i}\sum_{J\neq j}\left(U_{ijk}U_{IJk}\right). \quad (9)$$

Other expectations in (8) have analogous definitions.

The calculation of each E($U_{ijk}U_{mnl}$) in (8) requires four or five nested loops, which, as will be demonstrated later, takes a long time when the number of rating pairs is large. Therefore, we propose to implement a second method using a bootstrap approach [15], [16] to estimate the variance. The steps involved in this method are as follows.

Step 1) Formulate three empirical probability rating pair distributions, $\tilde{\mathrm{f}}_i(x,y)$, to estimate the real rating pair distribution, $\mathrm{f}_i(x,y)$, where I represents the ith class. For a standard Bootstrap approach, $\tilde{\mathrm{f}}_i(x,y)$ is expressed as a discrete distribution such that each of the n_i samples from $\tilde{\mathrm{f}}_i(x,y)$ from class i has a probability of $1/n_i$. Despite the fact that $\tilde{\mathrm{f}}_i(x,y)$ is discrete, each rating pair is a sample from the continuous rating distribution, $\mathrm{f}_i(x,y)$. In cases where the sample size is small, one might want to fit the distribution of the data with a normal distribution as is done in binary ROC analysis. However, we have not yet been able to derive a three-class counterpart of the binormal ROC curve fitting. Thus we provide an alternative approach using a simple parametric Bootstrap method, where the available experimental samples are used to fit bivariate Gaussian distributions to estimate $\tilde{\mathrm{f}}_1(x,y)$, $\tilde{\mathrm{f}}_2(x,y)$ and $\tilde{\mathrm{f}}_3(x,y)$, respectively. To be specific, using the experimentally obtained rating pairs of the ith class, we compute the mean, variance and the covariance of x and y, and then use these parameters to formulate the empirical probability distribution of the rating pairs of the ith class. Note that the parametric bootstrap approach is not essential, but does have the advantage of handling the case for discrete ratings and, to the extent the data are described by bivariate Gaussian distributions, more precise estimates of the VUS and its variance.

Step 2) Take n_1 random samples from $\tilde{\mathrm{f}}_1(x,y)$, n_2 random samples from $\tilde{\mathrm{f}}_2(x,y)$, and n_3 random samples from $\tilde{\mathrm{f}}_3(x,y)$ with replacement.

Step 3) Make all possible $n_1 \cdot n_2 \cdot n_3$ comparisons among the random samples from each class. Calculate the fraction, Γ, of the samples of comparisons that result in a correct categorization using (5)-(6).

Step 4) Repeat Steps 2 and 3 B times, to create B bootstrap samples. This results in a set of estimates of the percent correct $\Gamma_1, \Gamma_2, \Gamma_3, ..., \Gamma_B$.

Step 5) The distribution of Γ estimates the distribution of PC, which estimates the distribution of the VUS value. The variance of the VUS is estimated by the variance of Γ.

V. Experiments

We have implemented the algorithms for estimating the VUS and its variance. To test the algorithms, we used a set of data from a previous experiment on dual-isotope myocardial perfusion SPECT (MPS) image quality evaluation described in [17], [18], where the rest-stress MPS images were obtained from a simulated population of normal patients and patients with reversible or fixed defects. For each of the three classes, a total of 432 rating pairs were generated. Fig. 3(a) shows the decision plane and Fig. 3(b) shows the corresponding three-class ROC surface.

To validate the VUS calculation algorithm proposed, we developed an ad hoc VUS calculation algorithm, as illustrated in Fig. 4. The rating pair distributions were sampled on a grid with very small grid spacing [Fig. 4(a)]. In the ROC space, a relatively large bin size was defined on the (T1F, T2F) plane [Fig. 4(b)]. Moving the decision structure in the decision plane with very small grid spacing produced several $T3F$ values in each of the (T1F, T2F) bin in the ROC space. For each (T1F, T2F) bin in the ROC space, we average the T3F values to produce an ROC histogram. An estimate of the VUS value was obtained by averaging the histogram for all (T1F, T2F) bins. The accuracy of the VUS can be increased by decreasing the grid spacing.

The experimentally obtained data were too sparse to enable small grid spacing in the experimental decision plane. We thus fit the three rating pair distributions to bivariate Gaussian distributions for the ad hoc algorithm. This was done based on our previous unpublished study of the normality of the rating pairs. Note that the ad hoc method is a very coarse method, its resulting VUS value is affected by the bin sizes on both the decision plane and the histogram. We developed this ad hoc method only to obtain a rough idea of the magnitude of the VUS value to test whether the theoretical method provides a VUS that is similar. For the method based on (6), we did not use the bivariate Gaussian fitting. Table I shows the parameters of the fitted bivariate Gaussian distributions for the three classes. To validate the variance calculation, we simply compared the estimate from Dreiseitl's extension [13] in (8) and the proposed Bootstrap method. We also compared the computational times for (8) and the Bootstrap method.

Since the data were obtained from 432 triplets of rating values, we used $n_1 = n_2 = n_3 = 432$ in calculating VUS using (6). The results are shown in Table II, where we see that methods based on $VUS = PC$ agreed well with the ad hoc method for estimating the VUS value. To calculate the variance using the Bootstrap method, we used $B = 1000$ repetitions. It can be seen from Fig. 5 that VUS and its variance were well-converged after 200 repetitions. The resulting standard deviation using the Bootstrap method after 1000 iterations was 0.181 (Fig. 5), which is in the same order as the one obtained using (8). However, the algorithm for obtaining standard deviation using (8) took approximately eight days. This is because, as described above, each $E(U_{ijk}U_{mnl})$ in (8) involves a nested summation loop four or five levels deep [an example is given in (9)]. The five level deep loops dominate the computational time. Thus, when $n = n_1 = n_2 = n_3$, the nested summation loops result in a computational time that is roughly proportional to n^5. For example, using a 2.13 GHz AMD Opteron processor, when $n = 100$, it took 418 s to compute the variance, while when $n = 200$, it took 13, 358 s, approximately 31.96 times longer. In our study, with $n = 432$, the computational time was about eight days. The Bootstrap method, on the other hand, required only 15 min for 200 repetitions, as shown in Table II.

VI. Discussion

Current Status of the Proposed Three-Class ROC Analysis

Binary ROC analysis has been a standard method for assessing diagnostic performance. However, there are an increasing number of diagnostic tasks of interest for which binary classification is not sufficient. In particular, in many cases diagnosing disease using imaging techniques requires both detection and characterization of the disease instead of disease detection alone; analysis of these cases requires ROC analysis techniques for analyzing multi-class diagnoses. However, multiclass ROC analysis is a theoretical problem whose solution has been eluded the community ever since the introduction of the binary ROC in the 1950s [19], [20]. Much work has been devoted to understanding the nature of a multiclass classification problem, and many metrics have been proposed to assess the performance of a multiclass classification task [12], [13], [19]-[32].

Motivated by the medical problem of cardiac perfusion defect evaluation using simultaneous dual-isotope myocardial perfusion SPECT (MPS), where the assessment of a three-class diagnostic task is required to evaluate and optimize MPS imaging techniques, we have carried out a series of studies on the theoretical foundations for a practical three-class ROC analysis method. The present paper results in both an additional theoretical justification for the proposed method, and a practical method to calculate its figure of merit, the volume under a three-class ROC surface. In the following, we review the previous developments in order to present a more complete picture of the theoretical framework.

Model Development

We first developed an optimal three-class decision model that maximizes the expected utility under decision theory by assuming that incorrect decisions have equal utilities under the same hypothesis (the equal error utility assumption). This decision model produces a 2-D ROC surface in a 3-D ROC space and the volume under this surface (VUS) is a figure-of-merit for three-class task performance. We have compared the proposed three-class ROC analysis with conventional binary ROC analysis and concluded that they share many similar properties and that three-class ROC analysis reduces to binary ROC analysis for certain special cases [1].

Decision Theoretic Foundation

We thoroughly investigated the decision theoretic foundations of the proposed three-class ROC analysis and proved the optimality of the three-class ideal observer (3-IO) according to several decision theoretic criteria. In particular, we found that the 3-IO and the proposed decision model maximizes the expected utility (MEU) under equal error utility assumption, maximizes the probability of making correct decisions, provides the maximum likelihood (ML) decision, and satisfies the Neyman-Pearson (N-P) criterion in the sense that, given the sensitivities of two classes, the sensitivity of the third class is maximized [2]. We believe that the optimality with respect to N-P criterion is of particular importance for clinical applications, as explained in [2] and [3].

Linear Discriminant Analysis (LDA) Foundation

We then investigated the LDA foundation of three-class ROC analysis. We have shown that the conventional multiclass extension of LDA has significant limitations. In particular, the L-class Hotelling trace, which has been used as a figure-of-merit for multiclass task performance, cannot distinguish cases where all classes are perfectly classified from cases where only one of the classes can be correctly classified [33]. Using the proposed three-class ROC analysis method, we have found that when the data follow multivariate Gaussian distribution with equal covariance matrices, in addition to the optimality mentioned above, the proposed three-class decision model maximizes the SNR between each pairs of the classes, and likelihood ratios can be computed using a linear observer, i.e., the three-class Hotelling observer (3-HO) [3]. When the data are not Gaussian distributed, we have shown that 3-HO still maximizes the SNR between each pair of the classes given a certain data linearity condition [3].

Psychophysical Foundation

In this paper, we have investigated the relationship between a three-class rating procedure and the corresponding categorization procedure. The equivalence of VUS and the percent correct extends the psychophysical foundation of binary ROC analysis to the proposed three-class ROC analysis.

From the above, we conclude that the proposed three-class ROC analysis method extends and unifies the decision theoretic, linear discriminant analysis, and psychophysical foundations of binary ROC analysis in a three-class paradigm. The proposed three-class ROC methodology is practical mathematically in the sense that a figure-of-merit (FOM) was proposed along with practical numerical methods to obtain the FOM and its statistical properties [1]-[3]. Additionally, the proposed methodology might also prove to be a reasonable model of clinical decision making, as described in [2] and [3].

VII. Conclusion

In this paper, we investigated the psychophysical foundation of a previously-proposed three-class ROC methodology by presenting an intuitive meaning for the VUS value, i.e., the percent correct in a three-class categorization procedure for continuous rating data. This equivalence is neither dependent on the decision variables used, nor dependent on the actual distributions of the three classes. In other words, no matter what decision variables are used, the VUS obtained always equals the percent correct in a corresponding three-class categorization procedure when using the decision rules defined by the decision structure used in this work. Based on this psychophysical foundation, we developed and tested algorithms for calculating the VUS and its variance.

In light of this connection to the proposed psychophysical task, we reviewed the current status of the proposed three-class ROC analysis methodology, and concluded that it extends and unifies decision theoretic, linear discriminant analysis, and psychophysical foundations of binary ROC analysis in a three-class paradigm.

Acknowledgments

The authors would like to thank their colleagues Dr. D. S. Graff, Dr. J. M. Links, and Dr. B. M. W. Tsui for their helpful comments and thought-provoking discussions with regard to this work and manuscript.

This work was supported by the National Institutes of Health (NIH) under Grant K99-EB007620, Grant R01-EB000288, and Grant R01-HL068575. The content of this work is solely the responsibility of the authors and does not necessarily represent the official view of the NIH or its various institutes.

Appendix

We denote the rating pair distributions as $f_1(x, y)$, $f_2(x, y)$, and $f_3(x, y)$ for each of the three classes, respectively, and the corresponding true class fractions as

$$T1F = \mathrm{T1F}(x,y) = \int_{x}^{\infty} \int_{-\infty}^{y-x+x'} \mathrm{f}_1\left(x', y'\right)\, dy'\, dx' \quad \text{(A1)}$$

$$T2F = \mathrm{T2F}(x,y) = \int_{y}^{\infty} \int_{-\infty}^{y'-y+x} \mathrm{f}_2\left(x', y'\right)\, dx'\, dy' \quad \text{(A2)}$$

and

$$T3F = \mathrm{T3F}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} \mathrm{f}_3\left(x', y'\right)\, dy'\, dx'. \quad \text{(A3)}$$

Note that scalar variables are represented using italic fonts while functions are in regular fonts. In the following, we prove the equivalence of the VUS and the PC with respect to a categorization procedure.

A. Preparing Necessary Partial Derivatives and Three Lemmas

Before proving the equivalence of the VUS and the PC in a categorization procedure, we first introduce three lemmas and derive the partial derivatives of T2F(x, y) and T3F(x, y), which will be used later in the proof.

1) The Three Lemmas

Lemma 1—The purpose of Lemma 1 is to change the bound of the integrals, and is expressed as

$$\int_b^{\infty}\int_a^{y-b+a} \mathrm{f}(x,y)\ dx\,dy = \int_b^{\infty}\int_a^{\infty} \mathrm{f}(x,y+x-a)\ dx\,dy. \quad \text{(A4)}$$

Proof:

$$\int_b^{\infty}\int_a^{y-b+a} \mathrm{f}(x,y)\ dx\,dy = \int_b^{\infty}\int_a^{\infty} \mathrm{f}(x,y)\,\Theta(y-b+a-x)\ dx\,dy.$$

where Θ is the Heaviside step function. Now let $y' = y - x + a$

$$\begin{aligned}\int_b^{\infty} & \int_a^{\infty} \mathrm{f}(x,y)\,\Theta(y-b-a-x)\ dx\,dy \\ &= \int_b^{\infty}\int_a^{\infty} \mathrm{f}\left(x,y'+x-a\right)\Theta\left(y'-b\right)\ dx dy' \\ &= \int_b^{\infty}\int_a^{\infty} \mathrm{f}\left(x,y'+x-a\right)\ dx dy'.\end{aligned}$$

Let $y = y'$

$$\int_b^{\infty}\int_a^{\infty} \mathrm{f}\left(x,y'+x-a\right)\ dx dy' = \int_b^{\infty}\int_a^{\infty} \mathrm{f}(x,y+x-a)\ dx\,dy.$$

Lemma 2—Lemma 2 is very similar to Lemma 1, and is expressed as

$$\int_b^{\infty}\int_{y-b+a}^{\infty} \mathrm{f}(x,y)\ dx\,dy = \int_b^{\infty}\int_a^{\infty} \mathrm{f}(x+y-b,y)\ dx\,dy. \quad \text{(A5)}$$

Proof:

$$\int_b^{\infty}\int_{y-b+a}^{\infty} \mathrm{f}(x,y)\ dx\,dy = \int_b^{\infty}\int_{-\infty}^{\infty} \mathrm{f}(x,y)\,\Theta(x-y+b-a)\ dx\,dy$$

where Θ is the Heaviside step function, let $x' = x - y + b$

$$\begin{aligned}\int_b^{\infty} & \int_{-\infty}^{\infty} \mathrm{f}(x,y)\,\Theta(x-y+b-a)\ dx\,dy \\ &= \int_b^{\infty}\int_{-\infty}^{\infty} \mathrm{f}\left(x'+y-b,y\right)\Theta\left(x'-a\right)\ dx'\,dy \\ &= \int_b^{\infty}\int_a^{\infty} \mathrm{f}\left(x'+y-b,y\right)\ dx'\,dy\end{aligned}$$

Let $x = x'$

$$\int_b^{\infty}\int_a^{\infty} \mathrm{f}\left(x'+y-b,y\right)\,dx'\,dy = \int_b^{\infty}\int_a^{\infty} \mathrm{f}\,(x+y-b,y)\,dx\,dy.$$

Lemma 3—The purpose of Lemma 3 is to change the order of the double integral, and is expressed as

$$\int_{-\infty}^{\infty} \mathrm{f}\,(y)\left[\int_y^{+\infty} \mathrm{g}\left(y'\right)dy'\right]dy = \int_{-\infty}^{+\infty} \mathrm{g}\left(y'\right)\left[\int_{-\infty}^{y'} \mathrm{f}\,(y)\,dy\right]dy'. \quad \text{(A6)}$$

Proof:

$$\begin{aligned}\int_{-\infty}^{\infty} \mathrm{f}\,(y)\left[\int_y^{+\infty} \mathrm{g}\left(y'\right)dy'\right]\; dy \\ = \int_{-\infty}^{\infty} \mathrm{f}\,(y)\left[\int_{-\infty}^{+\infty} \mathrm{g}\left(y'\right)\Theta\left(y'-y\right)dy'\right]dy\end{aligned}$$

where Θ (*y*) is the Heaviside step function. Now the order of the integration can be changed.

$$\begin{aligned}&\int_{-\infty}^{\infty} \mathrm{f}\,(y)\left[\int_{-\infty}^{+\infty} \mathrm{g}\left(y'\right)\Theta\left(y'-y\right)dy'\right]dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{+\infty} \mathrm{f}\,(y)\,\mathrm{g}\left(y'\right)\Theta\left(y'-y\right)\,dy\,dy' \\ &= \int_{-\infty}^{+\infty} \mathrm{g}\left(y'\right)\left[\int_{-\infty}^{y'} \mathrm{f}\,(y)\,dy\right]dy'.\end{aligned}$$

2) Partial Derivatives of T2F (*x*, *y*)

For a decision structure centered on (*x, y*), T2F(*x, y*) is expressed in (A2), and the partial derivative of T2F(*x, y*) with respect to *x* is thus

$$\frac{\partial \mathrm{T2F}\,(x,y)}{\partial x} = \int_y^{\infty} \mathrm{f}_2\left(y'-y+x,y'\right)\,dy'. \quad \text{(A7)}$$

To calculate the partial derivative of T2F(*x, y*) with respect to *y,* we note that T2F(*x, y*) is an integral over a region bounded below by a horizontal ray extending from the origin of the decision structure to -∞ and a second ray from the origin of the decision structure along the 45° line from the origin of the decision structure toward (∞, ∞). We use the observation above to rewrite T2F(*x, y*) as

$$\begin{aligned}\mathrm{T2F}\,(x,y) = \int_y^{\infty}\int_{-\infty}^{x} \mathrm{f}_2 \quad \left(x',y'\right)dx'\,dy' \\ + \int_y^{\infty}\int_x^{y'-y+x} \mathrm{f}_2\left(x',y'\right)dx'\,dy'.\end{aligned} \quad \text{(A8)}$$

Applying Lemma 1 in (A4) to the second term of (A8), T2F(*x, y*) can be expressed as

$$\begin{aligned}\mathrm{T2F}\,(x,y) = \quad \int_y^{\infty}\left[\int_{-\infty}^{x} \mathrm{f}_2\left(x',y'\right)\,dx'\right. \\ \left. + \int_x^{\infty} \mathrm{f}_2\left(x',y'-x+x'\right)dx'\right]dy'.\end{aligned} \quad \text{(A9)}$$

Using this we find that

$$\frac{\partial \mathrm{T2F}(x,y)}{\partial y} = -\int_{-\infty}^{x} \mathrm{f}_2\left(x', y\right)\, dx' - \int_{x}^{\infty} \mathrm{f}_2\left(x', y - x + x'\right)\, dx'. \quad \text{(A10)}$$

3) Partial Derivatives of T2F(*x*, *y*)

For a decision structure centered on (*x*, *y*), T3F(*x*, *y*) is given by (3). Its partial derivatives with respect to *x* and *y* are thus

$$\frac{\partial \mathrm{T3F}(x,y)}{\partial x} = \int_{-\infty}^{y} \mathrm{f}_3\left(x, y'\right)\, dy'$$

and

$$\frac{\partial \mathrm{T3F}(x,y)}{\partial y} = \int_{-\infty}^{x} \mathrm{f}_3\left(x', y\right)\, dx'. \quad \text{(A11)}$$

B. Strategy for Computing the Percent Correct

We now describe the strategy for computing percent correct; a full mathematical derivation based on this strategy will be given in the next section. The percent correct on a three-class categorization procedure is given by

$$PC = \iint \iint \iint \{\mathrm{f}_1(x_1, y_1)\, \mathrm{f}_2(x_2, y_2)\, \mathrm{f}_3(x_3, y_3) \times \mathrm{c}(x_1, y_1, x_2, y_2, x_3, y_3)\}\, dx_1\, dy_1\, dx_2\, dy_2\, dx_3\, dy_3 \quad \text{(A12)}$$

where $c = \mathrm{c}(x_1, y_1, x_2, y_2, x_3, y_3)$ is 1 if there exists a position of the decision structure such that the triplet of rating pairs $((x_1, y_1), (x_2, y_2), (x_3, y_3))$can be correctly classified, and is 0 otherwise. Rearranging (A12), we obtain

$$PC = \iint \mathrm{f}_3(x_3, y_3) \left\{ \iint \mathrm{f}_2(x_2, y_2) \times \left[\iint \mathrm{f}_1(x_1, y_1)\, \mathrm{c}(x_1, y_1, x_2, y_2, x_3, y_3)\, dx_1\, dy_1 \right] \times dx_2\, dy_2 \right\} dx_3\, dy_3. \quad \text{(A13)}$$

Examination of (A13) reveals that the outermost double integral can be expressed as

$$PC = \iint \mathrm{f}_3(x_3, y_3)\, \mathrm{p}(c=1 \mid x_3, y_3)\, dx_3\, dy_3 \quad \text{(A14)}$$

where

$$\mathrm{p}(c=1 \mid x_3, y_3) = \iint \mathrm{f}_2(x_2, y_2) \left[\iint \mathrm{f}_1(x_1, y_1) \times \mathrm{c}(x_1, y_1, x_2, y_2, x_3, y_3)\, dx_1\, dy_1 \right] dx_2\, dy_2. \quad \text{(A15)}$$

Here, $\mathrm{p}(c = 1|x_3, y_3)$ is the probability density for a correct classification over all possible (x_1, y_1) and (x_2, y_2) combinations for the given (x_3, y_3). Similarly, $\mathrm{p}(c = 1|x_3, y_3)$ can be expressed as

$$\mathrm{p}(c=1 \mid x_3, y_3) = \iint \{\mathrm{f}_2(x_2, y_2) \times \mathrm{p}(c=1 \mid x_2, y_2, x_3, y_3)\}\, dx_2\, dy_2 \quad \text{(A16)}$$

where $\mathrm{p}(c = 1|x_2, y_2, x_3, y_3)$ is the probability density for correct classification over all possible (x_1, y_1) for the given (x_2, y_2) and (x_3, y_3) combination, and is expressed as

$$\mathrm{p}(c=1 \mid x_2, y_2, x_3, y_3) = \iint \mathrm{f}_1(x_1, y_1)\, \mathrm{c}(x_1, y_1, x_2, y_2, x_3, y_3)\, dx_1 dy_1. \quad \text{(A17)}$$

Equations (A14), (A16), and (A17) provide a natural strategy to evaluate the six-dimensional integral in (A12). We first select a (x_3, y_3) pair. Next, we recognize that all (x_2, y_2) for which correct classifications are possible will necessarily have $y_2 > y_3$, as seen in Fig. 6(a). For such pairs, the maximum region containing (x_1, y_1) that is correctly classifiable will be obtained if the decision structure is positioned such that the selected (x_2, y_2) and (x_3, y_3) lie on the rays comprising the decision structure. For a given (x_3, y_3), there are three subsets of (x_2, y_2) for which correct classification is possible, as described in Table III and shown in Fig. 6(b)–(d), depending on which of the rays the two pairs of decision variables are located on. Thus, what we need to do is to add the probability density for correct classification for each of these subsets. Integrating over all the (x_2, y_2) pairs for a given (x_3, y_3) and then over all the (x_3, y_3) will give the percent correct.

C. Computing the Percent Correct

First consider $\mathrm{p}(c = 1|x_3, y_3)$ for a particular (x_3, y_3). Note that

$$\mathrm{p}(c=1 \mid x_2, y_2, x_3, y_3) = 0 \quad \text{for} \quad y_2 \le y_3. \quad \text{(A18)}$$

Thus, (A16) can also be written as

$$\mathrm{p}(c=1 \mid x_3, y_3) = \iint_{y_2 > y_3} \mathrm{f}_2(x_2, y_2)\, \mathrm{p} \times (c=1 \mid x_2, y_2, x_3, y_3 l)\, dx_2\, dy_2 \quad \text{(A19)}$$

where the double integral is over the half plane where $y_2 > y_3$. As described above, there are three nonintersecting subsets of (x_2, y_2) that satisfy $y_2 > y_3$, defined by their relative locations; we label these Subset 1, Subset 2 and Subset 3, respectively, as shown in Table III and illustrated in Fig. 6(a).

Given the subsets defined in Table III, (A19) is expressed as

$$\begin{aligned}\mathrm{p}(c=1 \mid x_3, y_3) &= \iint_{\text{Subset1}} \mathrm{f}_2(x_2, y_2)\, \mathrm{p}(c=1 \mid x_2, y_2, x_3, y_3)\, dx_2\, dy_2 \\ &+ \iint_{\text{Subset2}} \mathrm{f}_2(x_2, y_2)\, \mathrm{p}(c=1 \mid x_2, y_2, x_3, y_3)\, dx_2\, dy_2 \\ &+ \iint_{\text{Subset3}} \mathrm{f}_2(x_2, y_2)\, \mathrm{p}(c=1 \mid x_2, y_2, x_3, y_3)\, dx_2\, dy_2.\end{aligned} \quad \text{(A20)}$$

Fig. 6 provides an intuitive illustration of the three subsets [Fig. 6(a)] and the corresponding strategies for identifying $\mathrm{p}(c = 1|x_2, y_2, x_3, y_3)$ for a (x_2, y_2) sampled from each subsets [Fig. 6(b)–(d)]. In particular, Fig. 6(b) shows the strategy for identifying $\mathrm{p}(c = 1|x_2, y_2, x_3, y_3)$ for a (x_2, y_2)sampled from Subset 1. We wish to find all (x_1, y_1) that may be correctly classified with this particular (x_2, y_2) and (x_3, y_3). To accomplish this, the decision structure should be moved toward $x = -\infty$, $y = +\infty$ to include as many (x_1, y_1)as possible. However, as shown in Fig. 6(b), the vertical line of the decision structure should not exceed $x = x_3$, and the

horizontal line should not exceed $y = y_2$. Otherwise, incorrectly classified triplets will be counted. As a result, the shaded area in Fig. 6(b) includes all (x_1, y_1) that form correct classification triplet with the particular (x_2, y_2), and (x_3, y_3) shown in Fig. 6(b). For this particular (x_2, y_2) and (x_3, y_3), $p(c = 1|x_2, y_2, x_3, y_3)$ is obtained by integrating $f_1(x_1, y_1)$ in the shaded area. Fig. 6(c)–(d) illustrate the strategies for identifying for $p(c = 1|x_2, y_2, (x_3, y_3)$ for (x_2, y_2) belonging to Subsets 2 and 3, respectively. It can be seen that the first, second and third term in (A20) are the probability densities for correct classification given a (x_3, y_3) for each of the three subsets of potentially-correctly-classifiable (x_2, y_2). In the following, we consider the computation of each of the three terms in (A20).

Let the first term be

$$A(x_3, y_3) = \iint_{\text{Subset1}} f_2(x_2, y_2)\, p(c=1 \mid x_2, y_2, x_3, y_3)\, dx_2\, dy_2. \quad \text{(A21)}$$

Now note that the true class 1 fraction, T1F(x, y), when the decision structure is located at (x, y) is equal to $p(c = 1|x_2, y_2, (x_3, y_3)$, i.e.,

$$\begin{aligned} &\text{if} \quad (x_2, y_2) \in \text{Subset} \quad 1 \\ &p(c=1 \mid x_2, y_2, x_3, y_3) = \text{T1F}(x_3, y_2). \end{aligned} \quad \text{(A22)}$$

Thus, referring to Fig. 6(b) for the computation of $A(x_3, y_3)$, we see that $A(x_3, y_3)$ can be expressed as

$$A(x_3, y_3) = \int_{y_3}^{+\infty} \int_{-\infty}^{x_3} f_2(x_2, y_2)\, \text{T1F}(x_3, y_2)\, dy_2 dx_2. \quad \text{(A23)}$$

Rearranging (A23) gives

$$A(x_3, y_3) = \int_{y_3}^{+\infty} \left[\int_{-\infty}^{x_3} f_2(x_2, y_2)\, dx_2 \right] \text{T1F}(x_3, y_2)\, dy_2. \quad \text{(A24)}$$

Similarly, the second term in (A20), is

$$B(x_3, y_3) = \iint_{\text{Subset2}} f_2(x_2, y_2)\, p_c(x_2, y_2, x_3, y_3)\, dx_2\, dy_2. \quad \text{(A25)}$$

Referring to Fig. 6(c) and again recognizing the true class 1 fraction, we see that we can rewrite this as

$$\begin{aligned} B(x_3, y_3) = \int_{y_3}^{\infty} \quad & \int_{x_3}^{y_2 - y_3 + x_3} f_2(x_2, y_2) \\ & \times \text{T1F}(x_3, y_2 - x_2 + x_3)\, dx_2\, dy_2. \end{aligned} \quad \text{(A26)}$$

We now apply Lemma 1 [given in (A2)] to change the bounds of the integrals in $B(x_3, y_3)$, obtaining

$$\begin{aligned} B(x_3, y_3) = \int_{y_3}^{\infty} \quad & \left[\int_{x_3}^{\infty} f_2(x_2, y_2 + x_2 - x_3)\, dx_2 \right] \\ & \times \text{T1F}(x_3, y_2)\, dy_2. \end{aligned} \quad \text{(A27)}$$

Thus, the sum of the first two terms of (A20) is

$$\begin{aligned} \mathrm{A}\quad & (x_3, y_3) + \mathrm{B}\,(x_3, y_3) \\ & = \int_{y_3}^{+\infty} \left[\int_{-\infty}^{x_3} \mathrm{f}_2\,(x_2, y_2)\; dx_2 \right. \\ & \quad \left. + \int_{x_3}^{\infty} \mathrm{f}_2\,(x_2, x_2 + y_2 - x_3)\; dx_2 \right] \mathrm{T1F}\,(x_3, y_2)\; dy_2. \end{aligned} \tag{A28}$$

Comparing (A28) and (A10) reveals that the line integral inside the square brackets in (A28) is equivalent to the negative of the partial derivative of T2F with respect to y_2.

Similarly, referring to Fig. 6(c), the third term in (A20) can be expressed as

$$\begin{aligned} \mathrm{C}\quad & (x_3, y_3) \\ & = \iint_{\mathrm{Subset3}} \mathrm{f}_2\,(x_2, y_2)\, \mathrm{p_c}\,(x_2, y_2, x_3, y_3)\; dx_2\, dy_2 \\ & = \int_{y_3}^{\infty} \int_{y_2 - y_3 + x_3}^{\infty} \mathrm{f}_2\,(x_2, y_2)\, \mathrm{T1F}\,(y_3 - y_2 + x_2, y_3)\; dx_2\, dy_2. \end{aligned} \tag{A29}$$

Applying Lemma 2 [given in (A3)] to change the limits of integration, (A29) becomes

$$\begin{aligned} \mathrm{C}\,(x_3, y_3) = \int_{x_3}^{\infty} & \left[\int_{y_3}^{+\infty} \mathrm{f}_2\,(y_2 - y_3 + x_2, y_2)\, dy_2 \right] \\ & \times \mathrm{T1F}\,(x_2, y_3)\; dx_2. \end{aligned} \tag{A30}$$

Inside the square brackets is the line integral of $\mathrm{f}_2(x, y)$ along the line of identity in Fig. 6(d), and function T1F is an area integration of $\mathrm{f}_1(x, y)$ over the shaded area. Comparing (A30) and (A7) reveals that the line integral inside the bracket in (A30) is the partial derivative of T2F with respect to x_2.

Substituting the partial derivatives of T2F given in (A7) and (A10), $\mathrm{p}(c = 1 | x_3, y_3)$ in (A20) is given by

$$\begin{aligned} \mathrm{p}\quad & (c{=}1 \quad | \quad x_3, y_3) \\ & = - \int_{y_3}^{+\infty} \frac{\partial \mathrm{T2F}(x_3, y_2)}{\partial y_2} \mathrm{T1F}\,(x_3, y_2)\; dy_2 \\ & \quad + \int_{x_3}^{\infty} \frac{\partial \mathrm{T2F}(x_2, y_2)}{\partial x_2} \mathrm{T1F}\,(x_2, y_3)\; dx_2. \end{aligned} \tag{A31}$$

Substituting (A31) into (A14), we obtain an expression for PC, i.e.,

$$\begin{aligned} PC \quad = & \\ & - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{f}_3\,(x_3, y_3) \\ & \times \int_{y_3}^{\infty} \frac{\partial \mathrm{T2F}(x_3, y_2)}{\partial y_2} \mathrm{T1F}\,(x_3, y_2)\; dy_2\, dx_3\, dy_3 \\ & + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{f}_3\,(x_3, y_3) \\ & \times \int_{x_3}^{\infty} \frac{\partial \mathrm{T2F}(x_2, y_2)}{\partial y_2} \mathrm{T1F}\,(x_2, y_3)\; dx_2\, dx_3\, dy_3. \end{aligned} \tag{A32}$$

Equation (A32) shows that the percent correct is a sum of two terms. For notational simplicity, we replace all the x_3 with x, y_3 with y, x_2 with x', and y_2 with y'. We first rewrite the first term as

$$- \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \mathrm{f}_3\,(x, y) \int_{y}^{\infty} \frac{\partial \mathrm{T2F}\left(x, y'\right)}{\partial y'} \mathrm{T1F}\left(x, y'\right)\; dy'\; dy \right] dx. \tag{A33}$$

Applying the Lemma 3 [given in (A4)] to the expression inside the square bracket in, (A33) becomes

$$-\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\frac{\partial \mathrm{T2F}(x,y')}{\partial y'}\mathrm{T1F}(x,y')\times\left(\int_{-\infty}^{y'}\mathrm{f}_3(x,y)\,dy\right)dy'\right]dx. \quad \text{(A34)}$$

Comparing (A11) and (A34) reveals that the term inside the parentheses is the partial derivative of T3F with respect to *x*. Substituting (A11) into (A34), gives

$$-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{T1F}(x,y')\frac{\partial \mathrm{T2F}(x,y')}{\partial y'}\frac{\partial \mathrm{T3F}(x,y')}{\partial x}dy'\,dx. \quad \text{(A35)}$$

Applying Lemma 3 and the expression for the partial derivative of T3F in (A11) to the second term of (A32) in a similar fashion gives

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{T1F}(x',y)\frac{\partial \mathrm{T2F}(x',y)}{\partial x'}\frac{\partial \mathrm{T3F}(x',y)}{\partial y}dx'\,dy. \quad \text{(A36)}$$

For notational simplicity, we rewrite (A35) and (A36) so that they are both integrals of *x* and *y*. Thus, (A32) becomes

$$PC=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{T1F}(x,y)\begin{vmatrix}\frac{\partial \mathrm{T2F}(x,y)}{\partial x} & \frac{\partial \mathrm{T2F}(x,y)}{\partial y}\\ \frac{\partial \mathrm{T3F}(x,y)}{\partial x} & \frac{\partial \mathrm{T3F}(x,y)}{\partial y}\end{vmatrix}dy\,dx. \quad \text{(A37)}$$

Inspection of (A37) reveals that the determinant is the Jacobian for the change of variables from (*T2F*, *T3F*) to (*x*, *y*). Noting that these fractions become 0 and 1, respectively, as the rating values move from -∞ to ∞, we thus have

$$PC=\int_0^1\int_0^1 T1F\,dT3F\,dT2F \quad \text{(A38)}$$

where *T1F*, *T2F*, and *T3F* are defined in (A1). Equation (A38) is the expression for VUS, and we thus have that

$$PC=\mathrm{VUS}. \quad \text{(A39)}$$

REFERENCES

1. He X, Metz CE, Links JM, Tsui BM, Frey EC. Three-class ROC analysis—A decision theoretic approach under the ideal observer framework. IEEE Trans. Med. Imag. May; 2006 25(5):571–581.
2. He X, Frey EC. Three-class ROC analysis—The equal error utility assumption and the optimality of three-class ROC surface/hypersurface with the ideal observer. IEEE Trans. Med. Imag. May; 2006 25(5):979–986.
3. He X, Frey EC. An optimal three-class linear observer derived from decision theory. IEEE Trans. Med. Imag. Jan.2007 26(1):77–83.
4. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143:29–36. [PubMed: 7063747]
5. Green, DM.; Swets, JA. Signal Detection Theory and Psychophysics. Krieger; Huntington, NY: 1973.
6. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J. Math. Psych. 1975; 12:387–415.

7. Barrett HH, Abbey CK, Clarkson E. Objective assessment of image quality. III. ROC metrics, ideal observers, and the likelihood-generating functions. J. Opt. Soc. Amer. 1998; 15:1520–1535.
8. Barrett, HH.; Myers, KJ. Foundations of Image Science. Wiley; New York: 2003.
9. Metz CE. Basic principles of ROC analysis. Sem. Nucl. Med. 1978; 8:283–298.
10. Metz CE. ROC methodology in radiologic imaging. Invest. Radiol. 1986; 21:720–733. [PubMed: 3095258]
11. Thompson ML, Zucchini W. On the statistical analysis of ROC curves. Statist. Med. 1989; 8:1277–1290.
12. Scurfield BK. Generalization of the theory of signal detectability to n-event m-dimensional forced-choice tasks. J. Math. Psych. 1998; 42:5–31.
13. 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]
14. Lehmann, EL. Nonparametrics: Statistical Methods Based on Ranks. McGraw-Hill; New York: 1975.
15. Pierre, C. M.S. thesis. Dept. Elect. Eng., Virginia Polytechnic Inst. State Univ.; Blacksburg: 1997. Confidence interval estimation for distribution systems power consumption by using the bootstrap method.
16. Efron, B.; Tsibshirani, RJ. An Introduction to the Bootstrap. Chapman & Hall; London, U.K.: 1993.
17. Song, X.; Frey, EC.; He, X.; Segars, WP.; Tsui, BMW. A mathematical observer study for evaluation of a model-based compensation method for crosstalk in simultaneous dual isotope SPECT. IEEE Med. Imag. Conf.; Portaland, OR. 2003.
18. He X, Frey EC, Links JM, Tsui BMW. Three-class ROC analysis and three-class hotelling observer for myocardial perfusion SPECT optimization and evaluation. J. Nucl. Med. 2004; 45:42P–42p.
19. Swets, J.; Pickett, R. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. Academic; New York: 1982.
20. Swets JA, Birdsall TG. The human use of information. 3. decision-making in signal-detection and recognition situations involving multiple alternatives. IRE Trans. Inf. Theory. 1956; 2:138–165.
21. Scurfield BK. Multiple-event forced-choice tasks in the theory of signal detectability. J. Math. Psychol. 1996; 40:253–269. [PubMed: 8979976]
22. 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]
23. Edwards DC, Metz CE. Restrictions on the three-class ideal observer's decision boundary lines. IEEE Trans. Med. Imag. Dec.2005 24(12):1566–1573.
24. 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.
25. Edwards DC, Metz CE, Kupinski MA. Ideal observers and optimal ROC hypersurfaces in n-class classification. IEEE Trans. Med. Imag. Jul.2004 23(7):891–895.
26. 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. Mar.2005 24(3):293–299.
27. Chan H-P, Sahiner B, Hadjiiski LM, Petrick N, Zhou C. Design of three-class classifiers in computer-aided diagnosis: Monte carlo simulation study. Proc. SPIE. 2003; 5032:567–578.
28. Mossman D. Three-way ROCs. Med. Decision Making. 1999; 19:78–89.
29. Nakas CT, Yiannoutsos CT. Ordered multiple-class ROC analysis with continuous measurements. Statist. Med. 2004; 23:3437–3449.
30. Xiong C, van Belle G, Miller JP, Morris JC. Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. Statist. Med. 2006; 25(7):1251–1273.
31. Sahiner B, Chan H-P, Hadjiiski LM. Performance analysis of 3-class classifiers: Properties of the 3-D ROC surface and the normalized volume under the surface. Proc. SPIE. 2006; 6146

32. Sahiner B, ChanL HP, Hadjilski M. Performance analysis of three-class classifers: Properties of a 3–D ROC surface and the normalized volume under the surface for the ideal observer. IEEE Trans. Med. Imag. Feb.2008 27(2):215–227.

33. He X, Frey EC. Describing three-class task performance: Three-Class volume under ROC surface (VUS) and three-class hotelling trace (3-HT) as figures of merit. Proc. SPIE: Image Perception, Observer Performance, Technol. Assessment. 2007; 6515

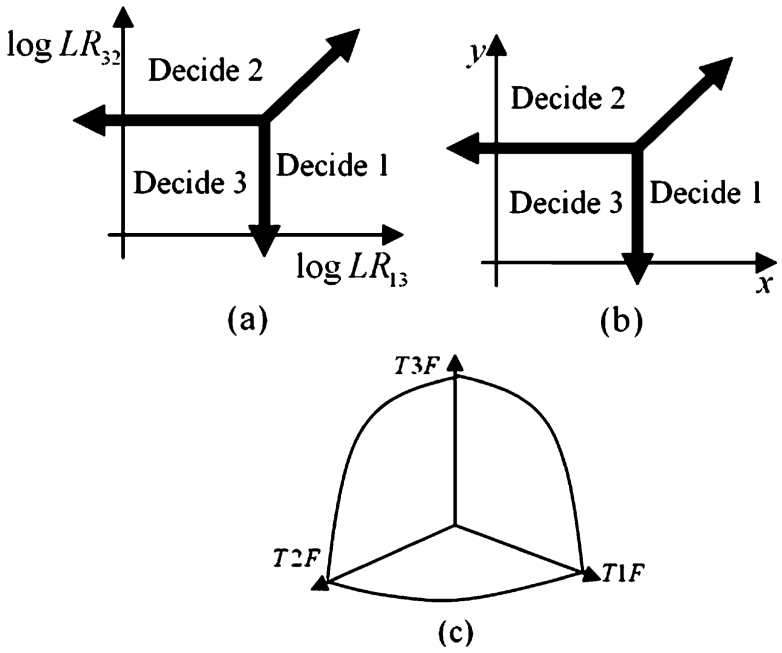

Fig. 1.

Decision plane of the three-class decision model and a three-class ROC surface. (a) Decision plane spanned by (log LR_{13}, log LR_{23}). (b) Decision plane spanned by (x, y). (c) Example of three-class ROC surface.

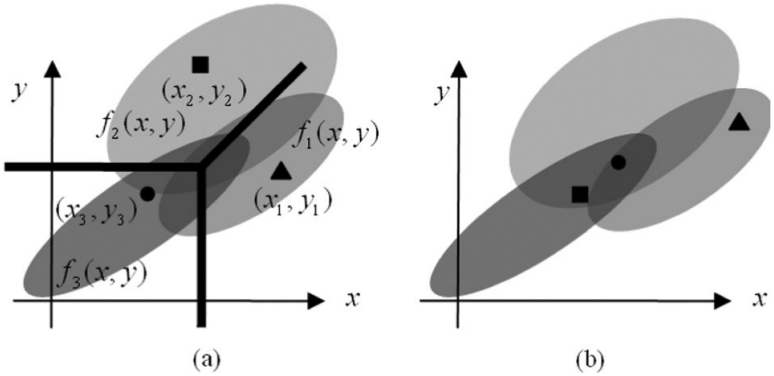

Fig. 2.

Illustration of a mathematical treatment of the categorization procedure that is equivalent to a rating procedure. Ellipses with different shadings schematically represent the distributions of decision variables for the three classes. Triangle is a random sample from class 1, the square is a random sample from class 2, and the disk represents a random sample from class 3. (a) Illustration of a case where the triplet can be correctly categorized. (b) Illustration of a case where correct categorization is not possible for any position of the proposed decision structure.

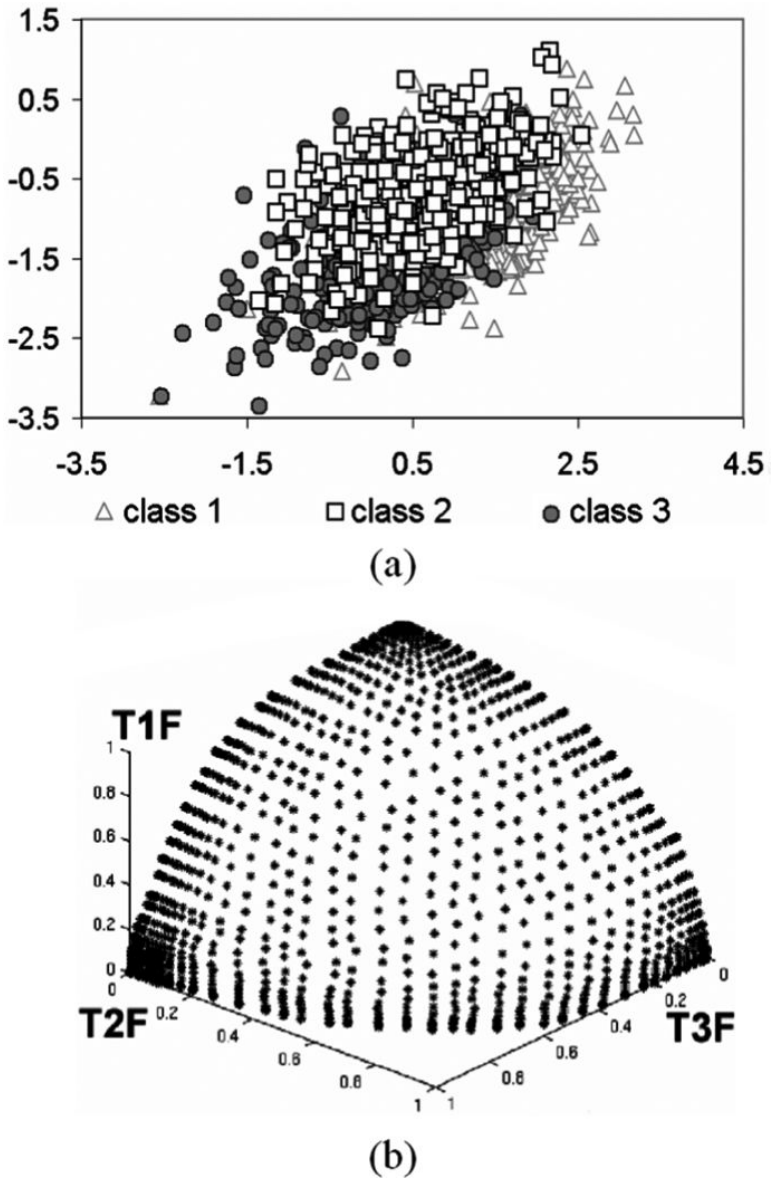

Fig. 3.

Decision planes and the ROC surface obtained for simulated dual-isotope MPS images. (a) Decision plane. (b) ROC surface traced out by moving the decision structure on the decision plane and compute the sensitivity triplets.

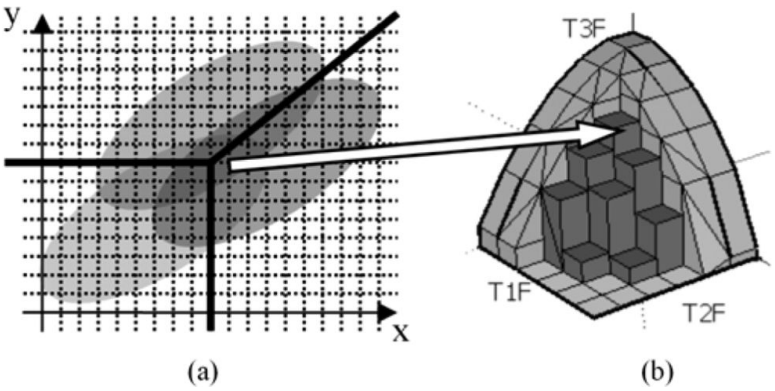

Fig. 4.
Calculation of the VUS. (a) Sample the distributions with very small grid spacing in the decision plane. (b) For each (T1F, T2F) bin in the ROC space, we average the T3F values to produce an ROC histogram.

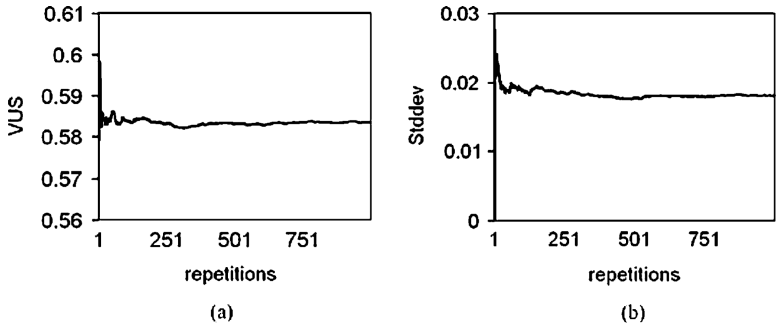

Fig. 5.
Convergence of VUS (a) and its standard deviation (b) using the Bootstrap method.

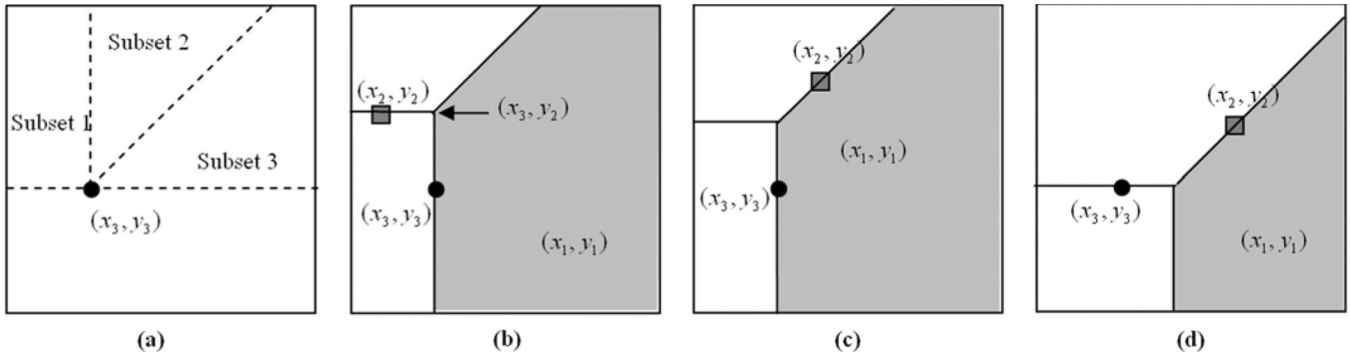

Fig. 6.

For a particular (x_3, y_3), (a)–(d) illustrates the strategies for finding correctly classifiable triplet. (a) We divide all (x_2, y_2) that may form a correct classification with (x_3, y_3), i.e., $y_2 > y_3$, into three mutually exclusive subsets, Subset 1, Subset 2, and Subset 3, respectively. Here, (b)–(d) shows examples of the set of (x_1, y_1), indicated by the shaded region, which will result in correct classification for the particular (x_3, y_3) and a (x_2, y_2) from each subset, respectively.

TABLE I

Parameters of the Fitted Bivariate Gaussian Distributions for the Three Classes

	μ_x	σ_x	μ_y	σ_y	ρ
Class 1	1.340	0.643	−0.878	0.386	0.532
Class 2	0.688	0.520	−0.642	0.328	0.492
Class 3	0.187	0.557	−1.358	0.390	0.528

TABLE II

Resulting VUS Values and the Standard Deviations Using Different Approaches

	***VUS* = *PC* method**			**Ad hoc method**
	Bootstrap		**Eqn. (6)-(8)**	
	1000 repetitions	**200 repetitions**		
VUS	0.5835	0.5835	0.5830	0.5815
Standard deviation	0.0181	0.0189	0.0182	N/A
Computational time	75min	15min	~8 days	N/A

TABLE III

Definition of the Three Subsets

Subset 1	$\{x_2 < x_3, y_2 > y_3\}$
Subset 2	$\{x_2 \geq x_3, y_2 > y_3, y_2 - x_2 \geq y_3 - x_3\}$
Subset 3	$\{x_2 \geq x_3, y_2 > y_3, y_2 - x_2 < y_3 - x_3\}$