Comparing Three-class Diagnostic Tests by Three-way ROC Analysis

Stephan Dreiseitl, Lucila Ohno-Machado and Michael Binder

Med Decis Making 2000; 20; 323
DOI: 10.1177/0272989X0002000309

Comparing Three-class Diagnostic Tests by Three-way ROC Analysis

STEPHAN DREISEITL, PhD, LUCILA OHNO-MACHADO, MD, PhD, MICHAEL BINDER, MD

Three-way ROC surfaces are based on a generalization of dichotomous ROC analysis to three-class diagnostic tests. The discriminatory power of three-class diagnostic tests is measured by the volume under the ROC surface. This measure can be given a probabilistic interpretation similar to the equivalence of the c-index to the area under the ROC curve. This article presents a method to calculate nonparametric estimates of the variance of the volume under the surface using Mann–Whitney U statistics. As a simple extension of this result, it is possible to calculate covariance estimates for the volume under the surface. This allows the statistical comparison of two tests used for diagnostic tasks with three possible outcomes. The formulas derived are validated on synthetic data and applied to a three-class data set of pigmented skin lesions. It is shown that a neural network algorithm trained on clinical data and lesion features performs better than one trained on only the lesion features. *Key words*: Receiver operating characteristic curves; trichotomous ROC analysis. **(Med Decis Making 2000; 20:323–331)**

Recently, Mossman[1] extended the dichotomous receiver operating characteristic (ROC) curve analysis to trichotomous diagnostic tasks. A trichotomous diagnostic task is the task of classifying a case as belonging to one of three possible classes. Generally, the classification is based on the outcome of a three-class diagnostic test.

In the dichotomous case, ROC curves are used to summarize the discriminatory performance of a test or rater by plotting sensitivity (true-positive rate) versus 1 − specificity (false-positive rate) across a spectrum of decision thresholds.[2] Because it is equivalent to the Mann-Whitney U-statistic, the area under the curve (AUC) is a measure of the discriminatory power of the test.[3] The equivalence to U-statistics allows to draw on the statistical literature to establish standard deviations and asymptotic normality of area measurements.[4,5]

Given these results, it is possible to statistically determine whether one diagnostic test is better than another.[6–8] A major research focus is the calculation of correlations in AUC estimates when two tests are applied to the same subjects.[9–12] Accounting for correlations increases a statistical test's power, making it easier to detect significant differences in AUC estimates.

In this article, we first briefly introduce three-way ROC curves, and point out the similarities to the dichotomous case. We then derive a nonparametric estimate of variance for the volume under the surface (VUS), the trichotomous analog to the AUC, by using the equivalence to U-statistics. The resulting formula is validated by experiments using normally distributed data. In order to compare three-class diagnostic tests, we show how to calculate nonparametric estimates of VUS correlations. We use this result to statistically test the hypothesis that a machine-learning algorithm trained on a data set including clinical information performs significantly better than one trained without this information.

Three-way ROCs

The work of Mossman[1] provides a clear presentation of three-way ROCs. For brevity, we summarize only those points relevant to the present discussion.

Generally, a test result or a case rating provides an assessment indicating to which of a number of classes a subject belongs. In the dichotomous case, there are only two classes, and the test result determines the location of a subject's status between these two classes. If the two classes are represented, as they usually are, by 0 for a negative and 1 for a positive condition, a test result's location x, $0 \leq x \leq$

Received May 28, 1999 from the Decision Systems Group, Brigham and Women's Hospital, Division of Health Sciences and Technology, Harvard Medical School, Massachusetts Institute of Technology, Boston, Massachusetts. Revision accepted for publication January 19, 2000. Supported in part by grant J 1661-INF from the Austrian FWF (SD), by the Max Kade Foundation (MB), by NLM/NHLBI contract 467-MZ-802289, and by NLM grant R29 LM06538-01 (LOM).

Address correspondence and reprint requests to Dr. Dreiseitl: Decision Systems Group, Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02115; e-mail: ⟨sdreisei@dsg.harvard.edu⟩.

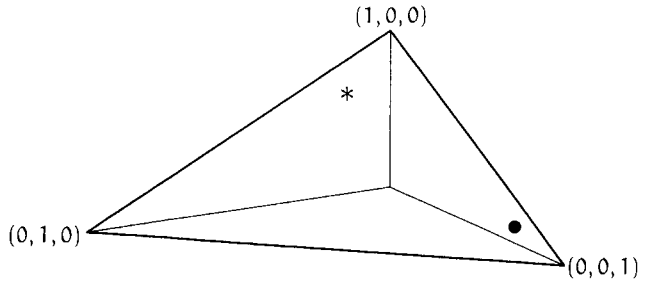

FIGURE 1. The three-dimensional estimate triangle that bounds all probabilities of a three-class diagnostic test outcome. The two points shown are * = (0.7, 0.2, 0.1) and • = (0.15, 0.05, 0.8).

1 gives a numerical estimate of the probability that the subject is positive. Consequently, $1 - x$ is the probability that the subject is negative.

In the trichotomous case, a test rates a subject's condition with regard to three possible classes $C = 1, 2, 3$, i.e., the test calculates three class-membership probabilities $P(C = 1)$, $P(C = 2)$, $P(C = 3)$. It is clear that these three classes cannot be represented in one dimension (as, e.g., 0, 1, and 2), because a test result's placement in one dimension could not provide the probability interpretation of the dichotomous case. It is however, possible to place the three classes in two dimensions, as corners of a equilateral triangle. In this case, the coordinates of a point inside the triangle determine the probability that a subject belongs to class 1, 2, or 3, respectively. Only two coordinates are required, as the third is determined by the condition that the probabilities must add to one. Nevertheless, it is easier to encode the corners of the triangle as (1, 0, 0), (0, 1, 0), and (0, 0, 1), respectively, so that the triangle connects three corners of the unit cube in three-dimensional space. Then, each three-dimensional coordinate can be directly interpreted as a class-membership probability triple. We identify corner (1, 0, 0) with class 1, (0, 1, 0) with class 2, and (0, 0, 1) with class 3. Note that since the probabilities have to add to one, the "estimate triangle" lies on a two-dimensional plane in three-dimensional space; see figure 1 for an illustration.

For a two-class diagnostic test, the points on the ROC curve are obtained by calculating test sensitivities and specificities at varying decision thresholds. Since "positive" (T^+) and "negative (T^-) are the only two outcomes for both disease-positive (D^+) and disease-negative (D^-) subjects, we know that $P(T^-|D^+) = 1 - P(T^+|D^+)$ and $P(T^+|D^-) = 1 - P(T^-|D^-)$. This means that there is exactly one alternative to sensitivity and specificity, and it makes sense to plot sensitivity versus 1 − specificity. This results in an ROC curve that runs through the points (0, 0) and (1, 1), tending towards (0, 1) for tests with increasingly good discriminatory performance. It is equally plausible, however, to plot sensitivity versus specificity. The resulting plot extends from (0, 1) to (1, 0), approaching (1, 1) for good tests. The area under this curve is the same as the area under a regular ROC curve.

For a three-class diagnostic test, a subject can belong to one of three disease classes ($D = 1, 2, 3$), and will be rated as belonging to one of three test-outcomes classes $C = 1, 2, 3$. There are two alternatives to each "true-class rate" $P(C = k|D = k)$, $k = 1, 2, 3$, so that it is possible to plot only $P(C = 1|D = 1)$ versus $P(C = 2|D = 2)$ versus $P(C = 3|D = 3)$. This is the trichotomous version of plotting sensitivity versus specificity for two-class diagnostic tests.

By varying decision criteria of how to assign estimate triples to classes, one can calculate several true-class rate triples and plot them in three-dimensional space, forming an ROC surface, the trichotomous analog to an ROC curve. The procedure for doing so is somewhat elaborate, and is not repeated here. The details can be found in Mossman's article.[1] An example of an ROC surface is shown in figure 2. The points (1, 0, 0), (0, 1, 0), (0, 0, 1) are on every ROC surface; connecting them with straight lines results in the surface corresponding to a test that cannot discriminate between the three classes. The VUS of such a test is 1/6.

Having established the analogy between ROC curves and ROC surfaces, it is interesting to consider how to interpret the volume under the ROC surface. The area under the ROC curve is equivalent to the probability that a randomly chosen D^+ subject will be rated higher than a randomly chosen D^- subject. Mossman established that, similarly, the volume under the ROC surface is equivalent to the probability that three chosen subjects, one each from classes 1, 2, and 3, will be rated correctly. This begs the question of what it means to rate three subjects correctly, given only their estimate triples. Mossman proposes two rules; in this work we use the following: Three estimate triples $p_1 = (p_{11}, p_{12}, p_{13})$, $p_2 = (p_{21}, p_{22}, p_{23})$,

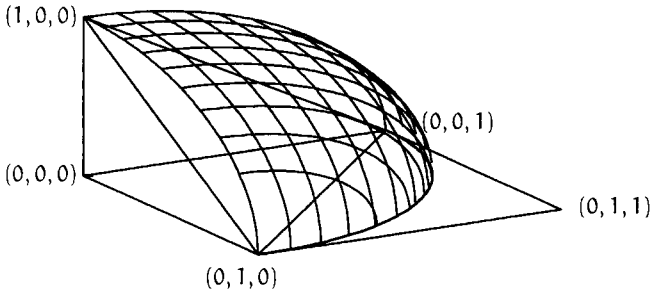

FIGURE 2. An ROC surface. The three axes are the true-class rates $P(C = k|D = k)$, $k = 1, 2, 3$. A point on the surface represents a contingency table with true-class rates given by the point's coordinates. Shown underneath the surface is the tetrahedron representing an uninformative ROC surface.

and $p_3 = (p_{31}, p_{32}, p_{33})$, from subjects in classes 1, 2, and 3, respectively, are correctly rated if the sum of the lengths of the lines joining each rating triple to the corner associated with its class is smaller than that of any other length-of-line combination joining the three triples to the three triangle corners. An example of a correctly ordered triple is shown in figure 3.

We define the function

$$cr(p_1, p_2, p_3) = \begin{cases} 1 & \text{if } (p_1, p_2, p_3) \text{ correctly rated} \\ 0 & \text{else} \end{cases}$$

to denote correctly rated triples. Note that for a non-discriminatory test, the rating triples p_1, p_2, and p_3 are randomly distributed in the estimate triangle. The chances of correctly rating these triples is 1 over 3! = 6, the number of all possible (equally likely) ways of connecting three points to three corners. As expected by the equivalence to the volume under the ROC surface, this probability is the same as the VUS for a non-discriminatory test.

Standard Deviation of the Volume under the Surface

For two-class diagnostic tests, it is possible to calculate standard deviations for the area under the ROC curve by using the equivalence of that measure to Mann-Whitney U-statistics. In this section, we use a similar equivalence that holds for the volume under an ROC surface to calculate standard deviations for this measure. This work proceeds similarly to the derivation of the standard deviation formula that can be found in Hanley and McNeil's article.[7]

It was already established by Mossman that the volume under the surface is equivalent to the probability of correctly rating three subjects, one from each class. For the following derivation, let X_i, $i = 1, \ldots, m$ be the estimate triples for subjects from class 1, Y_j, $j = 1, \ldots, n$ the estimate triples for subjects from class 2, and Z_k, $k = 1, \ldots, l$ the estimate triples for subjects from class 3. These triples are assumed to be independent and, within each class, identically distributed. An unbiased estimator of $\theta = P[cr(X, Y, Z) = 1]$, the probability of rating three estimate triples X, Y, Z from different classes correctly is given by

$$\hat{\theta} = W = \frac{1}{mnl} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{l} cr(X_i, Y_j, Z_k)$$

Thus, W gives the fraction of all possible three-subject combinations that are rated correctly. In much the same way as for two-class diagnostic tests, it is possible to calculate the variance of W (see the appendix for details) as

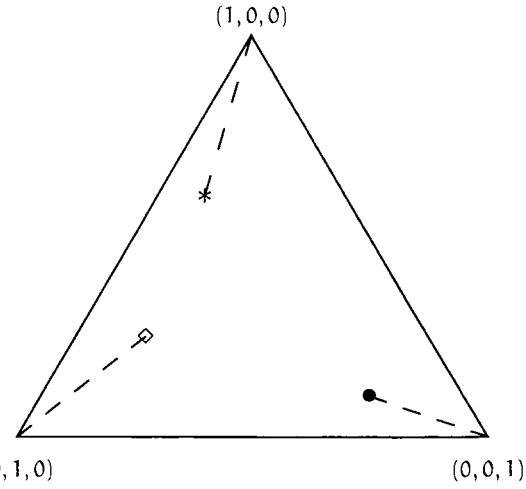

FIGURE 3. A correctly ordered estimate triple. The point * is in class 1, ◊ is in class 2, and • is in class 3. The sum of the dashed lines connecting the points to their corners is shorter than those of all other connections of these points to the corners.

$$\begin{aligned} \mathrm{Var}(W) = \frac{1}{mnl} [&\theta(1 - \theta) + (l - 1)(q_{12} - \theta^2) \\ &+ (n - 1)(q_{13} - \theta^2) + (m - 1)(q_{23} - \theta^2) \\ &+ (n - 1)(l - 1)(q_1 - \theta^2) \\ &+ (m - 1)(l - 1)(q_2 - \theta^2) \\ &+ (m - 1)(n - 1)(q_3 - \theta^2)] \end{aligned} \quad (1)$$

where the new symbols are defined as follows:

$$q_{12} = P[cr(X_i, Y_j, Z_k) = cr(X_i, Y_j, Z_K) = 1], \quad K \neq k$$

the probability of correctly rating three subjects X_i, Y_j, Z_k, and correctly rating X_i, Y_j, and a different class 3 subject Z_K,

$$q_{13} = P[cr(X_i, Y_j, Z_k) = cr(X_i, Y_J, Z_k) = 1], \quad J \neq j$$

$$q_{23} = P[cr(X_i, Y_j, Z_k) = cr(X_I, Y_j, Z_k) = 1], \quad I \neq i$$

$$q_1 = P[cr(X_i, Y_j, Z_k) = cr(X_i, Y_J, Z_K) = 1], \quad J \neq j, \quad K \neq k$$

$$q_2 = P[cr(X_i, Y_j, Z_k) = cr(X_I, Y_j, Z_K) = 1], \quad I \neq i, \quad K \neq k$$

$$q_3 = P[cr(X_i, Y_j, Z_k) = cr(X_I, Y_J, Z_k) = 1], \quad I \neq i, \quad J \neq j$$

The latter symbols have interpretations similar to q_{12}, with varying combinations of same and changing elements from the three classes. Estimates for these quantities can be obtained by counting the fraction of triples combinations for which the defining relation holds, e.g.,

Table 1 • Results of Comparing Estimated and Observed Standard Deviations for Nine Different Combinations of Sample Sizes (20, 50, 80) and Dispersions (0.4, 0.6, 0.8)*

	Sample Size 20			Sample Size 50			Sample Size 80		
	$\bar{\theta}$	σ_{100}	σ_{est}	$\bar{\theta}$	σ_{100}	σ_{est}	$\bar{\theta}$	σ_{100}	σ_{est}
Sample σ 0.4	0.880	0.0445	0.0467	0.880	0.0321	0.0307	0.882	0.0260	0.0244
Sample σ 0.6	0.533	0.0800	0.0805	0.536	0.0530	0.0528	0.532	0.0434	0.0421
Sample σ 0.8	0.356	0.0734	0.0744	0.362	0.0484	0.0493	0.352	0.0389	0.0390

*The values given for each combination are the average (over 100 runs) VUS estimate $\bar{\theta}$, the standard deviation σ_{100} of the 100 estimates, and the estimated standard deviations σ_{est}.

$$\hat{q}_{12} = \frac{1}{mnl(l-1)} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{l} \sum_{K=1, K \neq k}^{l} cr(X_i, Y_j, Z_k) cr(X_i, Y_j, Z_K)$$

An estimator $\hat{\sigma}$ of the standard deviation σ is then obtained by using estimators $\hat{q}_{12}, \ldots, \hat{q}_3$ for $q_{12}, \ldots, q_3$ in equation 1, and taking the square root.

To verify these calculations numerically, we performed a series of experiments with synthetic data. We generated 20, 50, and 80 estimate triples for each of the three classes from two-dimensional normal distributions centered at the corners of the estimate triangle, with the same degree of dispersion for each class. We used diagonal covariance matrices with constant σ values of 0.4, 0.6, and 0.8 for a total of nine combinations of sample sizes and dispersions. The results of running 100 calculations with each of the nine combinations are given in table 1. This table shows that there is good agreement between the estimated and measured standard deviations; the differences are on the order of 2/1,000 and can be explained by variance in the samples.

Several other properties of the VUS estimate can be obtained from the simulation measurements table. These observations were first given by Hanley and McNeil[7] for the AUC variance estimate, but hold for the VUS estimate as well: First, the variance decreases with increasing sample sizes, when holding the sample dispersion constant (seen in each row of table 1). Second, when increasing sample dispersions, i.e., when decreasing VUS, variance estimates increase (seen in the columns of table 1). Finally, variances are inversely proportional to the sample size, so that a fourfold increase in sample size decreases the variance by four, and the standard deviation by two.

Comparing Volumes under Surfaces

A further important consequence of the equivalence of VUS and Mann–Whitney U-statistics is asymptotic normality.[4,13] Asymptotic normality can be used to test the hypothesis that two VUS values are different. As pointed out by Hanley and McNeil[9] for the dichotomous case, it is advantageous to take possible correlations between volume estimates into account to increase the power of the test, i.e., to increase the ability to detect a difference in volume when it exists.

To be more precise, let X_i^1, $i = 1, \ldots, m$ be the rater 1 estimate triples for subjects from class 1, X_i^2, $i = 1, \ldots, m$, the same class 1 subjects rated by rater 2; Y_j^1, $j = 1, \ldots, n$ the rater 1 estimate triples for subjects from class 2, Y_j^2, $j = 1, \ldots, n$, the same class 2 subjects rated by rater 2; Z_k^1, $k = 1, \ldots, l$ the rater 1 estimates for class 3 subjects, and Z_k^2, $k = 1, \ldots, l$, the rater 2 estimates for the same class 3 subjects. Furthermore, let

$$W_1 = \frac{1}{mnl} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{l} cr(X_i^1, Y_j^1, Z_k^1)$$

$$W_2 = \frac{1}{mnl} \sum_{I=1}^{m} \sum_{J=1}^{n} \sum_{K=1}^{l} cr(X_I^2, Y_J^2, Z_K^2)$$

be two Mann-Whitney U-statistics based on the estimates of rater 1 and rater 2. Asymptotic normality implies that for large sample sizes, W_1 and W_2 will have normal distributions with parameters θ_1, σ_1^2 and θ_2, σ_2^2, respectively. The difference $W_1 - W_2$ is then normally distributed with mean $\theta_1 - \theta_2$ and variance $\sigma_1^2 + \sigma_2^2 - 2\,\mathrm{Cov}(W_1, W_2)$. When W_1 and W_2 are obtained from estimate triples of the same subjects (as is the case here), they will not be independent and the term $\mathrm{Cov}(W_1, W_2)$ cannot be dropped from the calculations.

A test for statistical significance of difference is then to calculate

$$z = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{r}\hat{\sigma}_1\hat{\sigma}_2}} \qquad (2)$$

and to determine whether this value lies outside the range that can be attributed to chance. In equation 2, $\hat{r}$ denotes an estimator for the correlation between W_1 and W_2, i.e., for

$$r = \frac{\mathrm{Cov}(W_1, W_2)}{\sigma_1 \sigma_2} \qquad (3)$$

The covariance of two volume measurements can be calculated similarly to their variances; the resulting expression is

$$\begin{aligned}\mathrm{Cov}(W_1, W_2) = \frac{1}{mnl}\,[&cq_{123} - \theta_1\theta_2 \\ &+ (l-1)(cq_{12} - \theta_1\theta_2) + (n-1)(cq_{13} - \theta_1\theta_2) \\ &+ (m-1)(cq_{23} - \theta_1\theta_2) \\ &+ (n-1)(l-1)(cq_1 - \theta_1\theta_2) \\ &+ (m-1)(l-1)(cq_2 - \theta_1\theta_2) \\ &+ (m-1)(n-1)(cq_3 - \theta_1\theta_2)] \end{aligned} \quad (4)$$

The definitions of the new symbols are

$$cq_{123} = P(cr_1(X_i, Y_j, Z_k) = cr_2(X_i, Y_j, Z_k) = 1)$$

The probability that both rater 1 and rater 2 correctly rate three subjects X_i, Y_j, Z_k,

$$cq_{12} = P[cr_1(X_i, Y_j, Z_k) = cr_2(X_i, Y_j, Z_K) = 1], \quad K \neq k$$

the probability that rater 1 correctly rates X_i, Y_j, Z_k, and that rater 2 correctly rates X_i, Y_j, and a different class 3 subject Z_K,

$$cq_{13} = P[cr_1(X_i, Y_j, Z_k) = cr_2(X_i, Y_J, Z_k) = 1], \quad J \neq j$$

$$cq_{23} = P[cr_1(X_I, Y_j, Z_k) = cr_2(X_I, Y_j, Z_k) = 1], \quad I \neq i$$

$$cq_1 = P[cr_1(X_i, Y_j, Z_k) = cr_2(X_i, Y_J, Z_K) = 1], \quad J \neq j, \quad K \neq k$$

$$cq_2 = P[cr_1(X_i, Y_j, Z_k) = cr_2(X_I, Y_j, Z_K) = 1], \quad I \neq i, \quad K \neq k$$

$$cq_3 = P[cr_1(X_I, Y_j, Z_k) = cr_2(X_I, Y_J, Z_k) = 1], \quad I \neq i, \quad J \neq j$$

The interpretations of the last symbols are similar to those of the first two. More information about the derivation of equation 4 can be found in the appendix.

As before, we can give estimators of these quantities by simply counting the number of combinations that satisfy the definitions; one example is

$$\hat{cq}_{12} = \frac{1}{mnl(l-1)} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{l} \sum_{\substack{K=1 \\ K \neq k}}^{l} cr(X_i^1, Y_j^1, Z_k^1) cr(X_i^2, Y_j^2, Z_K^2)$$

We can calculate $\hat{r}$, the estimator for the correlation coefficient, by using the estimators for the quantities $cq_{123}, \theta_1, \ldots, cq_3$ in equation 4 to obtain an estimator for $\mathrm{Cov}(W_1, W_2)$. Using equation 3, we then get an estimate for the correlation coefficient.

Clinical Application

As an application of the method derived above, we tested whether inclusion of clinical information helps to diagnose the malignancy of pigmented skin lesions. We analyzed a dataset containing 518 digital images of pigmented skin lesions that fell into three categories: 207 benign, common nevi, 195 morphologically atypical (dysplastic) nevi, and 116 cutaneous melanomas. Images were taken using the epiluminescence microscopy technique.[14,15] Digital image analysis was performed and 45 morphologic features, such as area, perimeter, shape factors, and color distributions, were extracted. Histopathologic findings for all lesions were used as the "gold standard" of truth. Seven pieces of clinical information were collected for each lesion: personal and family history of melanoma, the frequency of common nevi and the frequency of atypical nevi of the patient, degree of sun damage, skin type, and information about morphologic changes of the lesion observed and provided by the patient.

Two neural networks were constructed. One of the networks used only the features extracted from the images, while the other used the features and the clinical information. Both were trained on 260 images that were chosen randomly from the data set in such a way as to contain 100 instances each of common and dysplastic nevi, and 60 cutaneous melanomas. The remaining 258 images were taken as a test set to measure the performance of the trained network.

The networks were trained by using Markov-chain Monte Carlo methods to sample from the posterior distribution of parameters in a neural network model, using the algorithms and software developed by Neal.[16] This method incorporates automatic relevance determination (ARD) of inputs, so that it is possible to include all 45 features in the first model, and the 45 features and seven clinical information items in the second model. ARD automatically adjusts the contribution of inputs that are not relevant for calculating the output, so that good performances can be achieved even in the presence of inputs that do not influence the output. In contrast to standard neural network training such as backpropagation, this approach does not search for one set of network weights that minimizes an error function, but averages over several weight sets obtained by Monte Carlo sampling. We used 20 hidden neurons in the network and averaged over 2,000 sets of network weights.

To show that inclusion of clinical information has a significant influence on the ability to discriminate between the three lesion types, we set up the hypothesis H_0:$\theta_1 = \theta_2$ that there is no difference in VUS values. Here, θ_1 is the VUS of the test using only the

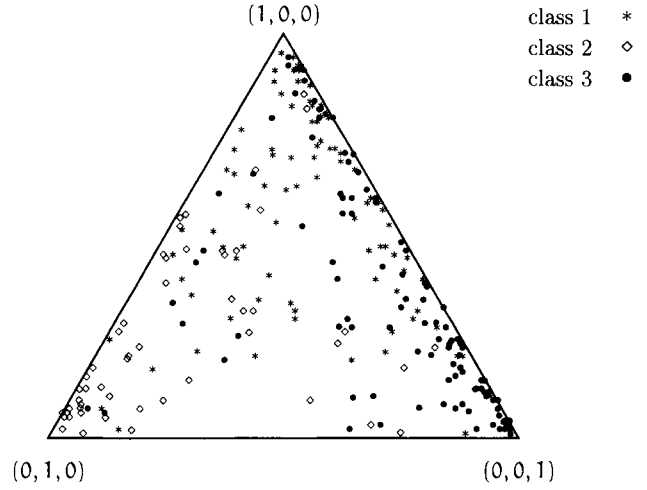

FIGURE 4. The distribution of the estimate triples on the triangle. Class 1 is associated with corner (1, 0, 0), class 2 with (0, 1, 0), class 3 with (0, 0, 1).

extracted image features, and θ_2 is the VUS of the test including clinical information. Our goal is then to discard this hypothesis based on the data.

The results obtained from the neural network algorithm give values of $\hat{\theta}_1 = 0.556$, $\hat{\sigma}_1 = 0.041$, $\hat{\theta}_2 = 0.649$, $\hat{\sigma}_2 = 0.038$, and $\hat{r} = 0.555$. Using equation 2, we obtain a z-value of $z = 2.49$. We can therefore discard the hypothesis that inclusion of clinical information does not help in diagnosis with a type I error probability of $p < 0.007$. Five different splits of training and tests sets produced similar results, with z-values between 2.21 and 3.10 (p-values between 0.014 and 0.001). The distribution of the estimate triples, as rated by one of the models including clinical information, is shown in figure 4. It is interesting to calculate the discrimination between any two of the three classes, using two-class AUC values based on the projections of the estimate triples onto the edges of the estimate triangle. Class 2 can be distinguished rather well from classes 1 and 3 (AUCs of 0.895 and 0.929, respectively), but it is not as easy to discriminate between classes 1 and 3 (AUC value 0.768).

Discussion

Two-class diagnostic tests are the tools of choice for classifying subjects as either normal or abnormal. In many cases, however, abnormality is not a one-sided alternative to normality, but actually two-sided. If normality of a condition is defined as an interval in the middle of a possible range of values, then there are actually two possible abnormalities: a value might lie below the normality range, or it might lie above it (corresponding, e.g., to thyroid hypofunction or hyperfunction). With three-way ROCs, it is possible to analyze tests for two distinct alternatives to abnormality; this is not possible with traditional dichotomous ROC analysis. As the example in the previous section shows, three-way ROC analysis is applicable to any test that distinguishes between three disease classes, not just those that can be ordered linearly.

In order to interpret results obtained from three-way ROC analysis, it is interesting to note what constitutes a "good" VUS value. We already know that the VUS of an uninformative three-class diagnostic test is 1/6; for an uninformative two-class diagnostic test, the AUC is 1/2. So what VUS value corresponds to an AUC value of, for example, 0.8? To answer this question, we consider dichotomous ROC curves that change continuously from the uninformative diagonal to more informative curves approaching the point (0, 1). For ROC surfaces, we know that the uninformative tetrahedron is bounded by the three coordinate axes and three lines connecting the points (1, 0, 0), (0, 1, 0), and (0, 0, 1). If we continuously change the three non-axes edges of the uninformative tetrahedron in the same way as dichotomous ROC curves, we get an ROC surface that tends towards the point (1, 1, 1). Figure 2 shows the uninformative tetrahedron and a more informative symmetric ROC surface with the same curves on all three boundary lines. Notice that the VUS grows slower than the AUC, since the higher dimensions provide for (relatively) more space in the unit cube that is not under the surface. The exact correspondence is given by the graph in figure 5. We can see that an AUC value of 0.8 corresponds to a VUS value of about 0.55. In our example, the neural network using clinical information for classifying pigmented skin lesions performed rather well, since its VUS value of 0.65 correspond to an AUC value of about 0.85.

We hope that this article can serve as a first step towards establishing a statistical framework for three-way ROC analysis. The results derived so far are applicable only to continuous distributions,

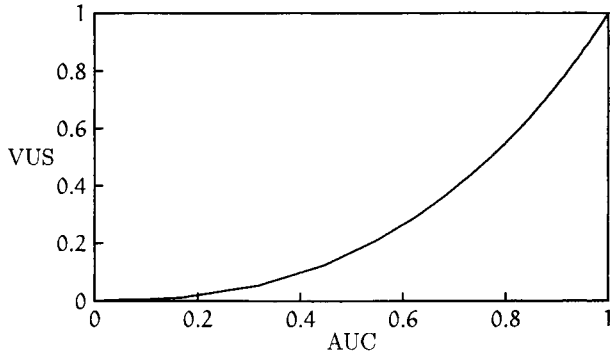

FIGURE 5. A plot of VUS against AUC values. Each point on the curve corresponds to a (AUC, VUS) pair obtained from using identical ROC curves as border lines of the ROC surface.

since they do not deal with the ties that would be obtained from discrete distributions. Furthermore, no distributional assumptions were made during the derivation of the formulas. Fitting appropriate distributions to the estimate data might lead to considerable simplifications of the calculations. Further research will be needed to alleviate these shortcomings: we will need to consider discrete data, and to incorporate distributional assumptions into the calculations.

The software used in the calculations is available for download from ftp://dsg.harvard.edu/pub/ThreewayROC.

References

1. Mossman D. Three-way ROCs. Med Decis Making. 1999;19: 78–89.
2. Swets JA, Pickett RM. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. New York: Academic Press, 1982.
3. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psychol. 1975;12:387–415.
4. Hoeffding W. A class of statistics with asymptotically normal distribution. Annals of Mathematical Statistics. 1948;19:293–325.
5. Lehmann EL. Consistency and unbiasedness of certain nonparametric tests. Annals of Mathematical Statistics. 1951;22: 165–79.
6. Metz CE, Kronman HB. Statistical significance tests for binormal ROC curves. J Math Psychol. 1980;22:218–43.
7. Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. Radiology. 1982;143:29–36.
8. Hanley JA, Hajian-Tilaki KO. Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update. Acad Radiol. 1997;4:49–58.
9. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983;148:839–43.
10. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837–45.
11. Swaving M, VanHouwelingen H, Ottes FP, Seerneman T. Statistical comparison of ROC curves from multiple readers. Med Decis Making. 1996;16:143–52.
12. Metz CE, Herman BA, Roe CA. Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets. Med Decis Making. 1998;18:110–21.
13. Randles RH, Wolfe DA. Introduction to the Theory of Nonparametric Statistics. Malabar, FL: Krieger, 1979.
14. Pehamberger H, Binder M, Steiner A, Wolff K. In vivo epiluminescence microscopy: improvement of early diagnosis of melanoma. J Invest Dermatol. 1993;100:356–62.
15. Binder M, Schwarz M, Winkler A, et al. Epiluminescence microscopy. A useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists. Arch Dermatol. 1995;131:286–91.
16. Neal RM. Bayesian Learning for Neural Networks. New York: Springer, 1996.
17. Lehmann EL. Nonparametrics: Statistical Methods Based on Ranks. New York: McGraw-Hill, 1975.

The appendix begins on the next page

Appendix

In this appendix, we show how to derive the variance of W, which is equivalent to the volume under the surface, and the covariance of two volume measures W_1 and W_2. We start with the variance calculation; a similar calculation for the easier case of two-class diagnostic tests is given by Lehmann.[17]

Recall that W is defined as

$$W = \frac{1}{mnl} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{l} cr(X_i, Y_j, Z_k)$$

For simplicity, we omit summation limits in the following. Indices i, I run from 1 to m, j, J from 1 to n, and k, K from 1 to l. Also, we introduce the shorthand notation $U_{ijk} = cr(X_i, Y_j, Z_k)$. Then, we can write

$$\begin{aligned}
\mathrm{Var}(W) &= \mathrm{Cov}(W, W) \\
&= \mathrm{Cov}\left(\frac{1}{mnl} \sum_i \sum_j \sum_k U_{ijk}, \frac{1}{mnl} \sum_I \sum_J \sum_K U_{IJK} \right) \\
&= \frac{1}{m^2 n^2 l^2} \sum_i \sum_j \sum_k \sum_I \sum_J \sum_K \mathrm{Cov}(U_{ijk}, U_{IJK})
\end{aligned}$$

We know that $E(U_{ijk}) = P[cr(X_i, Y_j, Z_k) = 1] = \theta$ and thus

$$\mathrm{Cov}(U_{ijk}, U_{IJK}) = E(U_{ijk}, U_{IJK}) - \theta^2 \tag{5}$$

Since the X_i, X_I, Y_j, Y_J, Z_k, Z_K are all independent, the covariance of U_{ijk} and U_{IJK} is zero if $i \neq I$ and $j \neq J$ and $k \neq K$. We are thus dealing with only $2^3 - 1 = 7$ cases where at least one of the index pairs (i, I), (j, J), and (k, K) is not equal. The sixfold sum above can thus be split into parts, depending on which indices are equal. We can then, using equation 5, write the sum as

$$\sum_i \sum_j \sum_k \sum_I \sum_J \sum_K \mathrm{Cov}(U_{ijk}, U_{IJK}) =$$

$$\sum_i \sum_j \sum_k [E(U_{ijk}, U_{ijk}) - \theta^2] + \qquad (I = i, J = j, K = k) \tag{6}$$

$$\sum_i \sum_j \sum_k \sum_{K \neq k} [E(U_{ijk}, U_{ijK}) - \theta^2] + \qquad (I = i, J = j, K \neq k) \tag{7}$$

$$\sum_i \sum_j \sum_k \sum_{J \neq j} [E(U_{ijk}, U_{iJk}) - \theta^2] + \qquad (I = i, J \neq j, K = k) \tag{8}$$

$$\sum_i \sum_j \sum_k \sum_{I \neq i} [E(U_{ijk}, U_{Ijk})\, \theta^2] + \qquad (I \neq i, J = j, K = k) \tag{9}$$

$$\sum_i \sum_j \sum_{J \neq j} \sum_{K \neq k} [E(U_{ijk}, U_{iJK}) - \theta^2] + (I = i, J \neq j, K \neq k) \tag{10}$$

$$\sum_i \sum_j \sum_k \sum_{I \neq i} \sum_{K \neq k} [E(U_{ijk}, U_{IjK}) - \theta^2] + (I \neq i, J = j, K \neq k) \tag{11}$$

$$\sum_i \sum_j \sum_k \sum_{I \neq i} \sum_{J \neq j} [E(U_{ijk}, U_{IJk}) - \theta^2] \qquad (I \neq i, J \neq j, K = k) \tag{12}$$

Looking at component 6 of the equality above, we see that the term

$$E(U_{ijk}, U_{ijk}) - \theta^2 = P(U_{ijk} = 1) - \theta^2 = \theta - \theta^2$$

is counted mnl times, so this line equals $mnl\theta(1 - \theta)$. Similarly, on line 7, we have $E(U_{ijk}, U_{ijK}) = P(U_{ijk}, U_{ijK} = 1) = q_{12}$, occurring $mnl(l - 1)$ times, so this line is equal to $mnl(l - 1)(q_{12} - \theta^2)$. The derivations for the remaining lines are analogous, down to line 12, which is equal to $mnl(m - 1)(n - 1)(q_3 - \theta^2)$. We obtain the result in equation 1 by dividing by $m^2 n^2 l^2$.

We can proceed in a manner to derive equation 4 for the covariance of two VUS estimates. Using the notation $U^1_{ijk} = cr(X^1_i, Y^1_j, Z^1_k)$ and $U^2_{IJK} = cr(X^2_I, Y^2_J, Z^2_K)$, we can write two Mann–Whitney U-statistics based on estimate triples from raters 1 and 2 as

$$W_1 = \frac{1}{mnl}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l} U^1_{ijk}$$

$$W_2 = \frac{1}{mnl}\sum_{I=1}^{m}\sum_{J=1}^{n}\sum_{K=1}^{l} U^2_{IJK}$$

We omit summation limits in the following and write

$$\mathrm{Cov}(W_1, W_2) = \mathrm{Cov}\left(\frac{1}{mnl}\sum_i\sum_j\sum_k U^1_{ijk}, \frac{1}{mnl}\sum_I\sum_J\sum_K U^2_{IJK}\right)$$

$$= \frac{1}{m^2n^2l^2}\sum_i\sum_j\sum_k\sum_I\sum_J\sum_K \mathrm{Cov}(U^1_{ijk}, U^2_{IJK})$$

The sixfold sum can be split based on possible combinations of index pairs. Independence of the estimates implies that $\mathrm{Cov}(U^1_{ijk}, U^2_{IJK}) = 0$ for $I \neq i$, $J \neq j$, $K \neq k$. The remaining combinations of index pairs can be split to yield

$$\sum_i\sum_j\sum_k\sum_I\sum_J\sum_K \mathrm{Cov}(U^1_{ijk}, U^2_{IJK}) =$$

$$\sum_i\sum_j\sum_k [E(U^1_{ijk}, U^2_{ijk}) - \theta_1\theta_2] + \qquad (I = i, J = j, K = k) \tag{13}$$

$$\sum_i\sum_j\sum_k\sum_{K\neq k} E(U^1_{ijk}, U^2_{ijK}) - \theta_1\theta_2] + \qquad (I = i, J = j, K \neq k) \tag{14}$$

$$\sum_i\sum_j\sum_k\sum_{J\neq j} E(U^1_{ijk}, U^2_{iJk}) - \theta_1\theta_2] + \qquad (I = i, J \neq j, K = k) \tag{15}$$

$$\sum_i\sum_j\sum_k\sum_{I\neq i} E(U^1_{ijk}, U^2_{Ijk}) - \theta_1\theta_2] + \qquad (I \neq i, J = j, K = k) \tag{16}$$

$$\sum_i\sum_j\sum_k\sum_{J\neq j}\sum_{K\neq k} E(U^1_{ijk}, U^2_{iJK}) - \theta_1\theta_2] + \quad (I = i, J \neq j, K \neq k) \tag{17}$$

$$\sum_i\sum_j\sum_k\sum_{I\neq i}\sum_{K\neq k} E(U^1_{ijk}, U^2_{IjK}) - \theta_1\theta_2] + \quad (I \neq i, J = j, K \neq k) \tag{18}$$

$$\sum_i\sum_j\sum_k\sum_{I\neq i}\sum_{J\neq j} E(U^1_{ijk}, U^2_{IJk}) - \theta_1\theta_2] \qquad (I \neq i, J \neq j, K = k) \tag{19}$$

The term

$$E(U^1_{ijk}, U^2_{ijk}) - \theta_1\theta_2 = P(U^1_{ijk} = U^2_{ijk} = 1) - \theta_1\theta_2 = cq_{123} - \theta_1\theta_2$$

in line 13 occurs mnl times in the above sum, so that line 13 is equal to $mnl(cq_{123} - \theta_1\theta_2)$. The remaining lines can be simplified similarly, using the definitions of cq_{12}, cq_{13}, cq_{23}, cq_1, cq_2, cq_3, respectively. Dividing by $m^2n^2l^2$ gives the result in equation 4.