Presented at the International Conference on Information Processing in Medical Imaging, Paris, July 2-6, 1979. To be published in PROCEEDINGS.

A TEST FOR THE STATISTICAL SIGNIFICANCE OF DIFFERENCES BETWEEN ROC CURVES

Metz, Charles E. and Kronman, Helen B.

Department of Radiology, The University of Chicago
and
The Franklin McLean Memorial Research Institute

950 East 59th Street, Chicago, Illinois 60637 U.S.A.

INTRODUCTION

Receiver Operating Characteristic (ROC) curve analysis provides a powerful and broadly applicable approach to the problem of evaluating the diagnostic performance of medical imaging systems.[1,2] In the past, there has been no statistical test with which one can adequately address the question: "Is an observed difference between two measured ROC curves statistically significant?" An ROC curve obtained in a visual detection experiment usually must be described by two parameters. Previous attempts to develop a significance test for ROC data have dealt with only a single index of signal detectability, by assuming that each of the ROC curves in question is described by a single parameter such as d' (which is inadequate for most visual detection experiments), by testing differences in only one of the two parameters describing an ROC curve, or by summarizing each ROC curve in terms of a single and incomplete index of performance, such as the area under the curve.

Recently we developed and evaluated a statistical test which simultaneously takes into account apparent differences in both parameters of the two ROC curves in question. The sets of rating scale data obtained for estimation of the two ROC curves are assumed to be statistically independent. Also, each ROC curve is assumed to be binormal -- that is, to be of a functional form that plots as a straight line with generally non-unit slope on double normal-deviate axes.[3] The literature of experimental psychology provides much empirical evidence that curves of this functional form provide good fits to ROC data from experiments in which decisions are based on subjective judgements.

On the basis of these assumptions, maximum likelihood estimates of the two ROC curve parameters associated with each set of rating scale data are computed using the "method of scoring."[4,5] This procedure yields not only maximum likelihood estimates of the two parameters of each ROC curve, but also estimates of the variances and the covariance of the parameter estimates.

DISCLAIMER

It is known theoretically, under rather general conditions, that maximum likelihood estimates of underlying parameters follow a multi-variate normal distribution in the limit of large numbers of trials in the experimental data.[6] By assuming that the maximum likelihood estimates of the ROC curve parameters are sampled from a bivariate normal distribution, we construct from the two pairs of ROC parameters a test statistic that should follow the Chi-square distribution (with two degrees of freedom) when the two sets of rating scale data arise, in fact, from the same ROC curve. The hypothesis that the two sets of rating scale data arise from a single underlying ROC curve is then rejected at the $100(1-\alpha)\%$ confidence level if the test statistic exceeds the critical value c for which Prob $(\chi^2_{\nu=2} \geq c) = \alpha$.

THEORY

Consider two independent pairs of normal random variables $(\hat{A}_1, \hat{B}_1)$ and $(\hat{A}_2, \hat{B}_2)$. Each pair represents the estimates of the two parameters of an ROC curve. Assume that the variances and covariances of the four parameters are known, with values

$$\text{Var}\{\hat{A}_1\} = \sigma_{A_1}^2 , \quad \text{Var}\{\hat{B}_1\} = \sigma_{B_1}^2 ,$$

$$\text{Var}\{\hat{A}_2\} = \sigma_{A_2}^2 , \quad \text{Var}\{\hat{B}_2\} = \sigma_{B_2}^2 ,$$

and

$$\text{Covar}\{\hat{A}_1, \hat{B}_1\} = \sigma_{A_1B_1} , \quad \text{Covar}\{\hat{A}_2, \hat{B}_2\} = \sigma_{A_2B_2} .$$

The covariance of $\hat{A}_1$ and $\hat{B}_1$ and the covariance of $\hat{A}_2$ and $\hat{B}_2$ are generally non-zero because the $\hat{A}_i$ and $\hat{B}_i$ are estimated jointly from a single set of observer response data. The covariances of $\hat{A}_1$ and $\hat{A}_2$, of $\hat{A}_1$ and $\hat{B}_2$, of $\hat{B}_1$ and $\hat{A}_2$, and of $\hat{B}_1$ and $\hat{B}_2$ must be zero, however, since the data sets used to estimate the pairs $(\hat{A}_1, \hat{B}_1)$ and $(\hat{A}_2, \hat{B}_2)$ are statistically independent.

If we wish to determine the statistical significance of an observed difference between two ROC curves, then we must test the null hypothesis that the two estimated ROC curves arose, in fact, from a single underlying ROC curve. This is equivalent to the null hypothesis that $E\{\hat{A}_1\}=E\{\hat{A}_2\}$ and $E\{\hat{B}_1\}=E\{\hat{B}_2\}$, or $E\{\hat{A}_1-\hat{A}_2\}=0$ and $E\{\hat{B}_1-\hat{B}_2\}=0$, where $E\{\ \}$ indicates "expected value of."

Let $A_d \equiv \hat{A}_1-\hat{A}_2$ and $B_d \equiv \hat{B}_1-\hat{B}_2$ represent the observed differences between the estimates of the A parameters and the B parameters of the two measured ROC curves. Then we must test the null hypothesis that $E\{A_d\}=0$ and $E\{B_d\}=0$. Since the two ROC curves were estimated from independent data,

$$\text{Var}\{A_d\} = \text{Var}\{\hat{A}_1\} + \text{Var}\{\hat{A}_2\}$$
$$= \sigma_{A_1}^2 + \sigma_{A_2}^2$$
$$\equiv \sigma_{A_d}^2$$

$$\text{Var}\{B_d\} = \text{Var}\{\hat{B}_1\} + \text{Var}\{\hat{B}_2\}$$
$$= \sigma_{B_1}^2 + \sigma_{B_2}^2$$
$$\equiv \sigma_{B_d}^2$$

and
$$\text{Covar}\{A_d, B_d\} = \text{Covar}\{A_1, B_1\} + \text{Covar}\{A_2, B_2\}$$
$$= \sigma_{A_1 B_1} + \sigma_{A_2 B_2}$$
$$\equiv \rho\, \sigma_{A_d}\, \sigma_{B_d} \quad ,$$

where ρ represents the correlation coefficient for A_d and B_d. These relationships can be summarized by the matrix equation $W = S_1 + S_2$, where W represents the "covariance matrix" for A_d and B_d:

$$W \equiv \begin{bmatrix} \sigma_{A_d}^2 & \sigma_{A_d B_d} \\ \sigma_{A_d B_d} & \sigma_{B_d}^2 \end{bmatrix} = \begin{bmatrix} \sigma_{A_d}^2 & \rho\sigma_{A_d}\sigma_{B_d} \\ \rho\sigma_{A_d}\sigma_{B_d} & \sigma_{B_d}^2 \end{bmatrix}$$

and S_i represents the covariance matrix for $\hat{A}_i$ and $\hat{B}_i$:

$$S_i \equiv \begin{bmatrix} \sigma_{A_i}^2 & \sigma_{A_i B_i} \\ \sigma_{A_i B_i} & \sigma_{B_i}^2 \end{bmatrix} .$$

An observed difference between two estimated ROC curves is represented by the pair of numbers (A_d, B_d). If each pair $(\hat{A}_1, \hat{B}_1)$ and $(\hat{A}_2, \hat{B}_2)$ arises from a bivariate normal distribution, then the pair (A_d, B_d) also follows a bivariate normal distribution. According to the null hypothesis, the expected value of this pair is (0, 0). Thus according to the null hypothesis (H_o), the probability density distribution of A_d and B_d is given by

$$f(A_d, B_d|H_o) = \frac{1}{2\pi\sigma_{A_d}\sigma_{B_d}\sqrt{1-\rho^2}} \cdot \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(A_d)^2}{\sigma_{A_d}^2} - \frac{2\rho A_d B_d}{\sigma_{A_d}\sigma_{B_d}} + \frac{(B_d)^2}{\sigma_{B_d}^2} \right] \right\}$$

in which σ_{A_d}, σ_{B_d}, and ρ are defined as above and can be calculated from the known (or estimated) variances and covariances of A_1, B_1, A_2, and B_2. The locus of points (A_d, B_d) for which this probability density is constant plots as an ellipse centered at (0, 0) in the (A_d, B_d) plane, with orthogonal major and minor axes given by the eigenvectors of the matrix W. The integral of $f(A_d, B_d|H_o)$ within any such ellipse represents the probability of observing a pair (A_d, B_d) within that ellipse of constant probability density. Thus if an observed pair (A_d, B_d) lies outside the ellipse within which the integral of $f(A_d, B_d|H_o)$ equals 0.95, for example, then we can conclude that the null hypothesis (i.e., that $E\{\hat{A}_1\}=E\{\hat{A}_2\}$ and $E\{\hat{B}_1\}=E\{\hat{B}_2\}$) should be rejected with $(1-\alpha) = 0.95$ confidence.

If a new set of coordinate axes in the (A_d, B_d) plane is chosen by rotation about the origin so that the directions of the new axes coincide with the major and minor axes of a constant probability density ellipse, and if these new axes are then rescaled so that the variance of each transformed coordinate is 1.0, then the ellipses of constant probability density are transformed into circles of constant probability density, and the probabilities within an original ellipse and the corresponding circle will be equal. In the new transformed coordinate system (x, y), the probability density distribution is given by

$$f(x, y|H_o) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)},$$

and the probability P that a pair (x, y) -- representing a transformed pair (A_d, B_d) -- lies inside a circle of radius r is given by

$$P \equiv \text{Prob}\ (x^2+y^2 < r^2)$$

$$= 1 - e^{-\frac{1}{2}(r^2)} \quad .$$

This is the cumulative probability distribution of a Chi-square random variable with 2 degrees of freedom, $\chi^2_{\nu=2}$. Thus the statistical significance of an observed difference between $(\hat{A}_1, \hat{B}_1)$ and $(\hat{A}_2, \hat{B}_2)$, which is the same as the statistical significance of an observed difference between (A_d, B_d) and (0, 0), can be determined by appropriately rotating and scaling the coordinate axes in the (A_d, B_d) plane to express the pair (A_d, B_d) in terms of the independent standard deviate pair (x, y), and by then finding the probability of observing a value of $\chi^2_{\nu=2}$ that is as large or larger than $r^2 \equiv x^2+y^2$:

$$\text{Prob}(\chi^2_{\nu=2} \geq x^2+y^2) = e^{-\frac{1}{2}(x^2+y^2)} \quad .$$

If this probability is less than a critical value α (e.g., 0.05), then we conclude that the observed difference between the two ROC curves is significant at the α level used.

To summarize the above discussion, mathematical formulation of the appropriate Chi-square statistic can be thought of in the following way. First, one determines the eigenvectors of the matrix W, which express the major and minor axes of the constant probability density ellipses. Next, one expresses a pair of parameter value differences (A_d, B_d) in terms of a coordinate system that has been rotated to coincide with these eigenvectors, and divides each transformed coordinate by the square root of the corresponding eigenvalue of W (i.e., by the standard deviation of the transformed variable on the eigenvector axis), so that the resulting coordinates equivalent to (A_d, B_d) are independent standard normal deviates (i.e., normal with zero mean and unit variance). Then one can show that the required χ^2 statistic is given in matrix form by

$$\chi^2_{\nu=2} = \begin{bmatrix} A_d & B_d \end{bmatrix} W^{-1} \begin{bmatrix} A_d \\ B_d \end{bmatrix}$$

$$= \begin{bmatrix} (A_1-A_2) & (B_1-B_2) \end{bmatrix} \begin{bmatrix} (\sigma_{A_1}^2+\sigma_{A_2}^2) & (\sigma_{A_1B_1}+\sigma_{A_2B_2}) \\ (\sigma_{A_1B_1}+\sigma_{A_2B_2}) & (\sigma_{B_1}^2+\sigma_{B_2}^2) \end{bmatrix}^{-1} \begin{bmatrix} (A_1-A_2) \\ (B_1-B_2) \end{bmatrix}$$

or in algebraic form by

$$\chi^2_{\nu=2} = \frac{1}{1-\rho^2}\left[\frac{(A_d)^2}{\sigma_{A_d}^2} - \frac{2\rho A_d B_d}{\sigma_{A_d}\sigma_{B_d}} + \frac{(B_d)^2}{\sigma_{B_d}^2}\right]$$

where $A_d \equiv A_1 - A_2$, $B_d \equiv B_1 - B_2$, $\sigma_{A_d} \equiv \sqrt{\sigma_{A_1}^2 + \sigma_{A_2}^2}$, $\sigma_{B_d} \equiv \sqrt{\sigma_{B_1}^2 + \sigma_{B_2}^2}$, and

$\rho = (\sigma_{A_1B_1} + \sigma_{A_2B_2}) / \sigma_{A_d}\sigma_{B_d}$.

All of the terms on the right-hand side of this equation can be calculated from the two pairs of estimated ROC curve parameters, $(\hat{A}_1, \hat{B}_1)$ and $(\hat{A}_2, \hat{B}_2)$, and their associated variances and covariances. All of these are provided by the computer program for maximum likelihood estimation from rating scale data.

Any difference between the estimated ROC curves $(\hat{A}_1, \hat{B}_1)$ and $(\hat{A}_2, \hat{B}_2)$ is then significant at the $100(1-\alpha)\%$ confidence level if the value of $\chi^2_{\nu=2}$ obtained by the above procedure exceeds $-2\ln(\alpha)$. Note that use of the $100(1-\alpha)\%$ confidence level for significance testing implies that one expects the null hypothesis (of no difference) to be rejected falsely in $100\alpha\%$ of all similar situations for which the two sets of rating scale data arose, in fact, from the same underlying ROC curve.

EVALUATION OF THE SIGNIFICANCE TEST

The statistical test described above was derived on the basis of an assumption that the differences between ROC curve parameter estimates, $A_d \equiv (\hat{A}_1 - \hat{A}_2)$ and $B_d \equiv (\hat{B}_1 - \hat{B}_2)$, are distributed as standard normal deviates after they are transformed on the basis of their estimated variances and covariance. Because this assumption is known to be valid only in the limit of large numbers of experimental trials, one must determine empirically the extent to which the test performs adequately for the numbers of trials typically used in psychophysical and medical applications of ROC analysis. To perform this evaluation, a digital computer was used to simulate sets of rating scale data from a binormal decision model. The statistical test was then applied to ROC curves estimated from these data, and observed performance of the test was compared to ideal performance.

In the following, we report some of the results of these simulations for the binormal ROC curves labeled as Curve 1 and Curve 3 in Figure 1. For each curve, the parameter A represents the Z_y intercept and the parameter B represents the slope when the ROC curve is plotted as a straight line on double normal-deviate axes. For each ROC curve, one thousand independent sets of rating scale data were simulated containing m "noise-only" trials and m "signal-plus-noise" trials for m=50, 250, and 500. Expected operating points on each ROC curve were held constant during the simulation of each data set but were varied randomly across data sets to yield realistic distributions of operating points. Typical sets of expected operating points on each ROC curve are shown in Figure 1 by similar symbols. Details of our computer simulation of rating scale data will be published elsewhere.

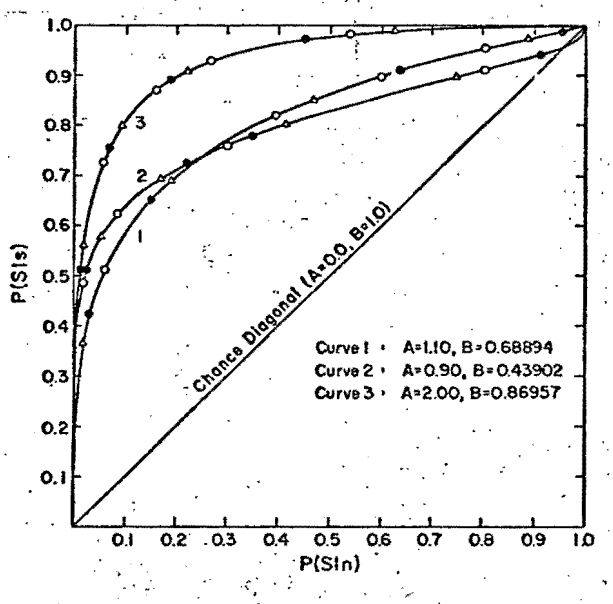

Fig. 1. The binormal ROC curves used to generate the simulated rating scale data. On each curve, similar symbols indicate typical expected operating points.

The algorithm described by Dorfman and Alf[4] and corrected by Grey and Morgan[5] was used to compute maximum likelihood estimates of the parameters A and B and their variances and covariance from each set of rating scale data. Then the statistical significance test was applied to arbitrarily selected pairs of estimated ROC curves that had arisen from the same underlying ROC curve. For each underlying ROC curve and for each number of trials, the fraction of curve pairs in which the difference was found significant at various α levels was tabulated. The differences found significant in this way were <u>falsely</u> significant (i.e., were type I errors) because both estimated ROC curves in each pair tested had arisen from the same underlying ROC curve.

To evaluate the performance of the test statistic, the fractions of (falsely) significant results (f_s) obtained at various α levels between 0 and 1 were compared to the expected proportions (α). If the proposed test performs well, f_s should be close to α for any α between 0 and 1.

RESULTS

We report here some typical results of our evaluation of the performance of the test statistic when it was applied to two sets of simulated rating scale data generated from the same underlying ROC curve and containing the same number of trials. A more extensive summary of our results will be published elsewhere.

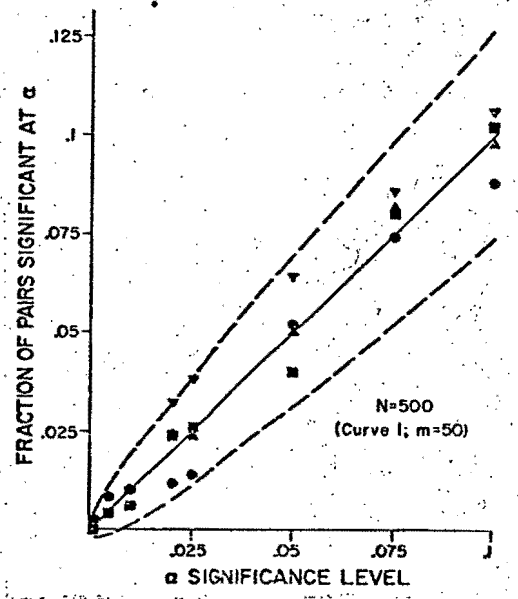

Fig. 2. Comparison of f_s and α for ROC curve #1 and 50 trials of each kind in each data set.

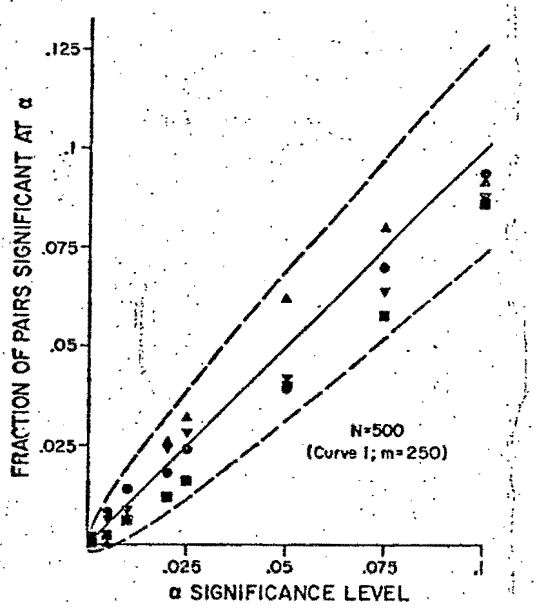

Fig. 3. Comparison of f_s and α for ROC curve #1 and 250 trials of each kind in each data set.

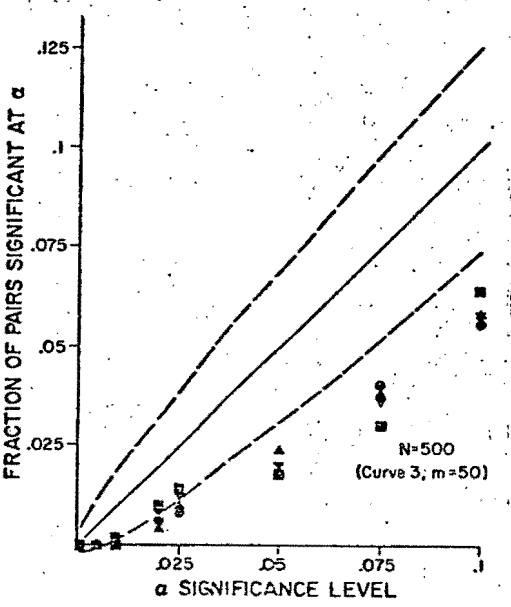

Fig. 4. Comparison of f_s and α for ROC curve #3 and 50 trials of each kind in each data set.

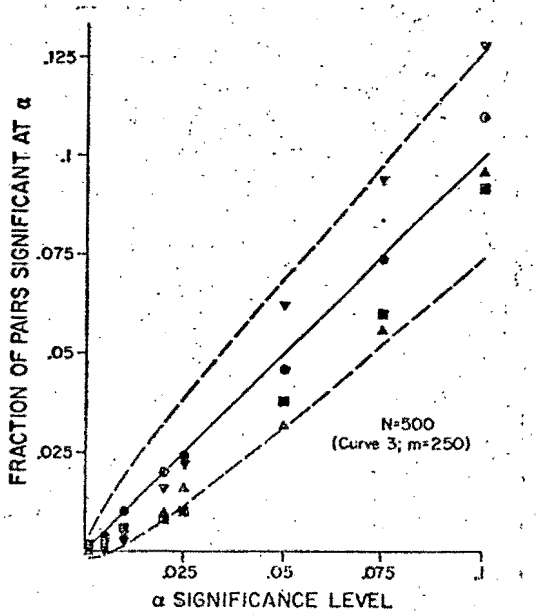

Fig. 5. Comparison of f_s and α for ROC curve #3 and 250 trials of each kind in each data set.

Figures 2 through 5 compare f_s, the fraction of 500 curve pairs found significant using our test, and α, the fraction expected with a perfect test, for values of α in the range $0<\alpha\leq 0.1$, which is the range of greatest interest in practical applications of statistical hypothesis testing. The broken lines indicate the 95% probability band for the 500 pairs tested here. In each figure, symbols of different shape indicate the results of four different arbitrary methods of pairing the estimated ROC curves. The results for paired estimates of ROC curve #1 are shown in Figure 2 for m=50 trials of each kind for each curve estimate and in Figure 3 for m=250. The results for paired estimates of ROC curve #3 are shown for m=50 in Figure 4 and for m=250 in Figure 5. The correspondence between f_s and α found for m=500, though not shown here, was similar to -- but slightly better than -- that obtained for m=250 in Figures 3 and 5. In general, for all ROC curves and all numbers of trials, the agreement between f_s and α improved for values of α greater than 0.1.

The selected results shown in Figures 2 through 5 are typical of those found for paired ROC curve estimates based on equal numbers of trials. Our statistical test for differences between ROC curves performed better -- that is, the observed values of f_s lay closer to α relative to the 95% probability bands -- (1) for larger numbers of trials in the rating scale data for each curve, and (2) for lower ROC curves (i.e., curves representing less detectability). Both observations can be explained by the related facts that (1) for a given underlying ROC curve, the distributions of ROC curve parameter estimates become more nearly normal as the asymptotic limit is approached by larger numbers of trials; and (2) for a given number of trials in the rating scale data, the parameter estimate distributions, especially for m=50, are less normal for higher ROC curves because uncertainties in the ROC curve operating points estimated from the data become larger (and nonlinear), relative to a given increment in the curve parameters, in the upper left portion of the ROC space.

The proposed test clearly performs very well for typical ROC curves when 250 or more trials of each kind are used to estimate each ROC curve. When only 50 trials of each kind are used to estimate each curve, the test performs well for ROC curves like #1 and #2 in Figure 1, but somewhat less reliably for higher ROC curves like curve #3. Even in this situation the test is probably adequate for most applications, however. The results of Figure 4, for example, suggest that when the test is used with a typical α level of 0.05, the fraction

of curve pairs found falsely significant will be somewhat smaller than expected, e.g., about 0.025 instead of 0.05. In order to fully understand the impact of such incorrect predictions of Type I error rate, one must determine the performance of the statistical test with regard to the rate of Type II errors (falsely accepting the null hypothesis of no actual difference between ROC curves when, in fact, a real difference exists). In general, the trade-offs that are possible between Type I error and Type II error rates by selection of different α levels are analogous to the compromises that can be made between false-positive and false-negative decision fractions in a detection experiment, which can be described by an ROC curve.

DISCUSSION

The statistical test that we propose here for differences between ROC curves perform well for typical ROC curves estimated from typical numbers of experimental trials in a rating scale experiment. For moderately high ROC curves (such as Curve 3 in Figure 1) and small numbers of trials (such as 50 of each kind), the test statistic yields Type I errors less frequently than would be expected on the basis of the α level used for the test. Even in this situation the discrepancy is small, however. Thus the proposed test appears to provide a useful assessment of the statistical significance of apparent differences between ROC curves estimated from independent rating scale data.

If a statistically significant difference is found between two independent ROC curves, one may wish to be able to state which ROC curve is "better" than the other. If the two ROC curves do not intersect in a region of the ROC space of interest, then the ROC curve closer to the upper left-hand corner of the ROC space can be considered better since that curve represents greater detectability. However, the decision is not always so clearcut, as in the case of Curves 1 and 2 in Figure 1, which intersect each other near the negative diagonal of the ROC space. Two approaches are possible in such seemingly ambiguous situations. One is to examine those portions of the curves where the decision maker is likely to operate and determine which portion exhibits greater detectability. The other is to calculate the areas under the estimated ROC curves, since the area is an overall performance index. This latter alternative can always be used because the area measure requires no assumption regarding where the decision maker is likely to operate; the resulting conclusion may be misleading, however, if particular operating points are of primary interest in an applied detection task.

In its present form, the statistical test we describe here requires that the data sets obtained for the two ROC curves in question be statistically independent. Thus the present test is applicable to sets of ROC data due to different sets of noisy images produced from the same phantom, or due to sets of clinical images made from different patients. This test is not applicable, however, when two sets of ROC data are generated by the same or different observers viewing the same set of images, or when the two data sets are obtained from different clinical images of the same set of patients. A generalized statistical test that will take such correlations into account is currently undergoing development.

ACKNOWLEDGEMENT

The Franklin McLean Memorial Research Institute is operated by The University of Chicago for the U. S. Department of Energy under Contract No. EY-76-C-02-0069.

REFERENCES

1. Metz, C.E.: Basic principles of ROC analysis. Semin. Nucl. Med., 8: 283-298, 1978.

2. Swets, J.A.: ROC analysis applied to the evaluation of medical imaging techniques. Invest. Radiol., 14: 109-121, 1979.

3. Green, D.M., and Swets, J.A.: Signal Detection Theory and Psychophysics (revised edition). Huntington, NY: Krieger, 1974, pp. 62-64.

4. Dorfman, D.D., and Alf, E.: Maximum likelihood estimation of parameters of signal-detection theory and determination of confidence intervals -- rating method data. J. Math. Psych., 6: 487-496, 1969.

5. Grey, D.R., and Morgan, B.J.T.: Some aspects of ROC curve-fitting: normal and logistic models. J. Math. Psych. 9: 128-139, 1972.

6. Rao, C.R.: Advanced Statistical Methods in Biometric Research (2nd edition). Darien, Connecticut: Hafner, 1970, pp. 151-173.

SUMMARY

A test for the statistical significance of observed differences between two measured ROC curves has been designed and evaluated. The set of observer response data for each ROC curve is assumed to be independent and to arise from an ROC curve having a form which, in the absence of statistical fluctuations in the response data, graphs as a straight line on double normal-deviate axes. Such a "binormal" ROC curve is defined by two parameters, which represent the slope and one axis intercept of the normal-deviate graph.

To test the significance of an apparent difference between two measured ROC curves, maximum likelihood estimates of the two parameters of each curve and the associated parameter variances and covariance are calculated from the corresponding set of observer response data. An approximate Chi-square statistic with two degrees of freedom is then constructed from the differences between the parameters estimated for each ROC curve and from the variances and covariances of these estimates.

This statistic is known to be truly Chi-square distributed only in the limit of large numbers of trials in the observer performance experiments. Performance of the statistic for data arising from a limited number of experimental trials was evaluated by simulating five-category rating scale data with 50, 250, and 500 each of noise and signal-plus-noise trials, and by applying the test to these data. Independent sets of rating scale data arising from the same underlying ROC curve were paired, and the fraction of differences found (falsely) significant was compared to the significance level, α, used with the test. Although test performance was found to be somewhat dependent on both the number of trials in the data and the position of the underlying ROC curve in the ROC space, the results for various significance levels showed the test to be reliable under practical experimental conditions.