[Reprinted from RADIOLOGY, Vol. 148, No. 3, Pages 839–843, September, 1983.]

James A. Hanley, Ph.D.
Barbara J. McNeil, M.D., Ph.D.

A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases[1]

Receiver operating characteristic (ROC) curves are used to describe and compare the performance of diagnostic technology and diagnostic algorithms. This paper refines the statistical comparison of the areas under two ROC curves derived from the same set of patients by taking into account the correlation between the areas that is induced by the paired nature of the data. The correspondence between the area under an ROC curve and the Wilcoxon statistic is used and underlying Gaussian distributions (binormal) are assumed to provide a table that converts the observed correlations in paired ratings of images into a correlation between the two ROC areas. This between-area correlation can be used to reduce the standard error (uncertainty) about the observed difference in areas. This correction for pairing, analogous to that used in the paired t-test, can produce a considerable increase in the statistical sensitivity (power) of the comparison. For studies involving multiple readers, this method provides a measure of a component of the sampling variation that is otherwise difficult to obtain.

Index term: Receiver operating characteristic curve (ROC)

Radiology 148: 839–843, September 1983

SEVERAL questions dealing with comparative benefits for alternative diagnostic algorithms, diagnostic tests, or therapeutic regimens have recently emerged in medicine. For example, how do we know whether one diagnostic algorithm is better than another in sorting patients into diseased and nondiseased groups? Whether the addition of a new test or procedure to an established algorithm improves its performance? Whether it matters who of several available readers interprets a mammogram? Whether one type of hard-copy unit in radiology is better than another? Whether reading a CT scan in conjunction with the patient's history allows a more accurate diagnosis than reading it without the history? The analyses of such problems have started with construction of receiver operating characteristic (ROC) curves (1–3). Generally these analyses have used as cutoff points either different posterior probabilities on a continuous scale or different thresholds on a discrete rating scale. The latter approach has been particularly popular in radiology.

Major gaps in the understanding of statistical properties of ROC curves have limited their usefulness, especially for questions involving comparisons of curves based on the same sample of subjects or objects. These comparative situations contrast with those involving a single data set and a single ROC curve. In such cases, the investigator generally only needs to know that a single modality or diagnostic approach has "poor", "moderate", or "good" accuracy, and the location of the ROC curve gives a rough assessment. However, when a comparison of two algorithms or modalities is relevant, more formal statistical criteria are needed in order to judge whether observed differences in accuracy are more likely to be random than real. Thus far these criteria have not been fully developed for ROC curves.

In a recent paper (4) we dealt with one popular accuracy index that can be derived from and used as a summary of the ROC curve. We showed that the relationship of the area under the ROC curve to the Wilcoxon statistic could be used to derive its statistical properties, such as its standard error (SE) and the sample sizes required to measure the area with a prespecified degree of precision (reliability) and to provide a desired level of statistical power (low type II error) in comparative experiments. This paper extends our statistical analysis to another large class of situations, where the two or more ROC curves are generated using the same set of patients. In these situations, it is inappropriate to calculate the standard error of the difference between two areas ($A\hat{r}ea_1$ and $A\hat{r}ea_2$) as

$$SE\,(A\hat{r}ea_1 - A\hat{r}ea_2) = \sqrt{SE^2(A\hat{r}ea_1) + SE^2(A\hat{r}ea_2)} \qquad (1)$$

since $A\hat{r}ea_1$ and $A\hat{r}ea_2$ are likely to be correlated. This correlation is likely to be positive; if the vagaries of random sampling of cases produce a higher/lower than expected accuracy index for one modality (*e.g.*, if the sample consisted of a larger than usual number of easy/difficult cases), then the accuracy of the second modality will probably also be correspondingly higher/lower than one would expect. In other words, while the two indices may fluctuate indepen-

[1] From the Department of Epidemiology and Health, McGill University, Montreal, Canada (J.A.H.) and the Department of Radiology, Harvard Medical School and Brigham and Women's Hospital, Boston, MA, USA (B.J.M.). Received June 3, 1981; revision requested July 21, 1981; final revision received and accepted Feb. 15, 1983.

Supported in part by the Hartford Foundation and the National Center for Health Care Technology. ht

dently by amounts SE_1 and SE_2 in separate samples, they will tend to fluctuate in tandem when derived from a single sample.

In this paper we have developed an approach to take account of this correlation. In brief, we indicate that the relevant standard error for such comparisons is not that shown in Equation 1 but rather

$$SE(A\hat{r}a_1 - A\hat{r}a_2) = \sqrt{SE^2(A\hat{r}a_1) + SE^2(A\hat{r}a_2) - 2rSE(A\hat{r}a_1)\,SE(A\hat{r}a_2)} \quad (2)$$

where r is a quantity representing the correlation introduced between the two areas by studying the same sample of patients. This paper reviews the calculations for comparing the ROC curves of two modalities and illustrates this new approach using data from a series of experiments involving phantoms.

METHODS

The general approach to assessing whether the difference in the areas under two ROC curves derived from the same set of patients is random or real is to calculate a critical ratio z, defined as

$$z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}} \quad (3)$$

where A_1 and SE_1 refer to the observed area and estimated standard error of the ROC area associated with modality 1; where A_2 and SE_2 refer to corresponding quantities for modality 2; and where r represents the estimated correlation between A_1 and A_2.[2] This quantity z is then referred to tables of the normal distribution and values of z above some cutoff, *e.g.*, $z \geq 1.96$, are taken as evidence that the "true" ROC areas are different. The importance of introducing the $2rSE_1SE_2$ term in the above equation is obvious: failure to subtract out from the sampling variability those fluctuations that the paired design has already eliminated will leave the denominator of Equation 3 too large and z too small, thereby reducing the chance of detecting a difference between two modalities.

Calculating Areas

Areas under ROC curves can be obtained in three ways: (*i*) by the trapezoidal rule; (*ii*) as output from the Dorfman and Alf maximum likelihood estimation program (5); or (*iii*) from the slope and intercept of the original data when plotted on binormal graph paper (3). As indicated in our companion paper (4) the trapezoidal approach systematically underestimates areas. Because the Dorfman and Alf approach is becoming readily accessible to those interested in this area, we will calculate areas using this approach. (For those limited to graphical methods, the area can be derived from the slope and intercept according to the rule Area = Percentage of Gaussian distribution to left of z_A, where z_A = Intercept/$\sqrt{1 + \text{slope}^2}$).

Calculating Standard Errors

The standard errors associated with areas can be obtained in three ways: (*i*) as output directly from the Dorfman and Alf maximum likelihood estimation program; (*ii*) from the variance of the Wilcoxon statistic as illustrated in detail in Reference 4; or (*iii*) from an approximation to the Wilcoxon statistic by making an assumption, shown to be conservative (compared with assuming a Gaussian-based ROC curve), that the underlying signal (diseased) and noise (nondiseased) distributions are exponential in type (4). We will use the standard errors estimated from the Dorfman and Alf program.

Calculating the Correlation Coefficient, r, Between Areas

Two intermediate correlation coefficients are required, which are then converted into a correlation between A_1 and A_2 *via* a table that we supply below. The first is r_N, the correlation coefficient for the ratings given to images from nondiseased patients by the two modalities. The second is r_A, the correlation coefficient for the ratings of diseased patients imaged by the two modalities. Each of these can be calculated in traditional ways using either the Pearson product-moment correlation method or the Kendall tau. The former approach is usually used for results derived from an interval scale whereas the latter is more appropriate for results obtained from an ordinal scale. ROC curves in radiology are derived from ordinal scale data and therefore we have used the Kendall tau for calculating r_N and r_A. Standard statistical packages (*e.g.*, SPSS, SAS) provide tau; when the number of rating categories is small, however, say four or less, the calculation can also be performed manually.

Once the correlations between the ratings (r_N among the normals, r_A among the abnormals) are obtained, it is necessary to calculate the correlation that they induce between the two areas A_1 and A_2; for ease of notation we have called this r (without any subscript). This is the coefficient present in Equations 2 and 3. Tabulation of r (TABLE I) is the fundamental contribution of this paper[3]; therefore, in our subsequent example we will illustrate its use.

Experimental Data for Illustrative Examples

We studied 112 phantoms that were specially constructed to evaluate the accuracy of two different computer algorithms used in image reconstruction for CT. Fifty-eight of these phantoms were of uniform density and were designated "normal"; the remaining 54 contained an area of reduced density to simulate a lesion and were designated "abnormal". Two images of each phantom were reconstructed using the two different algorithms, which we will refer to as modality 1 and modality 2. A single reader read each image and rated it on a 6-point scale: 1 = Definitely Normal; 2 = Probably Normal; 3 = Possibly Normal; 4 = Possibly Abnormal; 5 = Probably Abnormal; 6 = Definitely Abnormal. From the resulting data, we constructed two ROC curves. The data were submitted to the Dorfman and Alf maximum likelihood program to produce areas under the ROC curves and standard errors.

RESULTS

Our results will be divided into two parts. First, the analysis of the example involving CT phantoms will be illustrated. Then, in order to verify that the z statistic performs correctly, results of several simulations will be summarized.

CT Phantom Example

The basic data are presented in the Appendix, along with the calculations produced from them. The areas under the ROC curves were 89.45% (*SE* 3.0%) and 93.82% (*SE* 2.6%). The (Kendall tau) correlations between the paired ratings were $r_N = 0.39$ (nondiseased patients) and $r_A = 0.60$ (diseased patients), giving an "average" correlation between the ratings of 0.50. With this average correlation of 0.50 and with an average area of (89.45 + 93.82)/2 = 91.64, TABLE

[2] As we will see later, the SE of an estimated area depends on the magnitude of the underlying or "true" area. When calculating z to test the null hypothesis that this underlying area is the same for both modalities, one should equate SE_1 and SE_2, calculating them both from a common estimate of the area. In this case the denominator becomes $\sqrt{2SE^2(1-r)}$ or $SE\sqrt{2(1-r)}$.

[3] Mathematical derivation available upon request.

I indicates that the correlation r between areas is approximately 0.40.

Equation 3 was then used to calculate the critical ratio z in order to test the null hypothesis that the observed difference between observed areas was merely a result of random sampling. Using the above data

$$z = (0.9382 - 0.8945)/\sqrt{0.03^2 + 0.026^2 - 2(0.4)(0.03)(0.026)} = 0.0437/0.0309 = 1.41$$

As mentioned earlier, one might average the two areas to obtain a common area of 0.9164, and use the formula in Reference 4 to predict each of the standard errors as 0.0281; using $0.0281\sqrt{2(1-0.4)} = 0.0308$ in the denominator of Equation 3 yields an almost identical z value of 1.42.

If we have reason to believe *a priori* that modality 2 is likely to be better than modality 1 and are only interested in improvements, then a one-tailed test is appropriate. The Gaussian distribution indicates that a value of 1.41 or higher should occur roughly once in every 13 samples ($p = 0.079$); this evidence suggests that the observed difference may not be random. This contrasts with the weaker inference that would be drawn from a critical ratio of 1.10 (or a p value of 0.136 or 1 in 7) that would have been calculated had the correlation between areas been assumed to be equal to zero (in other words, had we failed to take into account the increased sensitivity induced by studying the same set of patients with both modalities). If we had no *a priori* interest in one particular direction, then a two-tailed test would have been appropriate.

TABLE I: Correlation Coefficients*

Average Correlation between Ratings†	Average Area‡											
	.700	.725	.750	.775	.800	.825	.850	.875	.900	.925	.950	.975
0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01
0.04	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02
0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.03	0.02
0.08	0.07	0.07	0.07	0.07	0.07	0.06	0.06	0.06	0.06	0.05	0.04	0.03
0.10	0.09	0.09	0.09	0.09	0.08	0.08	0.08	0.07	0.07	0.06	0.06	0.04
0.12	0.11	0.11	0.11	0.10	0.10	0.10	0.09	0.09	0.08	0.08	0.07	0.05
0.14	0.13	0.12	0.12	0.12	0.12	0.11	0.11	0.11	0.10	0.09	0.08	0.06
0.16	0.14	0.14	0.14	0.14	0.13	0.13	0.13	0.12	0.11	0.11	0.09	0.07
0.18	0.16	0.16	0.16	0.16	0.15	0.15	0.14	0.14	0.13	0.12	0.11	0.09
0.20	0.18	0.18	0.18	0.17	0.17	0.17	0.16	0.15	0.15	0.14	0.12	0.10
0.22	0.20	0.20	0.19	0.19	0.19	0.18	0.18	0.17	0.16	0.15	0.14	0.11
0.24	0.22	0.22	0.21	0.21	0.21	0.20	0.19	0.19	0.18	0.17	0.15	0.12
0.26	0.24	0.23	0.23	0.23	0.22	0.22	0.21	0.20	0.19	0.18	0.16	0.13
0.28	0.26	0.25	0.25	0.25	0.24	0.24	0.23	0.22	0.21	0.20	0.18	0.15
0.30	0.27	0.27	0.27	0.26	0.26	0.25	0.25	0.24	0.23	0.21	0.19	0.16
0.32	0.29	0.29	0.29	0.28	0.28	0.27	0.26	0.26	0.24	0.23	0.21	0.18
0.34	0.31	0.31	0.31	0.30	0.30	0.29	0.28	0.27	0.26	0.25	0.23	0.19
0.36	0.33	0.33	0.32	0.32	0.31	0.31	0.30	0.29	0.28	0.26	0.24	0.21
0.38	0.35	0.35	0.34	0.34	0.33	0.33	0.32	0.31	0.30	0.28	0.26	0.22
0.40	0.37	0.37	0.36	0.36	0.35	0.35	0.34	0.33	0.32	0.30	0.28	0.24
0.42	0.39	0.39	0.38	0.38	0.37	0.36	0.36	0.35	0.33	0.32	0.29	0.25
0.44	0.41	0.40	0.40	0.40	0.39	0.38	0.38	0.37	0.35	0.34	0.31	0.27
0.46	0.43	0.42	0.42	0.42	0.41	0.40	0.39	0.38	0.37	0.35	0.33	0.29
0.48	0.45	0.44	0.44	0.43	0.43	0.42	0.41	0.40	0.39	0.37	0.35	0.30
0.50	0.47	0.46	0.46	0.45	0.45	0.44	0.43	0.42	0.41	0.39	0.37	0.32
0.52	0.49	0.48	0.48	0.47	0.47	0.46	0.45	0.44	0.43	0.41	0.39	0.34
0.54	0.51	0.50	0.50	0.49	0.49	0.48	0.47	0.46	0.45	0.43	0.41	0.36
0.56	0.53	0.52	0.52	0.51	0.51	0.50	0.49	0.48	0.47	0.45	0.43	0.38
0.58	0.55	0.54	0.54	0.53	0.53	0.52	0.51	0.50	0.49	0.47	0.45	0.40
0.60	0.57	0.56	0.56	0.55	0.55	0.54	0.53	0.52	0.51	0.49	0.47	0.42
0.62	0.59	0.58	0.58	0.57	0.57	0.56	0.55	0.54	0.53	0.51	0.49	0.45
0.64	0.61	0.60	0.60	0.59	0.59	0.58	0.58	0.57	0.55	0.54	0.51	0.47
0.66	0.63	0.62	0.62	0.62	0.61	0.60	0.60	0.59	0.57	0.56	0.53	0.49
0.68	0.65	0.64	0.64	0.64	0.63	0.62	0.62	0.61	0.60	0.58	0.56	0.51
0.70	0.67	0.66	0.66	0.66	0.65	0.65	0.64	0.63	0.62	0.60	0.58	0.54
0.72	0.69	0.69	0.68	0.68	0.67	0.67	0.66	0.65	0.64	0.63	0.60	0.56
0.74	0.71	0.71	0.70	0.70	0.69	0.69	0.68	0.67	0.66	0.65	0.63	0.59
0.76	0.73	0.73	0.72	0.72	0.72	0.71	0.71	0.70	0.69	0.67	0.65	0.61
0.78	0.75	0.75	0.75	0.74	0.74	0.73	0.73	0.72	0.71	0.70	0.68	0.64
0.80	0.77	0.77	0.77	0.76	0.76	0.76	0.75	0.74	0.73	0.72	0.70	0.67
0.82	0.79	0.79	0.79	0.79	0.78	0.78	0.77	0.77	0.76	0.75	0.73	0.70
0.84	0.82	0.81	0.81	0.81	0.81	0.80	0.80	0.79	0.78	0.77	0.76	0.73
0.86	0.84	0.84	0.83	0.83	0.83	0.82	0.82	0.81	0.81	0.80	0.78	0.75
0.88	0.86	0.86	0.86	0.85	0.85	0.85	0.84	0.84	0.83	0.82	0.81	0.79
0.90	0.88	0.88	0.88	0.88	0.87	0.87	0.87	0.86	0.86	0.85	0.84	0.82

* Correlation coefficient r between two ROC areas A_1 and A_2 as a function of average correlation between ratings (rows) and average area (columns).
† $(r_N + r_A)/2$.
‡ $(A_1 + A_2)/2$.

General Performance of the Paired Test

A good statistical test should indicate a difference when one is really present, but it should minimize in a predictable way the number of instances in which a difference is said to exist when, in fact, none does exist (high sensitivity and specificity). To determine these characteristics for this new statistical test, we examined its diagnostic performance over a range of simulated situations, using methods analogous to those used by Pollack and Hsieh (6) and Metz and Kronman (7).

In order to calculate the specificity, 400 simulated analyses were performed for each of several combinations of underlying ROC areas and correlations. The tabulated distributions of the test statistic z obtained from these various simulations were for all practical purposes indistinguishable from Gaussian ones and had standard deviations acceptably close to 1. The false-positive rates were low, and close to what one should expect. Specifically, among the 4,800 trials (12 combinations each run 400 times) the average proportion of z values above 2.0 or below −2.0 (values often taken as indicating a statistically significant difference) was 5.1%, *i.e.*, a specificity of 94.9%. In a perfect Gaussian distribution, this would have been 4.6%, *i.e.*, a specificity of 95.4%.

Evaluation of the sensitivity (power) for this statistic required comparison of two modalities with different ROC areas. For this purpose, sets of 200 experiments were simulated from varying correlation coefficients. The performance was evaluated by tabulating the percentage of paired and unpaired tests indicating a significant difference. Over four combinations of correlations and baseline accuracies, one could project that the paired test would raise a 50% sensitivity (expected from an unpaired analysis) to 60–75%.

DISCUSSION

In this investigation we have described a method of comparing the areas under two ROC curves derived from the same sample of patients. Two immediate results are apparent. We have shown that the comparison can be made more sensitive if the investigator takes into account the smaller sampling variability of the difference in areas induced by studying each patient twice. Second, our data can be extrapolated to indicate the statistical economy that emerges from this kind of experimental design and analysis. We discuss these points in turn.

The larger the correlation between the areas, the more sensitive the paired z test will be. This observation may explain why a number of studies using

an unpaired z test that assumed the two areas were statistically independent failed to find a significant difference between the modalities. The degree of correlation expected between ROC areas obtained with different modalities varies considerably depending upon the types of modalities involved. For example, if the two images are obtained from the same machine with two different settings or if a radiologist reads a CT scan with and without extensive clinical history, high correlation can be expected. In this study involving different reconstruction algorithms with CT, the correlation between the paired ratings of abnormal phantoms was 0.60 and between paired ratings of normal phantoms was 0.39. We have observed similar results in a study of ours (8) involving the interpretation of CT studies of the head with and without extensive clinical history. On the other hand, when the only common denominator in the comparison is the patient, the correlations are likely to be weaker. For example, a study by Alderson *et al.*(9) comparing CT, ultrasound, and nuclear medicine imaging in the diagnosis of liver metastases found considerably lower rating-pair correlations (0.36 in abnormal patients and 0.28 in normal patients). Obviously, in the latter situation the gains from using a paired rather than an unpaired analysis are smaller.

Two other points must be made about correlation coefficients. First, in general we have noted that whatever the modalities under study, the ratings tend to be less correlated in the nondiseased patients than in the diseased patients. This suggests that in diagnostic imaging agreement tends to be greater if there is in fact underlying disease, and less if there is not. Second, if an investigator knew *a priori* that the correlations between the modalities under study were small, then an experimental design that did not involve pairing could be used, provided that it was no more difficult to separate (diagnose) the patients studied by one modality than it was to diagnose those studied by the other modality.

The statistical economy resulting from this new statistical test is large. Statistical economy relates to the question of how many more patients are required in an unpaired design then in a paired design to achieve the same sensitivity or statistical power. A comparison of Equations 1 and 2 provides an answer to this question. Each of the standard errors is inversely proportional to the square root of the sample size n. Also, the equations can be simplified by assuming that the standard errors of the two areas are equal; in this case, Equation 2 differs from Equation 1 only in the presence of the factor $(1 - r)$. When the sample sizes associated with the two techniques are arranged so that the paired and unpaired tests produce the same z value, then a simple algebraic identity emerges:

$$n_u = n_p/(1 - r)$$

or

$$n_p = (1 - r)n_u$$

where n_u and n_p are the numbers of patients per modality in the respective unpaired and paired designs[4]. For example, if r is anticipated to be roughly 0.3 and an unpaired design called for 100 patients per modality, then a paired design should require only 70 per modality. Thus the total number of images read would be 140 rather than 200. This efficiency is even more important if the limiting factor is the number of available patients with a proved outcome (rather than the number of images a reader can be expected to read), since the total of 140 paired images is obtained from just 70 patients, rather than from 200 patients in the unpaired design. The investigator must weigh very carefully the practical and statistical issues, keeping in mind that if one uses an unpaired design, one must establish (through case matching and/or random allocation) that the method of constructing two independent samples of subjects does not give one modality an inbuilt advantage.

The discussion thus far has centered on a rather restricted design where just one reader read the images generated by the two modalities being compared. The statistical test simply asked the question: if this one reader read an infinite rather than a finite number of images, would his/her accuracy be comparable in both modalities?[5] Clearly, a more general question is relevant: how do the modalities compare over many readers?

For the sake of completeness, we refer briefly to this problem of multiple readers and readings in each modality. This situation has been discussed extensively by Swets and Pickett (10); our main reasons for mentioning it here are to draw readers' attention to a very extensive treatment of the design and analysis of imaging experiments, and to point out that our method of obtaining r now allows the methods therein to be used with greater sensitivity. This is best appreciated by reproducing the formula that the authors give (Equation 2, Chapter 3) for the standard error of a difference between the value of an accuracy index (such as the area under an ROC curve) for one modality (averaged over l readers, each reading each image m times) and the value of the same accuracy index (again averaged over readers and readings) for a second modality. The expression involves three sources of variation: S_c^2, the variation in the index due to differences in mean difficulty of cases from case sample to case sample; S_{br}^2, between-reader variance due to differences in diagnostic capability from reader to reader; and S_{wr}^2, within-reader variance due to differences in an individual reader's diagnoses of the same case in repeated occasions. It also involves two correlation coefficients: r_c to denote the correlations introduced by using similar (or even the same) cases with both modalities and r_{br} to denote correlations between the accuracy index obtained by using matched (or possibly the same) readers. With this notation, the formula becomes

$$SE\ (difference) = \sqrt{2\{S_c^2(1 - r_c) + S_{br}^2(1 - r_{br})/l + S_{wr}^2/lm\}}$$

The authors describe fully *via* several worked examples how to evaluate each of these terms. They point out, however, that the estimation of the two components r_c and S_{wr}^2 creates problems. First, if $m = 1$, *i.e.*, if each image is read just once, then S_{br}^2 and S_{wr}^2 are not separable, and one is forced to overestimate the SE. The second, and more serious, problem is that if $m = 1$ and if one does not have a large number of cases, enough (for example) to split them into a number of subsamples and fit an ROC curve to each, one is unable to estimate r_c. In such cases, the authors explain that one has no alternative but to assume $r_c = 0$, thereby giving up any benefits attainable from case matching.

The method we have presented here means that if one uses the area under the ROC curve as an index of accuracy, one is not forced to assume $r_c = 0$. The quantity we have called r, which is obtainable *via* TABLE I from the area and from the correlations between ratings, is the same quantity r_{c-wr} mentioned in Equation 5, Chapter 4 of Swets and Pickett (8)[6]. The interested

[4] This simple relation allows the user to multiply the sample sizes in TABLE III of our first publication (4) by the appropriate $(1 - r)$ and use them for paired designs.

[5] One could also use the z test to compare two specific readers on one modality.

[6] If $m > 1$, one can correct the quantity r_{c-wr} (obtained from TABLE I) for the "attenuation" produced by S_{wr}^2, and estimate the "true" correlation r_c introduced by using similar (or the same) cases.

reader can consult that reference for full details on how it is used.

In summary, then, we have provided a method for estimating the correlation between the areas under two ROC curves derived from the same sample of patients and have shown how to use this correlation to perform a more sensitive comparison of the areas. Moreover, this provides an item that was previously only guessed at, or underestimated, in studies of multiple readers.

Acknowledgments: We are indebted to Richard Swensson and Philip Judy for providing the CT phantoms and rereading results. We are also indebted to Charles Metz for many helpful discussions in the course of this investigation. Irene McCammon typed the manuscript.

Department of Epidemiology and Health
McGill University
3775 University Street
Montreal, Quebec
Canada H3A 2B4

References

1. Green D, Swets JA. Signal detection theory and psychophysics. New York: John Wiley and Sons, 1966.
2. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978; 8:283–298.
3. Swets JA. ROC analysis applied to the evaluation of medical imaging techniques. Invest Radiol 1979; 14:109–121.
4. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29–36.
5. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals—rating-method data. J Math Psych 1969; 6:487–496.
6. Pollack I, Hsieh R. Sampling variability of the area under the ROC-curve and of d'_e. Psych Bull 1969; 71:161–173.
7. Metz CE, Kronman HB. Statistical significance tests for binormal ROC curves. J Math Psych 1980; 22:218–243.
8. McNeil BJ, Hanley JA, Funkenstein HH, Wallman J. Paired receiver operating characteristic curves and the effect of history on radiographic interpretation: CT of the head as a case study. Radiology 1983; in press.
9. Alderson PO, Adams DF, McNeil BJ, et al. Computed tomography, ultrasound, and scintigraphy of the liver in patients with colon or breast carcinoma: A prospective study. Radiology 1983; in press.
10. Swets JA, Pickett RM. Evaluation of diagnostic systems. New York: Academic Press, 1982.

APPENDIX

Ratings given to two images of each of 112 phantoms, together with calculations of z test to test whether one modality subtended a greater ROC area than the other:

(*a*) Basic data:

Rating* with Modality 1	Rating* with Modality 2 — Normal Phantoms							Abnormal Phantoms						
	1	2	3	4	5	6	Total	1	2	3	4	5	6	Total
1	9	3	—	—	—	—	12	—	—	1	—	—	—	1
2	17	9	2	—	—	—	28	1	—	2	—	—	—	3
3	3	4	1	—	—	—	8	1	1	1	3	—	—	6
4	1	2	2	1	—	—	6	1	1	1	9	1	—	13
5	1	1	—	2	—	—	4	—	—	—	7	10	5	22
6	—	—	—	—	—	—	—	—	—	—	—	4	5	9
Total	31	19	5	3	—	—	58	3	2	5	19	15	10	54

* Ratings: from 1 = Definitely Normal to 6 = Definitely Abnormal.

(*b*) Correlation between ratings (Kendall tau):

modality 1 *vs.* modality 2 $r_A = 0.60$

modality 1 *vs.* modality 2 $r_N = 0.39$

average correlation = 0.50

(*c*) ROC analysis:

Modality	Rating 1	2	3	4	5	6	Slope	Intercept	z_A	Area	Standard Error (Area)
Modality 1											
Normal	12	28	8	6	4	—	0.945	1.72	1.25	0.8945	0.030
Abnormal	1	3	6	13	22	9					
Modality 2											
Normal	31	19	5	3	—	—	0.467	1.70	1.54	0.9382	0.026
Abnormal	3	2	5	19	15	10					

Difference in areas = 0.0437.
Average area = 0.9146.
Correlation between areas = 0.40.

(*d*) Test statistic: $z = 0.0437 / \sqrt{0.030^2 + 0.026^2 - 2(0.040)(0.030)(0.026)} = 1.41$