[Reprinted from RADIOLOGY, Vol. 143, No. 1, Pages 29–36, April, 1982.]

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

The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve[1]

A representation and interpretation of the area under a receiver operating characteristic (ROC) curve obtained by the "rating" method, or by mathematical predictions based on patient characteristics, is presented. It is shown that in such a setting the area represents the probability that a randomly chosen diseased subject is (correctly) rated or ranked with greater suspicion than a randomly chosen nondiseased subject. Moreover, this probability of a correct ranking is the same quantity that is estimated by the already well-studied nonparametric Wilcoxon statistic. These two relationships are exploited to (*a*) provide rapid closed-form expressions for the approximate magnitude of the sampling variability, *i.e.*, standard error that one uses to accompany the area under a smoothed ROC curve, (*b*) guide in determining the size of the sample required to provide a sufficiently reliable estimate of this area, and (*c*) determine how large sample sizes should be to ensure that one can statistically detect differences in the accuracy of diagnostic techniques.

Index terms: Receiver operating characteristic curve (ROC)

Radiology 143: 29–36, April 1982

RECEIVER operating characteristic (ROC) curves (1–3) have become increasingly popular over the past few years as the radiologic community has become concerned with the measurement of the information content of a variety of imaging systems (4, 5). In addition, ROC curves are being used to judge the discrimination ability of various statistical methods that combine various clues, test results, etc. for predictive purposes (*e.g.*, to determine whether a patient will need hospitalization or will benefit from treatment). In most cases, however, ROC curves have been plotted and evaluated qualitatively with relatively little attention paid to their statistical characteristics. This has occurred for a number of reasons.

First, the most common quantitative index describing an ROC curve is the area under it, and there has not been a full description in the radiologic literature of an intuitive meaning of this area. Second, most quantitative measures used to describe an ROC curve are derived assuming that the varying degrees of normality/abnormality seen in the images can be represented by two separate but usually overlapping Gaussian distributions, one for the diseased group and one for the nondiseased group. This has led a number of investigators to question the validity of the ROC approach altogether. Third, iterative numerical methods rather than closed-form expressions are required to estimate the standard error of these quantitative indices; operationally, these methods are cumbersome and require a specialized computer program. Fourth, there has been no method shown to date to indicate the sample size necessary to ensure a specified degree of statistical precision for a particular quantitative index.

In this paper, we will elaborate on the meaning of the area under an ROC curve and, using the links between it and other, better known statistical concepts, we will develop analytic techniques for eliciting its statistical properties.

METHODS

Indices Used to Summarize ROC Curves

A large number of theoretically based measures has been proposed to reduce an entire ROC curve to a single quantitative index of diagnostic accuracy; all of these measures have been rooted in the assumption that the functional form of the ROC curve is the same as that implied by supposing that the underlying distributions for normal and abnormal groups are Gaussian (4). When an ROC curve, plotted on double probability paper, is fitted by eye to a straight line or when the ROC points are submitted to an iterative maximum likelihood estimation program, two parameters, one a difference of

[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 (B.J.M.). Received March 18, 1981; accepted and revision requested July 21; revision received Dec. 15.

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

See also the articles by Hessel *et al.* (pp. 129–133) and Abrams *et al.* (pp. 121–128) in this issue. cd

TABLE I: Rating of 109 CT Images

True Disease Status	Rating: Definitely Normal (1)	Probably Normal (2)	Questionable (3)	Probably Abnormal (4)	Definitely Abnormal (5)	Total
Normal	33	6	6	11	2	$n_N = 58$
Abnormal	3	2	2	11	33	$n_A = 51$
Totals	36	8	8	22	35	109

means and the other a ratio of variances, are obtained. From these, a number of indices can be calculated, the most popular being an estimate of the area under the fitted smooth curve (4). This index, denoted $A(z)$ to symbolize its Gaussian underpinnings, varies from 0.5 (no apparent accuracy) to 1.0 (perfect accuracy) as the ROC curve moves towards the left and top boundaries of the ROC graph.[2] When one fits the two parameters by maximum likelihood rather than by eye, one also obtains their standard errors, thereby allowing the area derived from the two parameters to be also accompanied by a standard error. This can be used to construct confidence intervals and to perform statistical tests of significance.

The Meaning of the Area under an ROC Curve

A precise meaning of the area under an ROC curve in terms of the result of a signal detection experiment employing the two-alternative forced choice (2AFC) technique has been known for some time. In this system, Green and Swets (6) showed that the area under the curve corresponds to the probability of correctly identifying which of the two stimuli is "noise" and which is "signal plus noise." In medical imaging studies, the more economical rating method is generally used: images from diseased and nondiseased subjects are thoroughly mixed, then presented in this random order to a reader who is asked to rate each on a discrete ordinal scale ranging from definitely normal to definitely abnormal. Very often a five-category scale is used. Although the points required to produce the ROC curve are obtained in a more indirect way, *i.e.*, by successively considering broader and broader categories of abnormal (*e.g.*, category 5 alone, categories 5 plus 4, categories 5 plus 4 plus 3), the important point is that on a *conceptual* level, and thus from a statistical viewpoint, the area under the curve obtained from a rating experiment has the same *meaning* as it has when it is derived from a 2AFC experiment. As we will explain below, the ROC area obtained from a rating experiment can be viewed, at least conceptually, in the same way as the area obtained from a 2AFC experiment. Basically, when an investigator calculates the area under the ROC curve directly from a rating experiment, he is in fact, or at least in mathematical fact, reconstructing random pairs of images, one from a diseased subject and one from a normal subject, and using the reader's separate ratings of these two images to simulate what the reader would have decided if these two images had in fact been presented *together* as a pair in a 2AFC experiment. Indeed, this mathematical equivalence (equivalent in the sense that the two areas are *measuring* the same quantity, even if the two curves are constructed differently) has also been verified empirically in a recognition memory experiment by Green and Moses (7).

More important, however, was the more recent recognition by Bamber (8) that this "probability of correctly ranking a (normal, abnormal) pair" is intimately connected with the quantity calculated in the Wilcoxon or Mann-Whitney statistical test. We now elaborate on this relationship in two ways. First, we show empirically that if one performs a Wilcoxon test on the ratings given to the images from the normal and diseased subjects, one obtains the same quantity as that obtained by calculating the area under the corresponding ROC curve using the trapezoidal rule. (If in fact the ratings are on a continuous scale, the area obtained by the Wilcoxon statistic or by the trapezoidal rule will be virtually identical to any smoothed area.) Second, and more important, we show how the statistical properties of the Wilcoxon statistic can be used to predict the statistical properties of the area under an ROC curve.

[2] An area can also be calculated by the trapezoidal rule; the area obtained in this way has been designated $P(A)$. As is seen in Figure 1, $P(A)$ is smaller than the area under any smooth curve, and is somewhat more sensitive to the location and spread of the points defining the curve than is the area $A(z)$ calculated as the smooth Gaussian estimate.

RESULTS

I. A Three-Way Equivalence

To amplify the three-way equivalence between the area under an ROC curve, the probability of a correct ranking of a (normal, abnormal) pair, and the Wilcoxon statistic, we present it as two pairwise relationships:

A. The area under the ROC curve measures the probability, denoted by θ, that in randomly paired normal and abnormal images, the perceived abnormality of the two images will allow them to be correctly identified.

B. The Wilcoxon statistic also measures this probability θ that randomly chosen normal and abnormal images will be correctly ranked.

We now deal with *A* and *B* in turn.

II. Mathematical Restatement of Relationship A

We make an implicit assumption that the sensory information conveyed by a radiographic image can be quantified by and ordered on a one-dimensional scale represented by x, with low values of x favoring the decision to call the image normal and high values favoring the decision to call it abnormal. The distributions of x values for randomly selected abnormal images, denoted by x_A, and those for normal images, denoted by x_N, will overlap; the x_A distribution will be centered to the right of the x_N one. In a rating experiment, the degree of suspicion, x, will actually be reported on an ordered categorical scale. TABLE I presents illustrative data showing how a single reader rated the computed tomographic (CT) images obtained in a sample of 109 patients with neurological problems. As expected, the x_A and x_N distributions overlap (*i.e.*, some nondiseased patients had abnormal readings and some diseased patients had normal readings). Thus, if the images from a randomly chosen normal and a randomly chosen abnormal case were paired, there would be less than a 100% probability that the sensory informa-

TABLE II: Computation of W and Its Standard Error

Row	Contents	Column (Rating) x = 1	x = 2	x = 3	x = 4	x = 5	Total	Remarks
1	Number of normals rated x	33	6	6	11	2	$58 = n_N$	Obtained from TABLE I
2	Number of abnormals rated $>x$	48	46	44	33	0		Obtained from 3 by successive substractions from $n_A = 51$
3	Number of abnormals rated x	3	2	2	11	33	$51 = n_A$	Obtained from TABLE I
4	Number of normals rated $<x$	0	33	39	45	56		Obtained from 1 by successive additions to 0
5	$(1) \times (2) + \frac{1}{2} \times (1) \times (3)$	$1{,}633\frac{1}{2}$	282	270	$423\frac{1}{2}$	33	2,642	W = Total (5) ÷ $(n_N \cdot n_A) = 0.893$
6	$(3) \times [(4)^2 + (4) \times (1) + \frac{1}{3} \times (1)^2]$	1,089	2,598	3,534	$28{,}163\frac{2}{3}$	107,228	$142{,}612\frac{2}{3}$	Q_2 = Total (6) ÷ $(n_A \cdot n_N^2) = 0.8313$
7	$(1) \times [(2)^2 + (2) \times (3) + \frac{1}{3} \times (3)^2]$	80,883	13,256	12,152	$16{,}415\frac{2}{3}$	726	$123{,}432\frac{2}{3}$	Q_1 = Total (7) ÷ $(n_N \cdot n_A^2) = 0.8182$

$W = \hat{\theta}$ = total (5) ÷ $(n_N \cdot n_A) = 2{,}642 \div (58 \cdot 51) = 0.893 = 89.3\%$

$$SE(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta}) + (n_A - 1)(Q_1 - \hat{\theta}^2) + (n_N - 1)(Q_2 - \hat{\theta}^2)}{n_A \cdot n_N}} = \sqrt{\frac{0.099551 + 0.037764 + 1.926686}{51 \cdot 59}} = 0.032 = 3.2\%$$

tion or, for want of a more precise term, the degree of suspicion, x_A, which one obtains from the abnormal image would in fact be greater than the corresponding x_N obtained from the normal image. As a statistical shorthand, we refer to this probability as $\theta = Prob(x_A > x_N)$. Then Green and Swets' result says that if we assume for the moment that we have an infinite sample of patients and that a reader is capable of reporting using the entire x continuum rather than only a finite number of category ratings, the area under the curve and the probability of a correct ranking are equal, or

"True" area under ROC curve = θ
= $Prob(x_A > x_N)$

From our viewpoint, the most important feature of the proof, which depends on integral calculus, is that it makes *no assumptions about the form of the x_A and x_N distributions.* Thus, the area under the curve can be thought of simply as measuring the probability of a correct ranking of a (normal, abnormal) pair; neither the nature (symmetric *vs.* right tailed *vs.* left tailed) nor the exact distributional form (Gaussian *vs.* negative exponential *vs.* gamma) need be explicitly quantified. However, there are practical reasons for the use of distributional assumptions: (*a*) the maximum of five or six rating categories a reader is capable of using means that the trapezoidal rule will tend to underestimate the area under what is in reality a smooth ROC curve; (*b*) the criteria for fitting a smooth curve are more easily agreed upon; and (*c*) the investigator is often interested in other aspects of the ROC curve, such as the trade-offs between sensitivity and specificity. Moreover, the extensive data from psychophysical and medical imaging studies tend to agree reasonably well with the ROC curve form implied by two Gaussian distributions.

For investigators interested in using ROC curves to describe the discrimination achieved with scores or probabilities constructed on a continuous scale from regression-type equations based on a patient's presenting symptomatology, the distributions of these scores or probabilities will not necessarily conform to Gaussian distributions. However, in this situation, the more continuous nature of the score or probability scale means that the empirical ROC curve will also be much smoother, and complex curve fitting will probably not be necessary.

III. Relationship B: The Wilcoxon Statistic and the Probability of Correct Pairwise Rankings

The Wilcoxon statistic, W, is usually computed to test whether the levels of some quantitative variable x in one population (A) tend to be greater than in a second population (N), without actually assuming how the x's are distributed in the two populations. The null hypothesis is that x is not a useful discriminator, *i.e.*, that an x value from an individual from A is just as likely to be smaller than an x value from an individual from N as it is to be greater than it, or that $\theta = Prob(x_A > x_N) = 0.5$. For x to be a good discriminator, this probability has to be much closer to unity. With a sample of size n_A from A and n_N from N, the procedure, *at least conceptually*, consists of making all $n_A \cdot n_N$ possible comparisons between the n_A sample x_A's and the n_N sample x_N's, scoring each comparison according to the rule

$$S(x_A, x_N) = \begin{cases} 1 \text{ if } x_A > x_N \\ \frac{1}{2} \text{ if } x_A = x_N \\ 0 \text{ if } x_A < x_N \end{cases} \quad \text{(discrete data only)}$$

and averaging the S's over the $n_A \cdot n_N$ comparisons, *i.e.*,

$$W = \frac{1}{n_A \cdot n_N} \sum_{1}^{n_A} \sum_{1}^{n_N} S(x_A, x_N)$$

In practice, the computation can be performed by a much faster method to be described below in section IV. Also, since the test is based on yes/no comparisons, W does not depend on the actual values of the x's but only on their rankings.

Relationship B should now be obvious from the very formulation of W, since it actually makes the kind of comparisons mentioned when describing $\theta = Prob(x_A > x_N)$. Since each comparison is scored as 1, $\frac{1}{2}$, or 0, the average score W lies between 0 and 1 and reflects, as it should, what proportion of the x_A's are greater than what proportion of the x_N's. Obviously not all $n_A \cdot n_N$ comparisons are independent; including them all is merely a convenience, and the standard error of W takes these interrelated comparisons into account.

Although, as stated above, W is usually used to test the (null) hypoth-

esis that variable x cannot be used to discriminate between A and N (*i.e.*, that θ equals 0.5), its behavior when θ exceeds 0.5 (*i.e.*, when x is actually of discriminatory value) is also well established. In that regard, for our purposes, the most important characteristic is its standard error, since our main interest is in quantifying how variable W (or its new alias, the area under the ROC curve) will be in different similar-sized samples. When $\theta > 0.5$, W is no longer nonparametric; its standard error, SE(W), depends on two distribution-specific quantities, Q_1 and Q_2, which have the following interpretation:

Q_1 = *Prob* (two randomly chosen abnormal images will both be ranked with greater suspicion than a randomly chosen normal image)

Q_2 = Prob (one randomly chosen abnormal image will be ranked with greater suspicion than two randomly chosen normal images)

If we assume for the moment (as Green and Swets do in their proof regarding θ and the area under the ROC curve) that the ratings are on a scale that is sufficiently continuous that it does not produce "ties," then SE(W), or equivalently *SE*(area underneath empirical ROC curve), can be shown (8) to be

$$\mathrm{SE}(W) = \sqrt{\frac{\theta(1-\theta) + (n_A - 1)(Q_1 - \theta^2) + (n_N - 1)(Q_2 - \theta^2)}{n_A\, n_N}} \quad (1)$$

The quantity W can be thought of as an estimate of θ, the "true" area under the curve, *i.e.*, the area one would obtain with an infinite sample and a continuous rating scale. In the rating category situation, of course, W will tend to underestimate θ, but Formula 1 will be useful nevertheless.

We now go on in the following section to calculate the Wilcoxon statistic for the data in TABLE I and to show that it does indeed correspond to an area under the ROC curve, albeit the area found by the trapezoidal rule. We also carry out the computations required to estimate directly from the data (*i.e.*, without any distributional assumptions) the two quantities Q_1 and Q_2. With these, and using W as an estimate of θ, we then use Formula 1 to estimate a standard error for what is in this case a somewhat biased area under the curve. As will become evident in subsequent sections, there is a second method, requiring fewer computations, for estimating Q_1 and Q_2 for use in Formula 1. We give both methods for the sake of completeness.

IV. W and SE(W) – Calculated without Distributional Assumptions

We illustrate the calculations using the data from TABLE I. Since the Wilcoxon statistic is based on pairwise comparisons, the specific values 1 through 5 that we have applied to the five rating categories are to be thought of simply as rankings. The computations can be conveniently carried out according to the scheme shown in TABLE II.[3] Rows 3 and 1 are taken directly from TABLE I, while rows 2 and 4 are derived from 3 and 1 by successive deletion and cumulation respectively. The quantity W can be computed in row 5 by using the entries in rows 1, 2, and 3; the SE requires calculation of the two intermediate probabilities Q_1 and Q_2 (see rows 6 and 7 for details), which are then used to compute an estimate of SE(W) from Formula 1.

TABLE II shows the detailed calculation of W and its standard error. The $W = \hat{\theta} = 0.893 = 89.3\%$ derived in this way agrees exactly with the area under the ROC curve calculated by the trapezoidal rate. By way of comparison, the area under the smooth Gaussian-based ROC curve fitted by the maximum likelihood technique of Dorfman and Alf (9) is 0.911 or 91.1%; the area under the smooth ROC curve derived from the parameters of a straight-line fit to the ROC plotted on double probability paper (see Swets [4], pp. 114–115) is 0.905 or 90.5%. The slightly lower estimate provided by W, or equivalently by the trapezoidal rule, merely reflects the fact that the rating scale does not have infinitely fine "grain." In another context, where the ratings might have been expressed on a more continuous scale (*i.e.*, without ties), the two would agree even better. What is more important is that TABLE II, using 89.3% as its estimate of θ, produces a standard error of 3.2%, compared with the SE of 2.96% predicted by the maximum likelihood parametric technique. Although this 3.2% appears to be a little high, it is not greatly so; moreover it is on the conservative side, and guards against the possibility that the distributional assumptions that produced the SE of 2.96% are not entirely justified.

The Wilcoxon statistic now provides a useful tool for the researcher who does not have access to the computer program described above, but who still wishes to use as an index of discrimination the area under a smooth ROC curve and to accompany it by an approximate standard error. He can use the parameters of the straight-line fit to produce the smooth ROC curve and the area under it and he can use SE(W) as a slightly conservative estimate of the SE of this smoothed area. In our example, simply by plotting the data on double probability paper, estimating a slope and intercept from a straight line fit, calculating from these a quantity that Swets (4) calls $z(A)$, and looking $z(A)$ up in Gaussian probability tables, one obtains a smoothed area of 90.5%, which is only 0.6% (in absolute terms) or 0.66% (in relative terms) different from the 91.1% obtained by a full maximum likelihood fit. By an equally straightforward approach, one can use the calculations in TABLE II to come reasonably close (3.2% compared with 2.96%) to the standard error produced by the maximum likelihood approach. As we will see later in section V, we will be able to improve even further on predicting the SE produced by this method.

All of the discussion thus far has centered on computing an area and its SE from observed data; we now turn to the commonly asked question, "How big a sample do I need?"

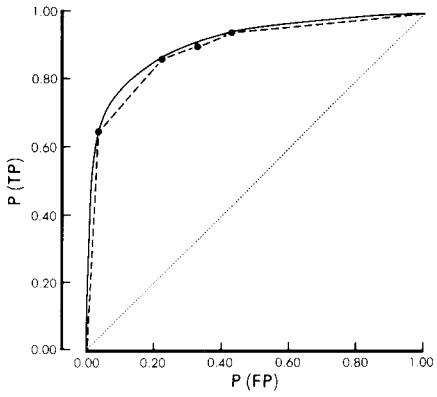
V. Planning Sample Sizes

Perhaps the more important use of the three-way equivalence between the area under the curve, the probability of a correct ranking, and the Wilcoxon statistic is in pre-experiment calculations. At this stage, one is often asked: "We wish to use θ, the area under the ROC curve, to describe the performance of an imaging system, and we would like to have this index accompanied by a measure of its uncertainty, *i.e.*, of the fluctuations in the index caused by the random sampling of cases. How many cases must be studied to ensure an acceptable level of precision?" This is equivalent to asking how large n_A and n_N must be so that the resulting SE is of a reasonably small magnitude, and the resulting confidence interval is correspondingly narrow.

In addition to the quantities n_A and

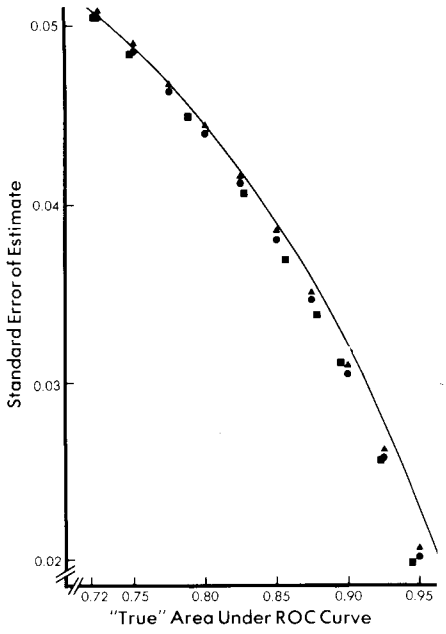
[3] The rationale for this scheme is conceptually simple but lengthy. Details can be obtained from the authors.

Figure 1



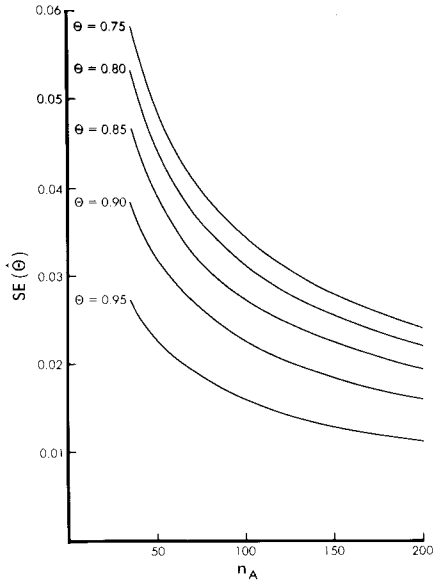
ROC curve for data in TABLE I. Dashed line = empirical curve; solid line = smoothed (Gaussian-based) curve; dotted diagonal line = no discrimination.

Figure 2



Anticipated standard error for area under ROC curve generated from different underlying distributions. Circles = Gaussian with variance ratio 1:1.5; triangles = Gaussian with variance ratio 1:0.5; squares = gamma with various degrees of freedom; solid line = negative exponential.

Figure 3



Standard error (SE) for estimated area under ROC curve ($\hat{\theta}$) in relation to sample size (n_A = number of abnormal cases) and true area under ROC curve (θ). Calculations assume an equal number (n_N) of normal cases.

n_N, Formula 1 contains three other parameters—θ, Q_1, and Q_2. While one can use anticipated values of the true area θ, the quantities Q_1 and Q_2 are complex functions of the underlying distributions for x_A and x_N. Fortunately, for any specified pair of distributions Formula 1 is almost entirely determined by θ, and only very slightly influenced by any further parameters of the distributions. As an example, Figure 2 shows how little the relationship between $SE(\hat{\theta})$ and θ changes as one postulates underlying Gaussian, gamma, or negative exponential distributions. Moreover, it seems that in the range of interest (areas of 80% or more), the negative exponential model yields SE's that are slightly more conservative than the other models considered. This is especially fortunate since under this model, the quantities Q_1 and Q_2 can be expressed as simple functions of θ, *i.e.*,

$$Q_1 = \theta \div (2 - \theta)$$
$$Q_2 = 2\theta^2 \div (1 + \theta) \qquad (2)$$

When these expressions are substituted into Formula 1, we obtain the SE to be expected[4] at any anticipated level of performance θ, and can vary the n's until $SE(\hat{\theta})$ is sufficiently small. For example, consider an experiment where the diagnostic accuracy is expected to be in the neighborhood of θ = 85%. Then $Q_1 = 0.85 \div 1.15 = 0.7391$ and $Q_2 = 2(0.85)^2 \div 1.85 = 0.7811$. Then with $n_A = n_N = 40$, Formula 1 predicts an SE of 4.37%, while $n_A = n_N = 60$ will reduce the SE to 3.56%. Figure 3 gives $SE(\theta)$ for various sample sizes and various anticipated θ's. A number of points should be noted:

1. As one expects with SE's, they vary inversely with $\sqrt{n}$, so that, for example, one must quadruple the sample size to halve the SE.
2. The SE's are smallest for very high θ's, *i.e.*, those close to 1.
3. The SE's are slightly more conservative than those obtained under the Gaussian model (Fig. 2).
4. One must resort to Formulas 1 and 2 to calculate SE's when the number of normal cases n_N does not equal the number of abnormal cases n_A.

In passing, it is of interest to compare the $SE(\hat{\theta})$ of approximately 3% obtained from our previously mentioned *rating* experiment example with what would have been obtained from an *actual* 2AFC experiment. In the latter, one would estimate θ simply by calculating the fraction $\hat{\theta}$ of m pairs of images where the normal and abnormal image were correctly identified, and one would accompany this estimate of θ with a standard error, based on the binomial distribution, of $\sqrt{\theta(1-\theta)/m}$. To achieve a standard error of 0.03 or 3%, and assuming that in fact $\hat{\theta}$ in the 2AFC experiment turned out to be 0.9 or 90%, one would need m = 100 *pairs* of (normal, abnormal) images, a considerably greater number of *images* ($2m$ or 200) than the $n_N + n_A = 109$ used in the rating experiment. For purposes of precision (low SE) in measuring θ, we can think of the $2m$ images (m pairs) as the statistical equivalent of our 109. Thus, even if one were interested only in estimating the *overall* percentage of patients that would be correctly classified by a medical imaging system, the rating method would be the method of choice. In addition to being more economical (*i.e.*, requiring fewer patients), it yields valuable data on the two separate components of diagnostic accuracy, namely, sensitivity and specificity.

For these and other reasons, the 2AFC method is not a serious competitor to the rating method in medical imaging experiments. However, the formal statistical ties between the two methods do help in seeing how to use an $SE(\hat{\theta})$ derived from a rating experi-

[4] This distribution-based approach can be used not only to *project a future SE* but also, as mentioned at the end of section III, to provide a second method of calculating an SE for *observed* data. For example, the data discussed in section IV produced an area of 0.905 or 90.5%. Using this for θ in Equation 2, we obtain $Q_1 = 0.8265$ and $Q_2 = 0.8599$. Substituting these three values into Equation 1 gives an SE of 0.0307 or 3.07%, even closer to the 2.96% produced by maximum likelihood estimation.

TABLE III: Number of Normal and Abnormal Subjects Required to Provide a Probability of 80%, 90%, or 95% of Detecting Various Differences between the Areas θ_1 and θ_2 under Two ROC Curves (Using a One-Sided Test of Significance with $p = 0.05$)

	θ_2									
θ_1	0.750	0.775	0.800	0.825	0.850	0.875	0.900	0.925	0.950	0.975
0.700	652	286	158	100	68	49	37	28	22	18
	897	392	216	135	92	66	49	38	29	23
	1131	493	271	169	115	82	61	46	36	29
0.725		610	267	148	93	63	45	34	26	20
		839	366	201	126	85	61	45	34	27
		1057	459	252	157	106	75	55	42	33
0.750			565	246	136	85	58	41	31	23
			776	337	185	115	77	55	41	31
			976	423	231	143	96	68	50	38
0.775				516	224	123	77	52	37	27
				707	306	167	104	69	49	36
				889	383	209	129	86	60	44
0.800					463	201	110	68	46	33
					634	273	149	92	61	43
					797	342	185	113	75	53
0.825						408	176	96	59	40
						557	239	129	79	52
						699	298	160	97	64
0.850							350	150	81	50
							477	203	108	66
							597	252	134	81
0.875								290	123	66
								393	165	87
								491	205	107
0.900								960	228	96
								1314	308	127
								1648	383	156
0.925									710	165
									966	220
									1209	272
0.950										457
										615
										765

80% probability = top number; 90% probability = middle number; 95% probability = bottom number.

ment to construct a confidence interval on θ. As one might expect, and indeed as we have found by repeatedly simulating data from two overlapping Gaussian distributions, constructing an ROC curve, and deriving the area $\hat{\theta}$ under the curve, the distribution of the $\hat{\theta}$'s we obtained is not entirely symmetric, but is instead somewhat skewed towards $\hat{\theta} = 0.5$. This skewness is more marked as the "true" θ approaches 1, and as the expected number of "misclassified pairs" $[m(1 - \theta)]$ falls below 5; this is identical to what occurs with the binomial distribution and a success probability close to unity. In such cases, one usually resorts to an exact (asymmetric) confidence interval for θ, rather than using the approximate (symmetric) one of ± 1.645 SE($\hat{\theta}$), 1.96 SE($\hat{\theta}$), . . . , provided by the normal distribution. In the example here $m(1 - \theta) = 10$ is considerably greater than the rule of thumb of 5; thus, the symmetric 95% confidence interval of $90.5\% \pm 1.96(3.07)$ or (84.5%, 96.5%) will be reasonably correct, compared with the exact, slightly asymmetric interval of (82.4%, 95.1%) obtained by consulting charted confidence limits for binomial sampling (10) with $\hat{\theta} = 0.905$ and $m = 100$.

Finally, we consider the question of obtaining sufficiently large sample ranges when one wishes to examine the difference between two areas, so that *if an important difference in performance exists*, it will be unlikely to go undetected in a test of significance.

VI. Detecting Differences between Areas under Two ROC Curves

Again, knowing in advance the approximate SE's that are likely to accompany an estimate of θ, we can calculate how many cases must be studied so that a comparison of two imaging systems will have any given degree of statistical power. This power or "statistical sensitivity" depends on how small the probabilities α and β of committing a type I or type II error are. Typically, one seeks a power $(100 - \beta)$ of 85% or 90% so that *if a specified difference exists*, it is 85% or 90% certain to be reflected in samples that will be declared "statistically different." Traditionally, one uses a type I error probability or α of 0.05 (5%) as the criterion for a significant difference.

TABLE III gives the numbers of normal and abnormal cases required for each ROC curve to have an 80%, 90%, or 95% assurance that various real differences δ between two areas, θ_1 and θ_2, will indeed result in sample curves showing a statistically significant dif-

ference ($p < 0.05$, one-sided). The calculations are based on an adaptation of the sample size formula given by Colton (11), namely,

$$n = \left[\frac{Z_\alpha \sqrt{2\,V_1} + Z_\beta \sqrt{V_1 + V_2}}{\delta}\right]^2 \quad (3)$$

where in our case

$Z_\alpha = 1.645$, for a 5% one-sided test of significance

$Z_\beta = 0.84$, 1.28, or 1.645 for 80%, 90%, or 95% power

$\delta = \theta_2 - \theta_1$

$V_1 = Q_1 + Q_2 - 2\theta_1^2$ (Q_1 and Q_2 obtained using θ_1 in Equation 2)

$V_2 = Q_1 + Q_2 - 2\theta_2^2$ (Q_1 and Q_2 obtained using θ_2 in Equation 2)

As an example, it shows that if indeed the "true" areas (*i.e.*, those that would be achieved with infinite populations) were 82.5% and 90.0%, one would need to plan on a sample of 176 normal subjects and 176 abnormal subjects for each curve to have a high assurance (*i.e.*, an 80% probability) that the test of significance on the samples would yield a statistically significant difference. Larger sample sizes allow one to detect smaller differences or have a greater assurance of detecting the same size difference, while smaller ones give less statistical power.

These considerations are not necessarily binding; after all, they are designed with the pessimistic attitude that the sampling will be "unkind" and apt to mask important differences. However, they do show that if a small study failed to show a statistically significant difference, there is a real possibility of a type II error.[5] On the other hand, if the "no difference" persists in spite of adequate sample sizes such as those shown in TABLE III, one can reasonably conclude that the stated differences do not in fact exist. Finally, TABLE III shows that a difference of 10% is more easily detected if it is a difference between 80% and 90% than if it is a difference between 70% and 80%.

DISCUSSION

The advent of new competitive imaging modalities (CT, ultrasound, nuclear medicine) for the same diagnostic problem has led to the performance of many studies involving comparisons of the information obtained from these imaging techniques. Many of these comparisons have used ROC curves in their analysis. However, it has become clear that an intuitive understanding of the statistical techniques proposed for comparing modalities with ROC curves is lacking. Also, there are special problems associated with the application of these techniques to medical problems and their associated small and usually heterogeneous data bases. Thus, we undertook this investigation, in part to aid our intuition in this area but more importantly to provide a firmer statistical basis for work in this field.

The intuitive results that have been shown in this paper are actually quite helpful. Basically, the results show that in the rating method, conventionally employed for analyzing imaging modalities using the ROC approach, the area under the ROC curve represents the probability that a random pair of normal and abnormal images will be correctly ranked as to their disease state. (We emphasize here that this probability of a correct ranking only conveys the intrinsic potential for discrimination with sensitivity and specificity weighted equally; other external [decision] factors that influence diagnostic performance include the *real* mixture [no longer 1:1] of diseased and nondiseased patients and the relative costs of the two types of diagnostic errors. However, as with other psychophysical processes, it is important to separate intrinsic discriminatory qualities of the imaging modality as much as possible from decision issues.)

Second, the combination of a graphic method for obtaining a smoothed area and a computational formula for its standard error now means that the investigator can obtain almost the full benefit of a parametric maximum likelihood without actual recourse to a large computer. Admittedly, the standard error may be slightly inexact, but it seems a small sacrifice.

Third, there is an increasing popularity of medical predictions using scores and probabilities derived from techniques such as discriminant analysis and logistic regression. Although these provide a scale that is much more "continuous" than the rating categories we have described above, the distributions of these composite indices may not be suitable for ROC analyses based on the binormal assumptions. However, the Wilcoxon statistic will now be a more bias-free method of estimating the "true" area θ under the ROC curve, and its standard error will continue to be valid, since it can be calculated directly from the data (as in TABLE II). Moreover, our calculations shown in Figure 2 suggest that it can largely be predicted from θ alone, even if the underlying distributions differ considerably from Gaussian.

Fourth and most important, the major contribution of this paper is the use of the Wilcoxon statistic for estimation of sample sizes and power calculations. Estimation of sample sizes is critically important. Generally, when a new imaging modality becomes available, a pilot study will provide some estimate of its approximate sensitivity and specificity in relationship to existing modalities. With some simplifying assumptions, the areas shown in TABLE III can be estimated. These data make it possible to calculate the approximate sample sizes necessary to detect differences between two areas (and hence their imaging modalities) with specific degrees of certainty. These numbers will let investigators know whether one institution can likely perform a research study alone, whether multiple institutions are required, or whether in fact so many institutions are required that the cost of obtaining the data is not worth the information obtained.

Two caveats are necessary in interpreting the results of this study. It should be noted that the sample size calculations in TABLE III are based on selecting a separate set of normal and abnormal subjects for each of the two experimental conditions being compared. These numbers are not suitable for situations when the two modalities are examining the same sets of cases or when one reader is evaluating the same set of cases under different conditions (*e.g.*, with and without history, with and without varying levels of contrast). Methods to deal with this latter situation are the subject of a subsequent article (12). Basically, they are based on the principle that when paired observations are available, the use of paired statistical analyses provides a much more powerful test than is obtained when paired observations are treated by statistical tests for independent samples. The difference between paired and unpaired t-tests is a familiar example of this problem.

Second, as mentioned at the outset, the *area* under an ROC curve is a parameter used to quantify in a single numerical value the overall *location* of an ROC curve relative to the (noninformative) diagonal. If one wishes to

[5] In fact, knowing the n's that were studied, one can solve Equation 3 for Z_β to find out how high the probability β of a type II error really was.

test statistically whether two *curves* are different (they could still subtend the same area but cross each other), then one must resort to a bivariate statistical test (13). This test simultaneously compares the two parameter values—a the difference between the x_A and x_N distributions and b the ratio of their variances, which describe one ROC curve with the corresponding values for the second curve. The test requires that one supply estimates of a and b, as well as estimates of their variances and covariance. These estimates are provided by the maximum likelihood estimation technique described by Dorfman and Alf (9).

Acknowledgments: We wish to thank Colin Begg, Charles Metz, and Stanley Shapiro for helpful suggestions and Irene McCammon for preparing the manuscript.

James A. Hanley, Ph.D.
Department of Epidemiology and Health
McGill University
3775 University Street
Montreal, Quebec
Canada H3A 2B4

References

1. Lusted LB. Decision-making studies in patient management. N Engl J Med 1971; 284:416–424.
2. Goodenough DJ, Rossmann K, Lusted LB. Radiographic applications of receiver operating characteristic (ROC) curves. Radiology 1974; 110:89–95.
3. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978; 8:283–298.
4. Swets JA. ROC analysis applied to the evaluation of medical imaging techniques. Invest Radiol 1979; 14:109–121.
5. Swets JA, Pickett RM, Whitehead SF, et al. Assessment of diagnostic technologies. Science 1979; 205:753–759.
6. Green D, Swets J. Signal detection theory and psychophysics. New York: John Wiley and Sons, 1966: 45–49.
7. Green DM, Moses FL. On the equivalence of two recognition measures of short-term memory. Psychol Bull 1966; 66:228–234.
8. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating graph. J Math Psych 1975; 12:387–415.
9. 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.
10. Pearson ES, Hartley HO, eds. Biometrika tables for statisticians. Vol. I. 3d ed., London: Biometrika Trust, 1976:228.
11. Colton T. Statistics in medicine. Boston: Little, Brown and Company, 1974:168.
12. Hanley JA, McNeil BJ. Comparing the areas under two ROC curves derived from the same sample of patients. Radiology (forthcoming).
13. Metz CE, Kronman HB. Statistical significance tests for binormal ROC curves. J Math Psych 1980; 22:218–243.