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

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 non-diseased 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 magnitudes 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: 29-30, April 1982

METHODS

Indices Used to Summarize ROC Curves

A large number of theoretically based measures have 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
TABLE 2: Rating of 109 CT Images

<table>
<thead>
<tr>
<th>Disease Status</th>
<th>Deficiency (1)</th>
<th>Probability (2)</th>
<th>Questionable (3)</th>
<th>Probably Abnormal (4)</th>
<th>Definitely Abnormal (5)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>38</td>
<td>6</td>
<td>4</td>
<td>7</td>
<td>2</td>
<td>53</td>
</tr>
<tr>
<td>Abnormal</td>
<td>32</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>49</td>
</tr>
<tr>
<td>Totals</td>
<td>70</td>
<td>8</td>
<td>6</td>
<td>10</td>
<td>4</td>
<td>109</td>
</tr>
</tbody>
</table>

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.

RESULTS

I. A Three-Way Equivalence

To amplify the three-way equivalence between the 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 $\theta$, 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 $\theta$ 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 $x_{\text{ab}}$, will overlap; the $x_{\text{ab}}$ distribution will be centered to the right of the $x_{\text{na}}$ one. In a ranking experiment, the degree of suspicion, $x_{\text{sc}}$, will actually be reported on an ordered categorical scale. TABLE 1 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 $x_{\text{sc}}$ and $x_{\text{ab}}$ distributions overlap (i.e., some non-diseased patients had abnormal readings, 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-

3 An area can also be calculated by the trapezoidal rule; the area obtained in this way has been designated $\text{PL}(A)$. As is seen in Figure 1, $\text{PL}(A)$ 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 $A(\text{C})$ calculated as the smooth Gaussian estimate.
tive or, for want of a more precise term, the degree of suspicion, \( x \), which one obtains from the abnormal image would, in fact be greater than the corresponding \( x_0 \) obtained from the normal image. As a statistical shorthand, we refer to this probability as \( \theta = P(\text{ROC curve}) \). 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 = \( \theta = P(\text{ROC curve}) \)

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 \( x \) and \( x_0 \) distributions. Thus, the area under the curve can be thought of as 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. normal, exponential vs. gamma) need be specified. 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 trapazoidal 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 statistician is often interested in other aspects of ROC curves, 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 Fairwise 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 \( \theta = P(\text{ROC curve}) = 0.5 \). For \( x \) to be a good discriminator, this probability must 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 \( n_A \times n_N \) possible comparisons between the \( n_A \) sample \( x \)'s and the \( n_N \) sample \( y \)'s, scoring each comparison according to the rule

\[
S(\text{x,y}) = \left\{ \begin{array}{ll}
1 & \text{if } x > y \\
0 & \text{if } x = y \\
-1 & \text{if } x < y
\end{array} \right. \\
\text{(discrete data only)}
\]

and averaging the \( S \)'s over the \( n_A \times n_N \) comparisons, i.e.,

\[
W = \frac{\sum_{x \in A} \sum_{y \in N} S(x,y)}{n_A \times n_N}
\]

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 \( \theta = P(\text{ROC curve}) \). Since each comparison is scored as 1, 0, or 0, the average score \( W \) lies between 0 and 1 and reflects, as it should, what proportion of the \( x \)'s are greater than what proportion of the \( y \)'s. Obviously not all \( n_A \times n_N \) 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-

---

**Table II:** Computation of \( W \) and Its Standard Error

<table>
<thead>
<tr>
<th>Row</th>
<th>Contents</th>
<th>Column (Rating)</th>
<th>Total</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Number of normals rated ( x )</td>
<td>( x = 1 )</td>
<td>33</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>Number of abnormals rated ( x )</td>
<td>( x = 2 )</td>
<td>48</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>Number of normals rated ( &lt;x_0 )</td>
<td>( x = 5 )</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>Number of normals rated ( \geq x_0 )</td>
<td>( x = 5 )</td>
<td>0</td>
<td>33</td>
</tr>
<tr>
<td>5</td>
<td>( 0 / (1,0) + \frac{x_0 \times (1,0)}{1,0} )</td>
<td>( 1,0 )</td>
<td>1,635%</td>
<td>282</td>
</tr>
<tr>
<td>6</td>
<td>( 0 / (1,0) + \frac{x_0 \times (1,0)}{1,0} )</td>
<td>( 1,0 )</td>
<td>1,089</td>
<td>2,598</td>
</tr>
<tr>
<td>7</td>
<td>( 0 / (1,0) + \frac{x_0 \times (1,0)}{1,0} )</td>
<td>( 1,0 )</td>
<td>60,883</td>
<td>13,256</td>
</tr>
</tbody>
</table>

**Note:** \( W = \text{total}\{(x_A \times y_N) + (x_N \times y_A)\} \) = 0.893

\( W = \sqrt{\frac{1}{(1,0) + \frac{x_0 \times (1,0)}{1,0}} - \frac{x_0 \times (1,0)}{1,0}} \) = 0.09955 + 0.035764 + 1.92686 |

**Meaning and Use of the Area under an ROC Curve**

**DIAGNOSTIC RADIOLOGY**

---

### Notes:

1. The table values are calculated based on the given data and probability assumptions.
2. The calculations for the Wilcoxon statistic and standard error follow the standard procedure for yes/no comparisons.
3. The expressions used for calculating the area under the ROC curve and the Wilcoxon statistic are derived from integral calculus and statistical theory.
4. The table values are rounded to the nearest whole number for simplicity.
5. The standard error of the Wilcoxon statistic is calculated using the formula provided in the notes.
6. The computations are performed to understand the relationship between the area under the ROC curve and the probability of correct fairwise rankings.
7. The Wilcoxon statistic is a non-parametric test used to compare the median values of two independent groups.
8. The null hypothesis assumes no difference between the two groups, while the alternative hypothesis suggests a difference.
9. The calculations are performed to assess the discriminative ability of a diagnostic test.
10. The Wilcoxon statistic is calculated by summing the differences in ranks, with positive differences indicating the first group has a higher rank on average.
11. The standard error of the Wilcoxon statistic is used to determine the confidence interval for the test's significance.
exists 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 purpose, 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 θ > 0.5, W is no longer nonparametric; its standard error, SE(W), depends on two distribution-specific quantities, Q1 and Q2, and which have the following interpretation:

\[ Q1 = \text{Prob}(\text{two randomly chosen abnormal images will both be ranked with greater suspicion than a randomly chosen normal image}) \]

\[ Q2 = \text{Prob}(\text{one randomly chosen abnormal image will be ranked with greater suspicion than two randomly chosen normal images}) \]

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(W) as used throughout this paper, can be shown to be

\[ SE(W) = \sqrt{\frac{\theta(1-\theta)}{N} \left( \frac{1}{N_1} - \frac{1}{N_2} \right)} \]

\[ \text{or SE(W)} \]

(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 become a more and more useless measure the more categories there are. 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 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 Q1 and Q2. 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 Q1 and Q2 for use in

IV. W and SE(W) - Calculated With 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 as simply as rankings. The computations can be conveniently carried out directly from Table I; Figures 3 and 4 are derived from 3 and 1 by successive deletion and cumulation respectively. The quantity W can be computed in no more than by using the entries in rows 1, 2, and 5; SW requires calculation of the two intermediate probabilities Q1 and Q2 (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 = 0.893 = 89.3% derived in this way agrees exactly with the area under the ROC curve calculated by the trapezoidal rule. By way of comparison, the area under the smooth Gaussian-based ROC curve fitted by the maximum likelihood technique of Dorfman and Allen (9) is 0.912 or 91.2% of the area under the smooth ROC curve derived from the parameters of a straight-line fit to the ROC plotted in double probability paper (see Swets (4), pp. 114-115) is 0.905 or 90.5%. The slightly lower estimate predicted by the Gaussian (or equivalent by the trapezoidal rule, merely reflects the fact that the rating scale does not have infinitely fine "grain." In other 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

\[ \text{the SE of 2.96% are not entirely justified.} \]

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 an index of discriminability of 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 (τ), and looking up (τ) in Gaussian probability tables, one obtains a smoothed area of 90.5%, which is only 0.6% (in absolute terms) or 0.6% (in relative terms) different from the 91.3% 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 even further on predicting the SE produced by this method.

All of the discussion thus far has centered on how 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 most 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 preexperiment sample planning. At this stage, one is often asked: "We wish to use θ, the area under the ROC curve, to determine the performance of an imaging system, and we would like to have this index accurate to within ±\epsilon, i.e., of the fluctuations in the index caused by the random sampling of cases. How many cases must we study to ensure a reasonable level of precision?" This is equivalent to asking how large nA and nB must be so that the resulting SE of θ is of a reasonably small magnitude, and the resulting confidence interval is correspondingly narrow.

In addition to the quantities nA and nB,
Formula 1 contains three other parameters—\( \theta \), \( Q_1 \), and \( Q_2 \). While one can use anticipated values of the true area \( \hat{\theta} \), the quantities \( Q_1 \) and \( Q_2 \) are complex functions of the underlying distributions for \( x_0 \) and \( x_0' \). Fortunately, for any specified pair of distributions \( F \) is almost entirely determined by \( \theta \), and only very slightly influenced by any further parameters of the distributions. As an example, Figure 2 shows how, little the relationship between \( SE(\theta) \) and \( \theta \) 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 \( SE(\theta) \) 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 \( \theta \), i.e.,

\[
Q_1 = \theta + (2 - \theta) \\
Q_2 = 2\theta + (1 + \theta)
\]

When these expressions are substituted into Formula 1, we obtain the SE to be expected at any anticipated level of performance \( \theta \), and can vary the \( n \) until \( SE(\theta) \) is sufficiently small. For example, consider an experiment where the diagnostic accuracy is expected to be in the neighborhood of \( \theta = 85\% \). Then \( Q_1 = 0.85 + 1.5 = 0.7391 \), and \( Q_2 = 2(0.85) + 1.85 = 0.7691 \). Then with \( n_0 = n_0' = 40 \), Formula 2 predict an SE of 4.35% while \( n_0 = n_0' = 60 \) will reduce the SE to 3.56%. Figure 3 gives \( SE(\theta) \) for various sample sizes and various anticipated \( \theta \). A number of points should be noted:

1. As one expects with SEs, they vary inversely with \( \sqrt{n} \), so that, for example, one must quadruple the sample size to halve the SE.
2. The SEs are smallest for very high \( \theta \), i.e., those close to 1.
3. The SEs are slightly more conservative than those obtained under the Gaussian model (Fig. 2).

4. One must resort to Formulas 1 and 2 to calculate \( n \) when the number of normal cases \( n_0 \) does not equal the number of abnormal cases \( n_0' \). In using these formulas, one is interested in comparing the SEs of approximately \( 3\% \) obtained from our previously mentioned rating experiment example with what would have been obtained from an actual 24FC experiment. In the latter, one would estimate \( \theta \) simply by calculating

\[
antilog(\text{sum of ratings})/	ext{(number of raters)}
\]

the fraction \( \theta \) of pairs of images where the normal and abnormal images were correctly identified, and one would accompany this estimate of \( \theta \) with a standard error, based on the bi-normal distribution, of \( \sqrt{(1-\theta)/\theta} \). To achieve a standard error of 0.03 or 3\%, and assuming that in fact \( \theta \) in the 2AF experiment turned out to be 0.9 or 90\%, one would need \( n = 100 \) pairs of (normal, abnormal) images, a considerably greater number of images (200 or 2000) than the \( n_0 + n_0' = 109 \) used in the rating experiment. For purposes of precision (low SE) in measuring \( \theta \), we can think of the \( k \) images (in 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 2AF 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 derived from a rating experi-
ment to construct a confidence interval on \( \theta \). 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 \( \theta \) under the curve, the distribution of the \( \theta \)'s we obtained is not entirely symmetric, but is instead somewhat skewed towards \( \theta < 0.5 \). This skewness is more marked as the "true" \( \theta \) approaches 1, and as the expected number of "misclassified pairs" [\( n(1 - \theta) \)] 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 \( \theta \), rather than using the approxi-
mation (symmetric) one of \( 1.645 \text{SE}(\theta) \), \( 1.96 \text{SE}(\theta) \), . . . , provided by the normal distribution. In the example here \( n(1 - \theta) = 10 \) is considerably greater than the rule of thumb of 5; thus, the symmetric 95% confidence interval of \( 90.5\% \pm 1.96(0.07) \) or \( 84.5\%, 96.5\% \) will be reasonably correct, compared with the exact, slightly asymmetric interval of \( 93.2\%, 95.1\% \) obtained by consulting chisquared confidence limits for binomial sampling (10) with \( \theta = 0.905 \) 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 \( \theta \), 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 "statis-
tical sensitivity" depends on how small the probabilities \( \alpha \) and \( \beta \) of committing a type I or type II error are. Typically, one seeks a power \( 1 - \beta = 0.85 \) of 85% or 90% so that if a specified dif-
ference 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 a 0.05 (5%) level to be the criterion for a significant difference.

TABLE II: Number of Normal and Abnormal Subjects Required to Provide a Probability of 80%, 90%, or 95% of Detecting Various Differences between the Areas \( \theta_1 \) and \( \theta_2 \) under Two ROC Curves (Using a One-Sided Test of Significance with \( p = 0.05 \))

<table>
<thead>
<tr>
<th>( \theta_1 )</th>
<th>0.700</th>
<th>0.750</th>
<th>0.800</th>
<th>0.825</th>
<th>0.850</th>
<th>0.900</th>
<th>0.925</th>
<th>0.950</th>
<th>0.975</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.700</td>
<td>452</td>
<td>286</td>
<td>158</td>
<td>100</td>
<td>68</td>
<td>49</td>
<td>37</td>
<td>28</td>
<td>22</td>
</tr>
<tr>
<td>0.750</td>
<td>897</td>
<td>392</td>
<td>216</td>
<td>135</td>
<td>92</td>
<td>66</td>
<td>49</td>
<td>36</td>
<td>29</td>
</tr>
<tr>
<td>0.800</td>
<td>3135</td>
<td>493</td>
<td>271</td>
<td>169</td>
<td>115</td>
<td>82</td>
<td>64</td>
<td>46</td>
<td>34</td>
</tr>
<tr>
<td>0.825</td>
<td>610</td>
<td>267</td>
<td>149</td>
<td>90</td>
<td>63</td>
<td>45</td>
<td>34</td>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td>0.850</td>
<td>839</td>
<td>366</td>
<td>218</td>
<td>136</td>
<td>89</td>
<td>64</td>
<td>45</td>
<td>34</td>
<td>27</td>
</tr>
<tr>
<td>0.900</td>
<td>1077</td>
<td>499</td>
<td>252</td>
<td>156</td>
<td>107</td>
<td>75</td>
<td>55</td>
<td>42</td>
<td>33</td>
</tr>
<tr>
<td>0.925</td>
<td>565</td>
<td>246</td>
<td>136</td>
<td>85</td>
<td>58</td>
<td>41</td>
<td>31</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>0.950</td>
<td>776</td>
<td>337</td>
<td>185</td>
<td>113</td>
<td>77</td>
<td>55</td>
<td>42</td>
<td>33</td>
<td></td>
</tr>
</tbody>
</table>

*0.80 probability = top number; 90% probability = middle number; 95% probability = bottom number.*
The calculations are based on an adaptation of the sample size formula given by Colton (11), namely:

\[ n = \frac{Z_1^2 + Z_2^2 + \sqrt{Z_1^4 + Z_2^4}}{\delta^2} \]

where

\[ Z_1 = 1.645, \text{ for a } 95\% \text{ one-sided test of significance} \]

\[ Z_2 = 0.84, 1.28, \text{ or } 1.649 \text{ for } 80\%, 90\%, \text{ or } 95\% \text{ power} \]

\[ \delta = \phi_1 - \phi_2 \]

\[ V_1 = Q_1 + Q_2 - 2\phi_1 Q_1 \text{ and } Q_2 \text{ obtained using } \phi_1 \text{ in Equation 2} \]

\[ V_2 = Q_1 + Q_2 - 2\phi_2 Q_1 \text{ and } Q_2 \text{ obtained using } \phi_2 \text{ in Equation 2} \]

As an example, it shows that if indeed the "true" area (i.e., the thing 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. On the other hand, if the "no difference" persists in spite of a precise sample sizes such as those shown in Table III, one can reasonably conclude that the stated differences are real. Table III and Table IV show 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 the imaging 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 firm 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 non-ROC indices are relevant to separate discriminatory power. We use the ROC index for convenience only.)

Second, the combination of a graphical method for obtaining a smoothed area and a computational formula for its standard error is shown to be a very effective tool, the investigator can obtain almost the full benefit of a parametric maximum likelihood analysis. However, it can be a large computer and a skilled computer operator. As a consequence, it is usually necessary to interpret the results of this study. It should be noted that this sample size for an ROC curve can be obtained from a table. The data may be more helpful to the statistician, e.g., with and without history, with and without varying levels of contrast. Methods to deal with this latter situation are the subject of an article in Table III may be compared with an experimental condition, e.g., with and without history, with and without varying levels of contrast. Methods to deal with this latter situation are the subject of an article in Table III may be compared with an experimental condition, e.g., with and without history, with and without varying levels of contrast.

In fact, knowing the classifier used, one can solve Equation 3 for Z to obtain how high the probability of a type II error really was.
test statistically whether two curves are different (they could still subside 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 $x_2$ and $x_3$ 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 Dornbusc and Alif (9).

Acknowledgments: We wish to thank Colin Hogg, Charles Most, and Stanley Sharp for helpful suggestions and Irene McCann for preparing the manuscript.

James A. Hanley, Ph.D.
Department of Epidemiology and Health
McGill University
3775 University Street
Montreal, Quebec
Canada H3A 2M4

References