Statistics in Radiology

Harold L. Kundel, MD, *Editor*

Sampling Variability of Nonparametric Estimates of the Areas under Receiver Operating Characteristic Curves: An Update

James A. Hanley, PhD[1,2], Karim O. Hajian-Tilaki, PhD[1]

Rationale and Objectives. Several methods have been proposed for calculating the variances and covariances of nonparametric estimates of the area under receiver operating characteristic curves (AUC). The authors provide an explanation of the relationships between them and illustrate the factors that determine sampling variability.

Methods. The authors investigated the algebraic links between two methods, that of "placements" and that of "pseudovalues" based on jackknifing. They also performed a numerical investigation of the comparative performance of the two methods.

Results. The "placement" method has a simple structure that illustrates the determinants of the sampling variability and does not require specialized software. The authors show that the pseudovalues used in the jackknife method are directly linked to the placement values.

Conclusion. Because of the close link, borne out in a numeric investigation of the sampling variation, and because of the ease of computation, the choice between the two methods can be based on users' preferences. For indexes other than the AUC, however, the use of pseudovalues holds greater promise.

Key Words. Nonparametric ROC analysis; area under the curve, DeLong method; jackknife pseudovalues

From the [1]Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada; and [2]Division of Clinical Epidemiology, Royal Victoria Hospital, Montreal, Canada.

Supported by funds from the Natural Sciences and Engineering Council of Canada and the Fonds de la recherche en santé du Québec.

Address reprint requests to J. A. Hanley, PhD, Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Ave W, Montreal, Quebec, Canada H3A 1A2.

Received March 25, 1996, and accepted for publication after revision September 18, 1996.

Acad Radiol 1997;4:49–58

The area under the receiver operating characteristic (ROC) curve (AUC) is commonly used as a measure of the accuracy of a diagnostic test. It can be estimated parametrically or nonparametrically [1–6]. Although this statistic has a helpful interpretation, the assessment of its sampling variability—especially in the nonparametric case—is less intuitive. At least four formulas or approaches have been proposed for calculating the variance of a nonparametric AUC estimate, two of which are extendable to the covariance between estimates from two curves.

The first of these four approaches was initially suggested by Bamber [7], who noted the connection between the AUC and the parameter estimated with the Wilcoxon statistic. Hanley and McNeil [6] used this link to give a

practical method for calculating the variance of the AUC. They relied on the traditional form of the nonnull variance for the Wilcoxon statistic, which involves three components, and showed how to estimate each component from the raw ROC data. A second, much simpler, variance formula that involves a closed form function of just the AUC estimate itself was also offered by Hanley and McNeil [6]. However, because it was derived from the assumption of two underlying negative exponential distributions, with just one free parameter, it can underestimate the variance when the AUC is close to 0.50 and overestimate it when the AUC is close to 1. Hanley and McNeil [8] also offered a method for calculating the covariance of two AUC estimates derived from the same sample of cases. The third method was given by DeLong et al [9]. Although developed primarily as a way to estimate the *co*variance between two or more AUC estimates, it can also be used to calculate the variance of a single estimate. It is similar to the method described by Wieand et al [10]. Although it provides the cleanest and most elegant approach to variances and covariances of AUCs, it has unfortunately not become widely used in radiology. This may have to do with the type of journal in which it was published or with the perceived complexity of the computations. The fourth method is that of jackknifing [11], first described by McNeil and Hanley [12] in the case of a different accuracy index but also applicable to the AUC index or two correlated AUC estimates. Most recently, it has been extended by Dorfman et al [13] to ROC data involving multiple readers.

From the description of each of the four methods, it is not easy to understand the relationships between them or to develop an intuition for the factors that determine the sampling variability. We believe that the method described by DeLong et al [9] should be the most intuitive and provide the greatest insights into the determinants of the sampling variability of the AUC. Conversely, the method of jackknifing is both simple and mysterious. The purpose of this article is to popularize both methods. We begin by explaining the method described by DeLong et al in enough detail for users to see its simplicity and its insights. Second, we try to demystify the concepts of jackknifing and pseudovalues. We then go on to show the link, only vaguely hinted at by DeLong et al, between the quantities used in their method and the pseudovalues used in the jackknifing method. In both approaches, the elemental quantities can be thought of as case-specific measures of "case difficulty." Finally, for those who are more convinced by data than by algebra, we provide a numeric investigation of the comparative performance of the two approaches.

TABLE 1: Disease Status and Ratings Based on Images Produced with Two Different Field Strengths

Patient No.	Disease Present	Rating*	
		Field Strength 1	Field Strength 2
1	Yes	1	1
2	No	2	1
3	Yes	5	5
4	No	1	1
5	No	1	1
6	Yes	1	1
7	Yes	2	4
8	No	1	1
9	No	2	2
10	Yes	2	2
11	No	1	1
12	No	1	1
13	Yes	5	5
14	No	1	1
15	No	1	1

*Rating scale ranged from 1(definitely negative) to 5 (definitely positive).

DATA USED FOR ILLUSTRATION

The data set we use for illustration comes from a small side study conducted in conjunction with a larger one to assess readers' diagnostic accuracy with images generated from two magnetic resonance (MR) imaging units with two different field strengths. A total of 15 patients suspected of temporal lobe epilepsy (six with disease and nine without) were assembled. After giving informed consent, each patient was imaged with two different field strengths. Both examinations were scheduled for the same day or on two consecutive days; the order was randomly determined for each patient. The interpretation involved several detection tasks. For each task, the reader used a five-point rating scale ("definitely normal," "probably normal," "uncertain," "probably abnormal," and "definitely abnormal") to describe his confidence about whether an abnormality was present. The true situation was established independently. The data concerning one reader's performance on one such task for one hemisphere are shown in Table 1. Although the numbers of diseased and non-

TABLE 2: DeLong Method: Calculation of Placements (and of the AUC and Its Variance) from Rating Data for Six Diseased and Nine Nondiseased Subjects

Ratings for $n = 9$ Nondiseased Subjects*	Ratings for $m = 6$ Diseased Subjects*						Placement
	$Y_1 = 1$	$Y_2 = 5$	$Y_3 = 1$	$Y_4 = 2$	$Y_5 = 2$	$Y_6 = 5$	V_X
$X_1 = 2$	0.0	1	0.0	0.5	0.5	1	0.50
$X_2 = 1$	0.5	1	0.5	1	1	1	0.83
$X_3 = 1$	0.5	1	0.5	1	1	1	0.83
$X_4 = 1$	0.5	1	0.5	1	1	1	0.83
$X_5 = 2$	0.0	1	0.0	0.5	0.5	1	0.50
$X_6 = 1$	0.5	1	0.5	1	1	1	0.83
$X_7 = 1$	0.5	1	0.5	1	1	1	0.83
$X_8 = 1$	0.5	1	0.5	1	1	1	0.83
$X_9 = 1$	0.5	1	0.5	1	1	1	0.83
Placement V_Y	0.39	1	0.39	0.89	0.89	1	. . .

Note.—Data indicate the placement of each Y with respect to each X, with 1 indicating the "correct" ordering, 0 an "incorrect" ordering, and 0.5 if Y and X are equal. The data in the right column and bottom row of the Table, obtained as the averages of the corresponding rows/columns, are the placements or pseudoaccuracies corresponding to each X and each Y. Calculations in this and later tables were performed with spreadsheet precision, but numbers were rounded for presentation. Data were obtained with the first of the two field strengths in Table 1. AUC = average of V_X's = average of V_Y's = 0.76. $\text{Var}(V_X)$ = 0.0216; $\text{Var}(V_Y)$ = 0.0848. Var(AUC) = 0.0216/9 + 0.0848/6 = 0.0165. SE(AUC) = $\sqrt{0.0165}$ = 0.13.

*Rating scale ranged from 1 (definitely negative) to 5 (definitely positive).

diseased subjects are small, and although the distribution of ratings is somewhat "lumpy," the advantage of this small data set is that we can show *all* calculations.

THE DELONG APPROACH

We begin with a single diagnostic test (eg, images obtained at one particular field strength). Let $X_1, X_2, \ldots, X_n$ denote the rating or test results for a sample of n nondiseased subjects and $Y_1, Y_2, \ldots, Y_m$ denote the test results for m diseased subjects. Suppose, as in Table 1, that the scale is such that larger test values constitute greater evidence of disease.

The approach is based on transforming each rating into a "placement" value. For a rating Y for a subject in the diseased group, its "placement," which DeLong et al called V_Y, is the fraction or percentage of X's that it exceeds (ie, one converts the Y into the percentile it would occupy in the X sample). Thus, if the test is informative the V_Y's will tend to be at the higher end of the 0-1 (or 0-100) scale.

Conversely, for the placement V_X of an X value for a subject in the nondiseased group, one calculates where—in the reverse percentile scale—it would lie in the Y distribution. Thus, if the X's tend to be less "positive" than the Y's, the V_X's will again tend to be at the upper end of the (0,1) scale.

One way to visualize these concepts and to see their connection with previous methods is to form an $n \times m$ matrix consisting of the placements of each Y with respect to each X. As is displayed in Table 2, matrix entries are scored as 1 if Y is greater than X, 0.5 if the two are equal, and 0 if Y is less than X. Then, the placement of a particular Y can be calculated by averaging the entries corresponding to that Y. For example, for Y_4, one averages the nine entries in column 4 to transform the original rating Y for a patient with disease into a placement $V_Y = 0.89$. Similarly, the placement (or V_X) for any X for a patient without disease is obtained by averaging the row of entries for that X. In fact, if calculations are done on a spreadsheet, one can calculate the placements directly without having to form the $n \times m$ matrix. To do so in Excel (Microsoft, Redmond, WA), for example, one can name the two ranges containing the X and Y values. If one calls them XRANGE and YRANGE, then the V's corresponding to the individual X values can be calculated with the following function: AVERAGE (IF $X <$ YRANGE, 1, IF $X >$ YRANGE, 0, 0.5), provided one indicates that the argument of the AVERAGE function is an array. The V's cor-

responding to the individual Y values are calculated in the same way.

The set of V_X's and V_Y's can be used in place of the original X and Y ratings to construct the empirical ROC curve. The average $\overline{V}_X = 0.76$ of the $n = 9$ V_X's and the average $\overline{V}_Y = 0.76$ of the $m = 6$ V_Y's are both equivalent to the nonparametric AUC. Thus, each V_X and each V_Y is an estimate (albeit a noisy one) of the AUC. Whereas DeLong et al call the V's the components of the "U" statistic, we prefer to call them "placements" or "patient-specific accuracies" because they are in the same (0,1) scale as the AUC itself and because one can think of a $V = 1$ as one of the easiest cases and a $V = 0$ as one of the most difficult.

Up to now, the reader may ask why one would bother to calculate these six individual V_Y's and nine V_X's, since one can simply calculate the AUC directly from the average of the $6 \times 9 = 54$ comparisons of each Y with each X. The answer is that the variations of these six and nine V's can be used directly to estimate the variance of the AUC estimate.

Variance of the AUC Estimate

In the method used by DeLong et al [9], the variance of the AUC estimate is calculated as the sum of two contributions, one relating to the number and variability of the V_X's, the other to the number and variability of the V_Y's, as follows:

$$\text{Var[AUC]} = \frac{\text{Variance of } V_x\text{'s}}{n\text{: number of } V_x\text{'s}} + \frac{\text{Variance of } V_y\text{'s}}{m\text{: number of } V_x\text{'s}}. \quad (1)$$

Those interested in the equivalence of this equation and the formula given in Hanley and McNeil's first article [6] can consult the textbook by Hettmansperger [14]. DeLong et al [9] omitted the third variance component, $\text{AUC}(1 - \text{AUC})/mn$, since it is negligible when n and m are large.

The attraction of Equation (1) is that each of the two contributions has the following form: variance of observations divided by number of observations, which is the well-known formula for the variance of a mean. We take the square root of this variance to obtain the standard error of a mean (SEM). However, there is one important difference in this particular instance. $\overline{V}_X$ and $\overline{V}_Y$ are both equivalent to the AUC statistic; however, when calculating the variance of AUC one cannot rely on the V_Y's alone and treat the V_X's as a set of constants or vice versa. The variability in the V_Y's and V_X's must both be used. One can see the degree to which the variability of the AUC estimate is influenced by both sets of V's if one considers an extreme example: If $m =$ 50 or even 500 but $n = 1$, then the estimate of the AUC depends entirely on the single $n = 1$ observation from the nondiseased population.

The bottom rows of Table 2 show the calculation of the variance of an AUC from the two component accuracies. The estimated variances of the pseudoaccuracies corresponding to the X's and Y's are 0.0216 and 0.0848. Thus, the estimated variance of the estimated accuracy index, that is, of AUC = 0.76, is as follows:

$$\text{Var[AUC]} = \frac{0.0216}{9} + \frac{0.0848}{6} = 0.0165,$$

so that the standard error (SE) is

$$\text{SE[AUC]} = \sqrt{0.0165} = 0.13$$

The structure of Equation (1) reveals one additional insight into the sampling variability (and its control) that does not appear to have been commented on previously. This insight comes from the nature of the component variances (0.0216 and 0.0848 in our example). These are estimates of the variance of the true-positive fraction (TPF) and false-positive fraction (FPF) points on the smooth ROC curve underlying the data. One can imagine the smooth ROC curve as a very large number (say 1,000 or 10,000) of TPF points corresponding to 100 or 10,000 equally spaced FPF points. If the ROC curve were the 45° diagonal line, these TPF points would be uniform on the (0,1) scale, and their variance would be 1/12 or 0.0833. The closer the curve is to the top left corner, the more concentrated and closer to the 1 than the 0 end of the (0,1) scale the TPF points will be and the smaller will be their variance. The V_Y's in the method used by DeLong et al are considered to be a random sample of these TPF points.

Likewise, one can think of the ROC curve as a large sequence of FPF points, measured at equal TPF spacings, and one can think of the variance of the V_X's as an estimate of the variance of this sequence. In practice, however, because actual sample sizes are finite and the data are recorded on a discrete rather than a continuous scale, the observed variance of the V_X's and V_Y's can be larger than it would be if one could observe

TABLE 3: DeLong Method: Calculation of Variances and Covariances of Two AUCs from the Variances and Covariances of the Individual Placement Values

Patient Group and No.	Placements	
	Field Strength 1	Field Strength 2
Patients with disease		
1	0.39	0.44
3	1.00	1.00
6	0.39	0.44
7	0.89	1.00
10	0.89	0.94
13	1.00	1.00
Average	0.76	0.81
Variance	0.0848	0.0787
Covariance*	0.0809	
Patients without disease		
2	0.50	0.83
4	0.83	0.83
5	0.83	0.83
8	0.83	0.83
9	0.50	0.58
11	0.83	0.83
12	0.83	0.83
14	0.83	0.83
15	0.83	0.83
Average	0.76	0.81
Variance	0.0216	0.0069
Covariance*	0.0081	

Note.—$Var(AUC_1) = 0.0848/6 + 0.0216/9 = 0.0165$; $Var(AUC_2) = 0.0787/6 + 0.0069/9 = 0.0139$; $Covar(AUC_1, AUC_2) = 0.0809/6 + 0.0081/9 = 0.0144$.

*Covariances for individual data pairs were calculated as follows: correlation × SD_1 × SD_2, where SD = standard deviation. Some spreadsheets can calculate the covariance directly, but they use a divisor of n, rather than $n - 1$. If this direct covariance function is used, one should multiply the result by $n/(n - 1)$.

them on a more refined (ie, truly continuous) scale. However, the main point remains: One can project the variance of the AUC from just the ROC curve itself and the sample sizes, m and n, used to estimate it. To do this, simply measure sufficient TPFs (corresponding to equally spaced FPFs) and sufficient FPFs (corresponding to equally spaced TPFs) to get a reasonably stable estimate of their variances; then divide these variances by m and n, respectively, and add the results to arrive at an estimate of the variance of the AUC. (Note that the variance for the FPFs is the same as that of their complements, namely, specificities.)

With this representation of the variance of the AUC, we can see that if the curve is symmetric, the two component variances will be equal. Thus, if one has a choice of how to allocate m and n, the variance of the AUC is minimized by taking $m = n$. If, however, the curve is asymmetric and rises more rapidly from the origin and then flattens out more as it turns toward the upper right corner, then the variance of the "sensitivities at equal specificities" will be smaller than the variance of the "specificities at equal specificities"; in this case, the optimal allocation would be to have m and n in the ratio of the two variances, so one uses the larger denominator to counteract the larger variance.

Variance and Covariances for Comparison of Two AUCs

Very often, the main purpose of a study is to compare the area under one ROC curve ($AUC_1 = 0.76$ in our example) with that from a second curve derived from the same sample of subjects ($AUC_2 = 0.81$). The SE used to judge this difference in AUCs involves not just the variances $Var(AUC_1) = 0.0165$ and $Var(AUC_2) = 0.0139$ (shown in Table 3, calculated in the same way as for AUC_1), but also the covariance (Covar) of these two estimates. The SE of the 0.81 - 0.76 is calculated as the square root of

$$\text{Var}[\text{AUC}_1 - \text{AUC}_2] = \text{Var}[\text{AUC}_1] + \text{Var}[\text{AUC}_2] - 2\text{Covar}[\text{AUC}_1, \text{AUC}_2]. \quad (2)$$

The main purpose of the method used by DeLong et al is the calculation of the covariance term. Following on the pattern already established for a variance, the covariance can again be expressed as the sum of two contributions, one involving the number and covariability of the pairs of V_x's, the other involving the number and covariability of the pairs of V_y's, as follows:

$$\text{Covar}[\text{AUC}_1, \text{AUC}_2] = \frac{\text{Covariance of pairs of } V_x\text{'s}}{n\text{: number of pairs}} + \frac{\text{Covariance of pairs of } V_y\text{'s}}{m\text{: number of pairs}}. \quad (3)$$

The calculation of the covariance of AUC_1 and AUC_2 is shown in Table 3. This covariance is then inserted in Equation (2) to obtain the SE of $\sqrt{(0.0165 + 0.0139 - 2 \cdot 0.0144)} = 0.04$ associated with the AUC difference of 0.05. Because the sample size was very small, the sampling variability of each AUC estimate is relatively

large. Thus, despite the positive correlation between the two, the confidence interval for their difference ($0.05 \pm 1.96 \times 0.04$ or -0.03 to 0.13) is large.

THE JACKKNIFE APPROACH

At the heart of the jackknife technique is the concept of a pseudovalue. One can think of a pseudovalue as a replacement of a raw data value by an equivalent value. The collection of pseudovalues gives the same summary statistic as the original values, and the variation of these pseudovalues allows one to calculate the sampling variability of the summary statistic. The pseudovalues are particularly valuable when the summary statistic has a complex form and the form of the sampling variation is not readily apparent. For example, it is not obvious what the form of the variance of the AUC estimate should be, other than that it must somehow involve m and n—or some function of them—in the denominator.

The pseudovalue corresponding to any one observation can be defined as the contribution of that observation to the summary statistic. This can be determined by calculating the summary statistic with and without the observation in question. For example, if the summary ROC index is the AUC, then the AUC pseudovalue (pAUC) corresponding to observation i is

$$\text{pAUC}_i = (m+n)\text{AUC} - (m+n-1)\text{AUC}_{(-i)}, \qquad (4)$$

where AUC is the area calculated with all $m + n$ observations and $\text{AUC}_{(-i)}$ the area calculated from the $(m + n - 1)$ observations, with observation i omitted. (A major advantage of jackknife pseudovalues over the pseudoaccuracy measures introduced in the method used by DeLong et al is that they can be calculated for any summary statistic, whether estimated parametrically or nonparametrically.)

In Table 4, we illustrate the calculation of the 15 AUC pseudovalues corresponding to the 15 original observations for the first diagnostic test. One begins with the AUC estimate of 0.76. Thus, if we work with three decimal places, the contribution of the first observation to this AUC of 0.759 can be calculated as $15 \cdot 0.759 - 14 \cdot 0.833 = -0.277$, or -0.28 when rounded back to two decimal places. The pseudovalues for observations 2–15 can be calculated in a similar way with Equation (4), yielding the last column of Table 4. One can now think of these 15 pseudovalues as the statistical equivalent of the 15 original (truth, rating) observations because (*a*) their mean (0.76) is the same as the AUC of the 15 original observations and (*b*) as we will see later, if one calculates the SE of this 0.76, by using the familiar formula used to calculate the SE of a mean of 15 independent observations, it gives virtually the same answer as if one calculates the SE with the more conventional formulas such as those of Bamber [7], Hanley and McNeil [8], or DeLong et al [9].

TABLE 4: Jackknife Pseudovalues and AUC

Patient No.	With Disease?	Rating	$\text{AUC}_{(-i)}$*	Jackknife Pseudovalues
1	Yes	1	0.833	−0.28
2	No	2	0.792	0.31
3	Yes	5	0.711	1.43
4	No	1	0.750	0.89
5	No	1	0.750	0.89
6	Yes	1	0.833	−0.28
7	Yes	2	0.733	1.12
8	No	1	0.750	0.89
9	No	2	0.792	0.31
10	Yes	2	0.733	1.12
11	No	1	0.750	0.89
12	No	1	0.750	0.89
13	Yes	5	0.711	1.43
14	No	1	0.750	0.89
15	No	1	0.750	0.89

Note.—Data were obtained with the first of two field strengths. Average = AUC = 0.76; variance = 0.2752; variance of AUC = 0.2752/15 = 0.0183; SE of AUC = $\sqrt{0.0183}$ = 0.14.

*AUC with i omitted.

One can see that the pseudovalues rank cases in terms of diagnostic difficulty or in terms of their contribution to the level of the AUC. For example, patient 3 has a pseudovalue of 1.43, above the average of 0.76. Because the image from this case—in the diseased group—was rated 5 (definitely abnormal), it makes sense that it be considered an "easier-than-average" case. In contrast, patient 1, with a below-average pseudovalue of −0.28, was the most difficult case (diseased, but rated as 1 [definitely normal]). Patient 2 was a nondiseased subject whose image was rated 2 (probably normal); thus, the far below average pseudovalue (0.31) seems rather harsh, until one considers that most of the images from nondiseased subjects were rated 1 (definitely normal). Unfortunately, the small number of cases and the presence of many ties in this illustrative example make the pseudovalue scale rather coarse.

TABLE 5: Comparison of Two Diagnostic Tests: Calculation of Variances and Covariances of Two AUCs from the Variances and Covariances of the Jackknife AUC Pseudovalues

Patient No.	Jackknife Pseudovalues	
	Field Strength 1	Field Strength 2
1	−0.28	−0.21
2	0.30	0.85
3	1.43	1.35
4	0.89	0.85
5	0.89	0.85
6	−0.28	−0.21
7	1.12	1.35
8	0.89	0.85
9	0.30	0.42
10	1.12	1.19
11	0.89	0.85
12	0.89	0.85
13	1.43	1.35
14	0.89	0.85
15	0.89	0.85
Average	0.76	0.81
Variance	0.2752	0.2325
Covariance	0.2406*	
Var(AUC)	0.2752/15 = 0.0183	0.2325/15 = 0.0155
Covar(AUC_1, AUC_2)	0.2406/15 = 0.0160	

*Covariance was calculated as follows: correlation × SD_1 × SD_2, where SD = standard deviation, or covariance = (15/14) × covariance calculated by using direct spreadsheet function.

Variance of the AUC Estimate

The variance of the AUC follows immediately from the representation of the AUC as the mean of $m + n$ "independent" pseudovalues, namely,

$$\text{Var[AUC]} = \text{Variance of mean of all } m+n \text{ pAUCs} = \frac{\text{Variance of all pAUCs}}{m+n\text{: number of pAUCs}}. \quad (5)$$

Thus, as is shown in Table 4, the variance associated with the AUC of 0.76 is 0.2752/15 = 0.0183; to two decimal places, this yields an SE of 0.14 (the method of DeLong et al yields an SE of 0.13).

Comparison of Two Correlated AUCs

Similarly, if one has two AUC estimates calculated from the same subjects, one can compute the pseudovalues for each patient with each test. One can then use the same techniques to calculate the covariance between the two AUCs; one simply treats the samples of pseudovalues as $m + n$ correlated pairs of observations and uses the equation for the covariance of two averages, namely,

$$\text{Covar}[AUC_1, AUC_2] = \frac{\text{Covar}(pAUC_1, pAUC_2)}{m+n}. \quad (6)$$

That is, one simply replaces the "squares" implicit in Equation (5) with products. This is illustrated in Table 5; the covariances are simply the average of the products of the deviations of each pair of pseudovalues from their own test-specific averages. Again, one can see that the covariance of AUC_1 and AUC_2, calculated with this method, is very close to that calculated with the method of DeLong et al. As before, the SE of the difference between the two AUCs can be calculated as the square root of the variance of their difference, calculated with Equation (2).

LINK BETWEEN PLACEMENT VALUES AND AUC PSEUDOVALUES

One of the difficulties for newcomers to pseudovalues is the seemingly unnatural scale on which they are distributed. The center (mean) of the AUC pseudovalues has a meaning—it is the AUC itself—but the range of variation is less intuitive. This is in contrast to the natural (0, 1) scale on which the placements are measured. That there is a direct one-to-one relationship between pseudovalues and placements can be seen from the following algebra.

Let $X_1, X_2, \ldots, X_n$ denote test results for n nondiseased subjects and $Y_1, Y_2, \ldots, Y_m$ denote results for m diseased subjects. If a larger value indicates a higher probability of a "signal," then for each (X_i, Y_j) pair an indicator function $I(X_i, Y_j)$, or I_{ij} for short, is defined as follows:

$$I(X_i, Y_j) = \begin{cases} 1 \text{ if } Y_j > X_i \\ 1/2 \text{ if } Y_j = X_i \\ 0 \text{ if } Y_j < X_i. \end{cases}$$

The average

$$\frac{1}{mn}\sum_i\sum_j I_{ij}$$

of these I's over all nm comparisons is the nonparametric AUC.

The jackknife AUC pseudovalue (pAUC) for X_i is defined as $\text{pAUC}_i = (m + n)\text{AUC} - (m + n - 1)\text{AUC}_{-i}$. Substituting for the two AUC statistics gives

$$\text{pAUC}_i = (m+n)\left[\frac{1}{mn}\sum_{k}^{n}\sum_{j}^{m} I_{kj}\right] - (m+n-1)\left[\frac{1}{m(n-1)}\sum_{k\neq i}^{n}\sum_{j}^{m} I_{kj}\right].$$

Then, after rearranging the second term we get

$$\text{pAUC}_i = \frac{m+n}{mn}\sum_{k}^{n}\sum_{j}^{m} I_{kj} - \frac{m+n-1}{m(n-1)}\left[\sum_{k}^{n}\sum_{j}^{m} I_{kj} - \sum_{j}^{m} I_{ij}\right]$$

$$\text{pAUC}_i = \left(\frac{m+n}{mn} - \frac{m+n-1}{m(n-1)}\right)\sum_{k}^{n}\sum_{j}^{m} I_{kj} + \frac{m+n-1}{m(n-1)}\sum_{j}^{m} I_{ij}.$$

Then, noting that

$$V_i = \frac{1}{m}\sum_{j}^{m} I_{ij}$$

and after some simplification of the leading term, we get

$$\text{pAUC}_i = \frac{m+n-1}{n-1}V_i - \frac{m}{n-1}\text{AUC}.$$

Similarly, the jackknife pseudovalue for the jth Y in the diseased group is as follows:

$$\text{pAUC}_j = \frac{m+n-1}{m-1}V_j - \frac{n}{m-1}\text{AUC}.$$

This link illustrates the difference in the scales and the reasons for the virtual equivalence of the two different forms of the variance of the AUC statistic.

In the simplest case where $m = n = N/2$,

$$\text{pAUC} = \frac{2N-2}{N-2}V - \frac{N}{N-2}\text{AUC}$$

for each of the $m + n$ observations.

The AUC pseudovalues fluctuate around the AUC with approximately twice the amplitude of the placements. This link also explains the difference in the form of the DeLong et al and jackknife variances of the AUC. The DeLong et al variance is a sum of two terms, S_Y^2/m and S_X^2/n, where S_Y^2 and S_X^2 are the variance of the placements (V's) for the Y and X samples. In contrast, the jackknife variance involves a single term $S^2/(m + n)$, where S^2 is the variance of all $m + n$ AUC pseudovalues. Because S^2 is approximately four times the size of S and S, then

$$\text{Var[AUC]} = \frac{S^2}{m+n} = \frac{4S_Y^2}{N} = \frac{4S_X^2}{N} = \frac{S_Y^2}{N/2} + \frac{S_X^2}{N/2} = \frac{S_Y^2}{m} + \frac{S_X^2}{n}.$$

The method used by DeLong et al has the advantage of directly showing the dependence of the SE on each of the two sample sizes m and n, whereas in the jackknife method, one simply sees the total sample size ($m + n$) in the denominator. If one of the two sample sizes were small, however, this would be reflected in a larger amplitude for the $m + n$ pseudovalues, since the observations from the smaller sample have a larger influence on the AUC.

PERFORMANCE OF TWO APPROACHES IN LARGER DATA SETS: A SIMULATION STUDY

In the worked example, the DeLong and jackknife formulas for the variance of the AUC produced slightly different answers. One may wonder whether this is a random phenomenon peculiar to this data set or whether the pattern would persist for other and for larger data sets. Thus, we conducted a Monte Carlo simulation to assess the performance of each method in the prediction of the sampling variability of the AUC estimates. As is shown in the four leftmost columns of

TABLE 6:Comparison between DeLong and Jackknife Methods in Estimating the SE of Nonparametric AUCs: Average SE Over 1,000 Data Sets Generated from Various Configurations of the Bivariate Binormal Model ($m = n = 50$)

Degree of Correlation	AUC			SE of the AUC or Difference in AUC		
	1	2	Difference	DeLong	Jackknife	Empirical
High	0.61	...	...	0.0566	0.0569	0.0581
	...	0.75	...	0.0488	0.0490	0.0487
	...	...	0.14	0.0387	0.0383	0.0378
Moderate	0.60	...	...	0.0568	0.0571	0.0583
	...	0.72	...	0.0508	0.0511	0.0501
	...	...	0.12	0.0507	0.0507	0.0502
Low	0.55	...	...	0.0578	0.0581	0.0586
	...	0.63	...	0.0557	0.0560	0.0541
	...	...	0.08	0.0725	0.0728	0.0721
High	0.75	...	...	0.0487	0.0490	0.0502
	...	0.90	...	0.0307	0.0308	0.0307
	...	...	0.15	0.0349	0.0348	0.0351
Moderate	0.72	...	...	0.0508	0.0511	0.0525
	...	0.86	...	0.0357	0.0359	0.0353
	...	...	0.14	0.0439	0.0439	0.0443
Low	0.63	...	...	0.0557	0.0559	0.0568
	...	0.74	...	0.0496	0.0498	0.0481
	...	...	0.11	0.0676	0.0678	0.0678

Table 6, the data concerning the presence of a signal were generated by a bivariate binormal model with different patterns of location parameters and various correlations (high, moderate, low). We generated 1,000 data sets with sample sizes of $m = n = 50$ for each configuration studied.

In each data set, we calculated the SE of the two accuracy indexes AUC_1 and AUC_2 and the SE of the difference in accuracies (ΔAUC) by using both the DeLong and the jackknife method. The average of the 1,000 calculated SEs was compared with the standard deviation of the 1,000 estimates of each accuracy index.

The results are given in the three rightmost columns of Table 6. The jackknife and DeLong methods yielded very similar estimates of the SEs of accuracy indexes AUC_1, AUC_2, and $AUC_2 - AUC_1$, confirming the pattern one would expect on the basis of the link between the two. We suspect the minor differences stem from the fact that $m + n - 1$ degrees of freedom are used in the jackknife method and $m - 1$ and $n - 1$ are used in the DeLong method. The ratio of average estimates of the SEs to the corresponding empirical SE ranged from 0.97 to 1.05 over the various configurations studied. When sample sizes of $m = n = 100$ were used, the ratio varied from 1.00 to 1.05.

DISCUSSION

For researchers who wish to calculate the SE associated with a nonparametric AUC or the difference between two such AUCs, the choices seem bewildering. The purpose of this review was to make sense of the two main approaches and show users an approach to calculation that is readily extendable to more than one AUC.

We have no strong preference between the method used by DeLong et al, which is based on placements, and the AUC pseudovalue approach, which is based on jackknifing. The structure of the former emphasizes the need for adequate samples from each of the two states to be distinguished (ie, m and n), whereas the pseudovalue method appears to emphasize the total sample size ($m + n$). In the end, however, both m and n do influence the SE. One feature that makes the approach used by DeLong et al more intuitive is that each placement is measured on the (0,1) scale, just like the AUC itself. Moreover, one can relate the "component variances" directly to the shape of the ROC curve. Although the individual jackknife pseudovalues also average out to the AUC, they vary over a wider range, from below 0 to above 1, and so are less "natural."

From a computational viewpoint, there is little reason to choose one over the other. While we have described the DeLong method first, and then showed that the jackknife pseudovalues could be obtained from the placements with Equations (7) and (8), one could just as easily have presented them in reverse order and calculated the placements from the pseudovalues.

Given the twinlike nature of the two methods, which approach should one adopt? We suggest that the jackknife approach has greater payoff in the longer run. The reason has to do with extensibly: The DeLong method deals only with AUCs; the jackknife method can deal with any ROC index (AUC, TPF at a given FPF, TPF averaged over a range of FPFs, etc), whether estimated parametrically or nonparametrically [14]. Indeed, the use of pseudovalues has already been extended to the case of multiple readers by Dorfman et al [13], and in another work [15] we propose that it can be used to advantage for dealing with multiple signals.

Finally, because we see a future for jackknifing, we want to comment on terminology. As with many statistical terms (such as "standard" deviation), the term pseudovalue is not entirely descriptive. One must emphasize therefore that "pseudo" values are not "false" values. If one were to be classic, the term "quasi" values would convey more of the meaning, namely, that they are "equivalent" values that can be substituted for the original values, and that from their distribution one is able to use conventional estimators for the uncertainty of a mean to calculate the reliability of the summary statistic in question. In the simplest case, the pseudovalues for the best understood of all summary statistics—the mean—are none other than the original observations themselves.

REFERENCES

1. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. *Crit Rev Diagn Imaging* **1989**; 29:307–335.
2. Swets JA. ROC analysis applied to the evaluation of diagnostic techniques. *Invest Radiol* **1979**; 14:109–121.
3. Metz CE. ROC methodology in radiologic imaging. *Invest Radiol* **1989**; 21:720–733.
4. Begg CB. Advances in statistical methodology for diagnostic medicine in the 1980's. *Stat Med* **1991**; 10:1887–1895.
5. Green DM, Swets JA. Signal detection theory and psychophysics. New York, NY: Wiley, **1966**.
6. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**; 143:29–36.
7. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating graph. *J Math Psychol* **1975**; 12:387–415.
8. Hanley JA, McNeil BJ. A method of comparing the area under receiver operating characteristic curves derived from the same cases. *Radiology* **1983**; 148:839–843.
9. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **1988**; 44:837–845.
10. Wieand S, Gail, MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* **1989**; 76:585–592.
11. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York, NY: Champan & Hall, **1993**.
12. McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Med Decis Making* **1984**; 4:137–150.
13. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. *Invest Radiol* **1992**; 27:723–731.
14. Hettmansperger TP. Statistical inference based on ranks. New York, NY: Wiley, **1984**.
15. Hajian-Tilaki KO, Hanley JA, Joseph L, Collet JP. An extension of ROC analysis to data concerning multiple "signals." *Acad Radiol* (in press).