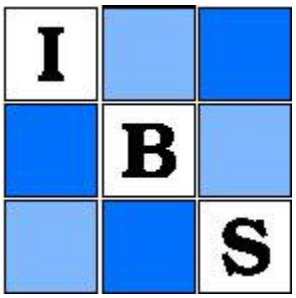

BIOMETRICS 53, 567–578
June 1997

Nonparametric Analysis of Clustered ROC Curve Data

Nancy A. Obuchowski

Department of Biostatistics and Epidemiology, The Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, Ohio 44195–5196, U.S.A.

SUMMARY

Current methods for estimating the accuracy of diagnostic tests require independence of the test results in the sample. However, cases in which there are multiple test results from the same patient are quite common. In such cases, estimation and inference of the accuracy of diagnostic tests must account for intracluster correlation. In the present paper, the structural components method of DeLong, DeLong, and Clarke-Pearson (1988, *Biometrics* **44,** 837–844) is extended to the estimation of the Receiver Operating Characteristic (ROC) curve area for clustered data, incorporating the concepts of design effect and effective sample size used by Rao and Scott (1992, *Biometrics* **48,** 577–585) for clustered binary data. Results of a Monte Carlo simulation study indicate that the size of statistical tests that assume independence is inflated in the presence of intracluster correlation. The proposed method, on the other hand, appropriately handles a wide variety of intracluster correlations, e.g., correlations between true disease statuses and between test results. In addition, the method can be applied to both continuous and ordinal test results. A strategy for estimating sample size requirements for future studies using clustered data is discussed.

1. Introduction

Receiver Operating Characteristic (ROC) curves are important tools for describing the accuracy of diagnostic tests. The ROC curve and its associated indices take into account both of the traditional measures of diagnostic accuracy, sensitivity and specificity. Sensitivity is the probability of correctly detecting the condition of interest among subjects with the condition. Specificity is the probability of correctly ruling out the condition among subjects without the condition. An ROC curve is a plot of a test's sensitivity (or true positive rate) versus false positive rate (or 1-specificity). The curve is constructed by changing the decision criterion that defines a positive test result. Although the area under the ROC curve is not the only measure of a test's accuracy, it is preferred over simple estimates of sensitivity or specificity because the area under the curve incorporates both of these measures of accuracy and accounts for the inherent trade-offs between them as the decision criterion changes (see Metz, 1978).

Both parametric and nonparametric methods have been developed to estimate and compare ROC areas (Dorfman and Alf, 1968, 1969; Metz and Kronman, 1980; Hanley and McNeil, 1982, 1983; Metz, Wang, and Kronman, 1984; DeLong et al., 1988; Wieand et al., 1989; Zhou and Gatsonis, 1996). However, all of these methods assume that the test results are independent observations, which, as illustrated by the following example, is not always the case.

Three-dimensional Magnetic Resonance Angiography (MRA) is a relatively new diagnostic procedure for quantifying the degree of arterial atherosclerotic stenosis. The procedure has been proposed as a screening tool for atherosclerosis of the carotid arteries. In a study by Masaryk et al. (1991), two radiologists evaluated 65 carotid arteries (left and right) in 36 patients using MRA. These patients also underwent intraarterial digital subtraction angiography (DSA), which is considered the gold standard for characterizing the degree of stenosis. The goals of the study were to estimate the accuracy of MRA for each reader using the area under the ROC curve as the index of diagnostic accuracy, and to compare the accuracy of the two radiologists. Since there is no biological reason why the accuracy of MRA would differ for left and right arteries, the precision of the estimate of

Key words: Effective sample size; Intracluster correlations; Mann–Whitney test; ROC curve.

accuracy could be improved by incorporating the results of both arteries into a single measure of accuracy.

A similar problem of clustered observations occurs when describing the accuracy of tests that require both detection and localization of disease. A good example is screening mammography, where there can be multiple lesions (some benign and some malignant) in the same patient. Treatment of malignant lesions is usually location-specific; thus, both detection and localization are critical. To characterize accuracy in a clinically meaningful way, images can be divided into small pre-defined sections, which could then replace the patient as the unit of analysis.

In studies such as these, in which there are multiple test results per subject, the intracluster correlation must be considered when estimating and drawing inferences about the accuracy of the diagnostic test. Rao and Scott (1992) presented a simple nonparametric method for the analysis of clustered binary data. Their procedure is based on the concepts of design effect and effective sample size used in sample surveys and can be applied directly for estimating sensitivity or specificity. Similarly, Smith and Hadgu (1992) described how generalized estimating equations can be used to estimate sensitivity or specificity from correlated binary data.

In the present paper, we describe a method for estimating the area under the ROC curve in the presence of clustered data. We use DeLong et al.'s (1988) structural components approach to ROC curve estimation but extend their method to clustered ROC data using the ideas of Rao and Scott (1992). The proposed method is nonparametric and thus does not require specification of the intracluster correlation structure. Furthermore, the method can be applied to either continuous or ordered categorical test results. Results from a Monte Carlo simulation study suggest that the method is robust to a variety of intracluster correlation patterns, as well as to non-normally distributed test results.

In Section 2, we develop methods for estimation and inference for a single ROC curve and for two correlated ROC curves. The latter case is important in comparing the accuracy of two readers or two diagnostic tests estimated from the same sample of subjects. Results of the Monte Carlo simulation study are presented in Section 3. We illustrate the proposed methods in Section 4 using the MRA data. In Section 5, we discuss sample size estimation for ROC studies with clustered data.

2. Notation

Let X_{ij} denote the test result of the jth affected unit in the ith cluster ($j = 1, 2, \ldots, m_i$), ($i = 1, 2, \ldots, I$). Similarly, let Y_{ik} denote the test result of the kth unaffected unit in the ith cluster ($k = 1, 2, \ldots, n_i$). These test results can be either continuous or ordered discrete random variables. The test results reflect a range of confidences about the presence of disease. Without loss of generality, we define larger values of the test results to indicate greater confidence about the presence of disease.

Let $s_i = m_i + n_i$ be the total number of units in cluster i. The total number of affected units in cluster i is given by $M = \sum_i (m_i)$, and the total number of unaffected units is given by $N = \sum_i (n_i)$. The total number of clusters with at least one affected unit is denoted by I_{10}, and the total number of clusters with at least one unaffected unit is denoted by I_{01}. Note that $I_{10} + I_{01}$ will often be greater than I, the total number of clusters in the sample.

2.1 *Estimation and Inference for a Single ROC Curve Area*

For independent data (one unit per cluster), the nonparametric estimate of the area under the ROC curve is (see Bamber, 1975)

$$\hat{\theta} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{i'=1}^{N} \psi(X_i, Y_{i'}), \tag{1}$$

where

$$\Psi(X, Y) = \begin{array}{ll} 1.0 & \text{if } Y < X \\ 0.5 & \text{if } Y = X \\ 0.0 & \text{if } Y > X. \end{array} \tag{2}$$

Note that the estimator in (1), rewritten as

$$\hat{\theta} = \frac{\displaystyle\sum_{i=1}^{M} \left(\frac{\displaystyle\sum_{i'=1}^{N} \psi(X_i, Y_{i'})}{N} \right)}{M},$$

is simply the average proportion of observations in the sample of unaffected units exceeded by each member of the sample of affected units. It is analogous to the Mann–Whitney (1947) formulation of the well-known Wilcoxon–Mann–Whitney U-statistic.

For clustered data, we consider the following estimate of the area under the curve:

$$\hat{\theta}_c = \frac{1}{MN} \sum_{i=1}^{I} \sum_{i'=1}^{I} \sum_{j=1}^{m_i} \sum_{k=1}^{n_{i'}} \psi(X_{ij}, Y_{i'k}). \tag{3}$$

This estimate gives equal weight to all pairwise rankings within and between clusters. Note that $\hat{\theta}_c$ reduces to $\hat{\theta}$ when all clusters have a single unit. For convenience, we drop the subscript c notation in the remainder of the paper.

Hanley and McNeil (1982), DeLong et al. (1988), and Wieand et al. (1989) have proposed different approaches to estimating the variance of $\hat{\theta}$ in the absence of clusters. Hanley and McNeil's estimate is based on work by Bamber (1975), and involves empirically estimating the quantities B_{XXY} and B_{YYX}. B_{XXY} is the probability that two randomly chosen affected units will both be ranked higher than a randomly chosen unaffected unit, and B_{YYX} is the probability that one randomly chosen affected unit will be ranked higher than two randomly chosen unaffected units. DeLong et al. use the method of structural components for U-statistics developed by Sen (1960) to provide consistent estimates of variance. Rockette et al. (1990) showed, by means of a Monte Carlo simulation, that the variance estimates of Hanley and McNeil and DeLong et al. are equivalent for large samples. The approach of Wieand et al. is the most cumbersome of the three; it does, however, handle other indices associated with the ROC curve. For the full ROC curve area index, Wieand et al. show that their variance is asymptotically equivalent to that of DeLong et al.

We extend the method of DeLong et al. to handle clustered data because their method involves computing sums of squares, which enables us to apply the ideas that Rao and Scott (1992) used for binary data. In our method, we first transform the test scores of affected units into X-components, and the test scores of unaffected units into Y-components. These X- and Y-components are computed from the kernel in (2). The X- and Y-components for clustered data are

$$V_{10}(X_{ij}) = \frac{1}{N} \sum_{i'=1}^{I_{01}} \sum_{k=1}^{n_{i'}} \psi(X_{ij}, Y_{i'k})$$

for all X_{ij}, and

$$V_{01}(Y_{i'k}) = \frac{1}{M} \sum_{i=1}^{I_{10}} \sum_{j=1}^{m_i} \psi(X_{ij}, Y_{i'k})$$

for all $Y_{i'k}$. Note $\sum_i \sum_j V_{10}(X_{ij})/M = \hat{\theta}$. Similarly, $\sum_i \sum_j V_{01}(Y_{ij})/N = \hat{\theta}$.

Next, the sums of squares of the X- and Y-components are computed, applying the ideas developed by Rao and Scott (1992) to deal with clustered binary data. Using results of Scott and Wu (1981), they provided a consistent estimate of the variance of a proportion by taking sums of squares of the number of affected units in a cluster. Similarly, we compute two sums of squares: the first involves the sum of the X-components for cluster i; the second involves the sum of the Y-components for cluster i.

Let $V_{10}(X_{i.})$ and $V_{01}(Y_{i.})$ be the sum of the X- and Y-components, respectively, for the ith cluster. When there are no affected units in cluster i, m_i and $V_{10}(X_{i.})$ equal zero. Similarly, when there are no unaffected units in cluster i, n_i and $V_{01}(Y_{i.})$ equal zero. Using the notation of DeLong et al., we denote the sum of squares of the X-components by S_{10}, and the sum of squares of the Y-components by S_{01}.

$$S_{10} = \frac{I_{10}}{(I_{10} - 1)M} \sum_{i=1}^{I_{10}} \left[V_{10}(X_{i.}) - m_i\hat{\theta}\right]^2$$

and

$$S_{01} = \frac{I_{01}}{(I_{01} - 1)N} \sum_{i=1}^{I_{01}} \left[V_{01}(Y_{i.}) - n_i\hat{\theta}\right]^2,$$

where $m_i\hat{\theta}$ is the mean sum of the X-components in clusters with m_i affected units, and $n_i\hat{\theta}$ is the mean sum of the Y-components in clusters with n_i unaffected units.

Note that we can write $\hat{\theta}$ in (3) as the ratio of two sample means,

$$\hat{\theta} = \frac{\sum_{i=1}^{I_{10}} V_{10}(X_{i.})/I_{10}}{\sum_{i=1}^{I_{10}} m_i/I_{10}}, \quad \text{or similarly,} \quad \hat{\theta} = \frac{\sum_{i=1}^{I_{01}} V_{01}(Y_{i.})/I_{01}}{\sum_{i=1}^{I_{01}} n_i/I_{01}}.$$

Then, following Scott and Wu (1981), Rao and Scott (1992), and DeLong et al. (1988) $(S_{10}/M + S_{01}/N)$ is a consistent estimator of the variance of $\hat{\theta}$ in the special case where there is no correlation between affected and unaffected units within the same cluster. However, for the more general case, we consider the following cross-product, S_{11}, which takes into account the correlation between affected and unaffected units within the same cluster,

$$S_{11} = \frac{I}{(I-1)} \sum_{i=1}^{I} \left(\left[V_{10}(X_{i.}) - m_i\hat{\theta} \right] \left[V_{01}(Y_{i.}) - n_i\hat{\theta} \right] \right).$$

Our estimator of the variance of $\hat{\theta}$ is then

$$\widehat{\text{var}}(\hat{\theta}) = \frac{1}{M} S_{10} + \frac{1}{N} S_{01} + \frac{2}{MN} S_{11}. \tag{4}$$

Following DeLong et al. (1988), $(\hat{\theta} - \theta)/\sqrt{\widehat{\text{var}}(\hat{\theta})}$ is asymptotically $N(0,1)$ if $\lim_{I \to \infty} I_{10}/I_{01}$ is bounded and nonzero.

2.2 *Estimation and Inference for Two Correlated ROC Curve Areas*

When comparing two ROC curve areas estimated from the same sample of clustered data, as when comparing the accuracy of the two readers in the MRA example, the covariance between the two estimated areas must be taken into account. DeLong et al. (1988) use results from Sen (1960) to derive consistent estimates of the elements of the variance–covariance matrix of a vector of ROC areas. We use the results of DeLong et al., again applying the ideas of Rao and Scott (1992) to handle the clustered data.

Let $\hat{\theta}^1$ and $\hat{\theta}^2$ denote the two estimated ROC areas. Define

$$S_{10}^{1,2} = \frac{I_{10}}{(I_{10}-1)M} \sum_{i=1}^{I_{10}} \left[V_{10}^1(X_{i.}) - m_i\hat{\theta}^1 \right] \left[V_{10}^2(X_{i.}) - m_i\hat{\theta}^2 \right]$$

and

$$S_{01}^{1,2} = \frac{I_{01}}{(I_{01}-1)N} \sum_{i=1}^{I_{01}} \left[V_{01}^1(Y_{i.}) - n_i\hat{\theta}^1 \right] \left[V_{01}^2(Y_{i.}) - n_i\hat{\theta}^2 \right],$$

where $V_{10}^r(X_{i.})$ is the sum of the X-components in cluster i from the rth ROC curve, and $V_{01}^r(Y_{i.})$ is the sum of the Y-components in cluster i from the rth ROC curve. The terms to account for the correlation between affected and unaffected units within a cluster are

$$S_{11}^{1,2} = \frac{I}{(I-1)} \sum_{i=1}^{I} \left[V_{10}^1(X_{i.}) - m_i\hat{\theta}^1 \right] \left[V_{01}^2(Y_{i.}) - n_i\hat{\theta}^2 \right]$$

and

$$S_{11}^{2,1} = \frac{I}{(I-1)} \sum_{i=1}^{I} \left[V_{10}^2(X_{i.}) - m_i\hat{\theta}^2 \right] \left[V_{01}^1(Y_{i.}) - n_i\hat{\theta}^1 \right].$$

The estimated covariance between between $\hat{\theta}^1$ and $\hat{\theta}^2$ is

$$\widehat{\text{cov}}(\hat{\theta}^1, \hat{\theta}^2) = \frac{S_{10}^{1,2}}{M} + \frac{S_{01}^{1,2}}{N} + \frac{S_{11}^{1,2}}{MN} + \frac{S_{11}^{2,1}}{MN}. \tag{5}$$

Our estimator of the variance of the difference between two correlated ROC curves is given in equation (6), where the variance of $\hat{\theta}^1$ and $\hat{\theta}^2$ are given in equation (4).

$$\widehat{\text{var}}(\hat{\theta}^1 - \hat{\theta}^2) = \widehat{\text{var}}(\hat{\theta}^1) + \widehat{\text{var}}(\hat{\theta}^2) - 2\,\widehat{\text{cov}}(\hat{\theta}^1, \hat{\theta}^2). \tag{6}$$

Following DeLong et al. (1988), $([\hat{\theta}^1 - \hat{\theta}^2] - [\theta^1 - \theta^2])/(\widehat{\text{var}}(\hat{\theta}^1 - \hat{\theta}^2))^{1/2}$ has a standard normal distribution if $\lim_{I\to\infty} I_{10}/I_{01}$ is bounded and nonzero.

3. Monte Carlo Simulation Study

A Monte Carlo simulation study was conducted to assess the size and power of Wald tests and the coverage of 95% confidence intervals (CIs) derived from the proposed estimators. For each assessment, 2000 simulated data sets were generated. Each data set consisted of 100 independent clusters ($I = 100$). This sample size of 100 clusters was chosen because it is a realistic size for many ROC curve studies.

For each cluster of size k, a random vector was generated from a k-variate normal distribution with a zero mean vector and a covariance matrix having a compound symmetry structure with a common correlation of 0.0, 0.4, 0.8, or 0.95. If the simulated value for the jth unit within a cluster was greater than zero, then the status of that unit was classified as affected; otherwise, the unit was classified as unaffected. This resulted in a prevalence rate of disease of 50%. Test results for units within each cluster were then generated from a second k-variate normal distribution with mean vector zero and a compound symmetry covariance matrix. A constant, δ, was added to the test results of units identified as affected, where δ was chosen in order to make the area under the ROC curve equal to either 0.7 or 0.8. Correlations for this second k-variate distribution of 0.0 (representing the case of independence of test results within clusters), 0.1, 0.4, and 0.8 were considered. For simulations where the intracluster correlation differed among clusters, correlations of 0.0, 0.1, 0.4, and 0.8 were randomly assigned with equal probability to the clusters. Finally, in simulations where the cluster size varied, 10% of the units were randomly and independently deleted.

This process of data simulation enabled us to specify the correlation between disease statuses as well as the correlation between test results of units within a cluster. This distinction is important. In particular, it is likely that disease statuses of units within a cluster will be correlated. However, test results within a cluster, conditional on disease status, may or may not be correlated. Both possibilities were assessed here.

We also assessed the size, power, and coverage of 95% CIs when the test results are not normally distributed. We chose to model the distribution of test results that occurs when the test results are readers' confidence scores. Confidence scores are used when the diagnostic test requires a subjective interpretation, such as a mammogram for detecting breast cancer. Reader confidence is often recorded using a 0–100% scale, where 0% is no confidence that the unit is affected, and 100% is complete confidence that the unit is affected. Thus, the test results range from 0–100%, often with modes at 0% and 100%. To model this distribution, we simulated data from normal distributions, but we truncated the tails of the distributions. For example, for unaffected units, the test results were simulated from a standard normal distribution, but all realized values less than -0.84 (cumulative value of the standard normal distribution at 20%) were set to -0.84. Similarly, all affected units with test results less than -0.84 were set to -0.84. We used an analogous strategy for the test results of the affected units, in this case truncating the upper tail.

3.1 *Single ROC Curve Area*

For the single ROC curve, the coverage of the 95% CIs for the true area was assessed. The CIs were constructed from the variance estimate given in equation (4). For this assessment, the true area under the curve was 0.7. Table 1 summarizes the estimated standard error of $\hat{\theta}$ from (4), the empirical standard error of $\hat{\theta}$, and the 95% CIs for the coverage. The results are compared to those obtained when independence is assumed between units of a cluster (i.e., results based on the method of DeLong et al. (1988) without the modification for clustered data). Table 1 presents the case of normally distributed test results and (1) equal-sized clusters (i.e., two units per cluster) with constant intracluster correlation for all clusters, (2) unequal-sized clusters (1–3 units per cluster) with constant intracluster correlation for all clusters, and (3) unequal-sized clusters with variable intracluster correlation. Table 1 also presents the case of non-normally distributed test results and unequal-sized clusters with constant intracluster correlation for all clusters.

When the test results are independent, conditional on the true disease status, both estimates of the variance of $\hat{\theta}$, the one based on the conditional independence assumption and the one for

Table 1
Monte Carlo simulation results for single ROC curve area: 95% *CIs for coverage*

Correlation between		Empirical	Independence assumed	Equation (4)
Disease status	Test results	SE	coverage (SE)	coverage (SE)
Normally distributed test results				
Two units/cluster				
0.0	0.0	0.036	94.6–96.4 (0.037)	92.7–94.9 (0.037)
0.4	0.0	0.036	94.5–96.3 (0.037)	93.7–95.7 (0.037)
0.8	0.0	0.037	93.6–95.6 (0.037)	93.4–95.4 (0.037)
0.4	0.1	0.037	94.3–96.1 (0.037)	93.6–95.6 (0.037)
0.4	0.4	0.039	92.9–94.9 (0.037)	93.2–95.2 (0.039)
0.4	0.8	0.041	90.8–93.2 (0.037)	93.3–95.3 (0.041)
0.8	0.1	0.038	93.2–95.2 (0.037)	93.4–95.4 (0.038)
0.8	0.4	0.041	90.4–92.8 (0.037)	93.5–95.5 (0.042)
0.8	0.8	0.045	86.9–89.7 (0.037)	93.4–95.4 (0.045)
1–3 units/cluster				
0.0	0.0	0.031	95.6–97.2 (0.032)	93.9–95.9 (0.032)
0.4	0.0	0.031	94.8–96.6 (0.032)	94.3–96.1 (0.032)
0.8	0.0	0.031	93.7–95.7 (0.032)	93.2–95.2 (0.032)
0.4	0.1	0.032	94.0–96.0 (0.032)	94.5–96.3 (0.032)
0.4	0.4	0.034	91.6–93.8 (0.032)	94.0–96.0 (0.034)
0.4	0.8	0.038	88.8–91.4 (0.032)	93.0–95.0 (0.038)
0.8	0.1	0.033	92.2–94.4 (0.032)	92.6–94.8 (0.033)
0.8	0.4	0.038	87.8–90.6 (0.032)	93.2–95.2 (0.037)
0.8	0.8	0.044	82.6–85.8 (0.032)	93.2–95.2 (0.043)
1–3 units/cluster				
0.4	variable	0.034	90.7–93.1 (0.032)	92.1–94.3 (0.034)
0.8	variable	0.037	88.8–91.4 (0.032)	93.0–95.0 (0.036)
Nonnormally distributed test results				
1–3 units/cluster				
0.4	0.1	0.032	94.0–96.0 (0.032)	94.7–96.5 (0.033)
0.4	0.4	0.034	92.0–94.2 (0.032)	95.1–96.9 (0.036)
0.4	0.8	0.038	87.9–90.7 (0.032)	95.0–96.8 (0.041)
0.8	0.1	0.033	92.7–94.9 (0.032)	93.5–95.5 (0.033)
0.8	0.4	0.037	89.4–91.0 (0.032)	93.8–95.8 (0.038)
0.8	0.8	0.043	83.4–86.6 (0.032)	94.2–96.0 (0.044)

clustered data, provide appropriate coverage for θ. This is true even when disease statuses are highly correlated within clusters. However, when the test results are correlated, the estimator based on the conditional independence assumption leads to insufficient coverage. The coverage decreases as the correlation between test results increases. Also, note that for a constant correlation between test results, the coverage decreases as the correlation between the true disease statuses increases. Intuitively, this makes sense because the correlation between test results of like units (both affected or both unaffected) decreases the precision of estimates of sensitivity and specificity much more than does the correlation between test results of unlike units, where correlation in test results is dampened by differences in disease status.

The estimator for clustered data [equation (4)] provides appropriate coverage for θ for equal and unequal cluster sizes and for a wide range of correlations between test results as well as between disease statuses.

3.2 *Two Correlated ROC Curve Areas*

For two correlated ROC curves, the size and power of the Wald test comparing the areas and the coverage of the 95% CI for the difference in areas were assessed. The Wald tests and CIs were constructed from the estimated variance of the difference, given in equation (6). When assessing the size of the Wald tests, the true areas under the two curves were 0.7; a significance level of 0.05 was used. When assessing the power of the Wald tests and coverage of the 95% CIs, the true areas under the two curves were 0.7 and 0.8. The correlations between the results of the two diagnostic

tests (or two readers) were specified as follows: the correlation between the test results of the same unit within a cluster was arbitrarily set to 0.5; for different units within the same cluster, the correlation between test results was set to one-half of the between-unit within-cluster within-test correlation (i.e., one-half of 0.1, 0.4, or 0.8).

Table 2 summarizes the 95% CIs for the size, power, and coverage. In the presence of intracluster correlation of tests results, the coverage is inadequate when independence is assumed and the size of the test exceeds 5%. On the other hand, Wald tests constructed from the estimator in (6) have appropriate size and coverage.

4. Illustrative Example

The data from the MRA example are given in Table 3. The MRA test results according to DSA findings are summarized: both arteries with hemodynamically significant disease (>70% by DSA), left artery with significant disease, right artery with significant disease, and neither artery with significant disease. Of the 36 patients studied, seven had test results from only one carotid artery, resulting in unequal-sized clusters.

Using a criterion of >70 to define a positive test result, the estimated sensitivity and specificity of Reader 1 are 86% (25/29) and 92% (33/36), respectively. A surgeon who wants to use the MRA results to identify surgical patients may require a greater specificity than 92%. Using a criterion

Table 2

Monte Carlo simulation results for two correlated ROC curve areas: 95% CIs for size, power, and coverage

Correlation between		Independence assumed			Equation (6)		
Disease status	Test results	Size	Power	Coverage	Size	Power	Coverage
Normally distributed test results							
Two units/cluster							
0.0	0.0	3.3–5.2	77.2–80.8	93.4–95.4	3.4–5.2	77.1–80.7	93.0–95.0
0.4	0.0	3.4–5.2	77.4–81.0	93.8–95.8	3.7–5.5	77.7–81.3	93.7–95.7
0.8	0.0	3.5–5.3	77.5–81.1	95.3–96.9	3.9–5.7	77.6–81.2	95.1–96.9
0.4	0.1	3.5–5.3	77.0–80.6	93.5–95.5	4.1–6.1	76.2–79.8	93.8–95.8
0.4	0.4	4.6–6.6	75.9–79.5	92.9–94.9	4.0–5.8	72.7–76.5	93.7–95.7
0.4	0.8	6.2–8.4	74.2–78.0	90.7–93.1	3.9–5.7	67.0–71.0	93.8–95.8
0.8	0.1	4.1–6.1	77.3–80.9	94.6–96.4	3.9–5.7	76.1–79.7	94.8–96.6
0.8	0.4	5.2–7.4	74.7–78.5	92.1–94.3	3.8–5.6	69.5–73.5	94.3–96.1
0.8	0.8	8.5–11.1	72.4–76.2	87.4–90.2	3.7–5.5	59.2–63.4	93.9–95.9
1–3 units/cluster							
0.0	0.0	4.0–5.8	89.3–91.9	95.0–96.8	3.9–5.7	89.2–91.8	94.5–96.3
0.4	0.0	4.1–6.1	89.2–91.8	94.6–96.4	4.5–6.5	88.9–91.5	94.0–96.0
0.8	0.0	4.2–6.2	89.3–91.9	94.8–96.6	4.5–6.5	89.2–91.8	94.5–96.3
0.4	0.1	4.5–6.5	88.9–91.5	94.2–96.0	4.4–6.4	87.8–90.6	94.3–96.1
0.4	0.4	5.6–7.8	87.0–89.8	91.7–93.9	4.4–6.4	84.0–87.0	93.8–95.8
0.4	0.8	8.5–11.1	85.1–88.1	88.4–91.0	3.5–5.3	76.2–79.8	93.7–95.7
0.8	0.1	5.3–7.5	88.2–90.8	93.7–95.7	4.7–6.7	87.2–90.0	94.4–96.2
0.8	0.4	7.9–10.5	85.1–88.1	90.0–92.4	4.3–6.3	76.8–80.4	94.3–96.1
0.8	0.8	12.7–15.7	81.2–84.6	83.3–86.5	4.6–6.6	64.7–68.9	94.0–96.0
1–3 units/cluster							
0.4	variable	7.2–9.6	85.7–88.7	90.5–92.9	5.1–7.3	82.7–85.9	92.5–94.7
0.8	variable	8.3–10.9	84.4–87.4	89.5–92.1	5.1–7.3	77.2–80.8	93.0–95.0
Nonnormally distributed test results							
1–3 units/cluster							
0.4	0.1	4.0–6.0	89.1–91.7	93.9–95.9	4.9–6.9	87.3–90.1	93.5–95.5
0.4	0.4	6.1–8.3	87.3–90.1	92.0–94.2	4.8–6.8	82.4–85.6	93.1–95.1
0.4	0.8	8.8–11.4	84.4–87.4	87.8–90.6	4.1–6.1	73.3–77.1	93.2–95.2
0.8	0.1	4.6–6.6	88.4–91.0	92.9–94.9	4.8–6.8	88.1–88.9	92.9–94.9
0.8	0.4	8.2–10.8	85.1–88.1	88.8–91.4	4.7–6.7	76.2–79.8	93.2–95.2
0.8	0.8	12.7–15.7	81.4–84.6	82.2–85.4	4.3–6.3	63.1–67.3	94.2–96.0

of >75 to define a positive test result, the estimated sensitivity and specificity of Reader 1 are 83% (24/29) and 97% (35/36), respectively. In Figure 1, the ROC curve for Reader 1 depicts the trade-off between sensitivity and specificity as the decision criterion changes.

The area under the curve describes the inherent ability of MRA to distinguish between arteries with significant versus insignificant disease. Masaryk et al. (1991) estimated the area under the ROC curve separately for both the right and left arteries for both readers. For Reader 1, the estimated ROC areas are: left, 0.986 (SE = 0.016); right, 0.988 (SE = 0.014). We now apply the methods in Section 2 to show how greater precision can be gained by incorporating the data from both arteries into a single estimate of the area under the curve.

The MRA results, conditional on the true disease status, are correlated (for Reader 1: Pearson correlation coefficient (r) = 0.53, p value = 0.043; for Reader 2: r = 0.83, p value < 0.001), although the true disease statuses are not correlated (contingency coefficient = 0.044, $p = 0.812$). The estimator in equation (4) is needed whenever the test results are correlated, regardless of whether the true disease statuses are correlated. The estimated area for Reader 1 is 0.984 (SE = 0.011) with 95% CI of [0.963, 1.00]; and for Reader 2 is 0.985 (SE = 0.010) with 95% CI of [0.966, 1.00].

Next, we compare the accuracy of the two readers. From equation (5), the estimated covariance of the two ROC areas is 0.00008. The estimated difference between the ROC areas is −0.001 (SE = 0.007). The Wald statistic is −0.14 with associated p value of 0.89. We conclude that there is insufficient evidence to suggest that the accuracy of the readers differ. The 95% CI for the difference between the two readers' accuracy is $[-0.015, 0.013]$.

5. Sample Size Estimation for Study Designs with Clustered Data

Estimates of the required sample size for studies with clustered data can be derived from the proposed method. All that is needed for computing sample size is: (1) an estimate of the sample size required if all observations were independent [see Hanley and McNeil (1982, 1983) and Obuchowski (1994)], (2) the average number of units per cluster, and (3) an estimate of the average correlation between test results of units within a cluster. Given these, an upper bound on the required sample size can be determined.

The required sample size for clustered data is a function of the design effect, d, which is the ratio of the variance of correlated data to the variance of uncorrelated data. Thus, M'/d is the effective sample size for the clustered study design, where M' is the number of units required for an unclustered study design [see Rao and Scott (1992) and Kish (1965)].

Table 3
MRA test results according to DSA finding

Both arteries[a]				Left artery[b]				Right artery[c]				Neither artery[d]			
Reader 1		Reader 2		Reader 1		Reader 2		Reader 1		Reader 2		Reader 1		Reader 2	
Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right
87	79	87	83	77	—	77	—	—	75	—	75	−3	0	−14	−11
88	95	94	93	94	45	89	44	70	70	73	81	47	−94	−3	−90
100	68	100	79	89	−11	92	−17	26	86	27	93	−16	75	17	85
65	89	61	91	95	30	91	14	0	65	8	70	5	−1	−22	−14
97	100	97	100	95	10	95	−122	6	100	23	100	−6	—	−52	—
100	89	99	88	100	19	100	−3	76	96	79	95	37	47	38	35
								44	100	55	100	42	4	40	6
								73	97	67	94	55	55	23	45
								—	85	—	86	4	−13	−87	−65
								58	100	60	99	49	35	24	9
								−2	100	−2	100	—	53	—	66
												2	—	−15	—
												−48	—	−78	—

$I = 36$, $I_{10} = 23$, $I_{01} = 27$, $M = 29$, and $N = 36$. For Reader 1: $S_{10} = 0.00132$, $S_{01} = 0.00224$, and $S_{11} = 0.00518$. For Reader 2: $S_{10} = 0.00093$, $S_{01} = 0.00226$, and $S_{11} = -0.00050$. For comparing the accuracy of the two readers: $S_{10}^{1,2} = 0.00085$, $S_{01}^{1,2} = 0.00192$, $S_{11}^{1,2} = 0.00286$, and $S_{11}^{2,1} = -0.00151$.

[a] Both arteries with hemodynamically significant disease (> 70%).

[b] The left artery with hemodynamically significant disease (> 70%).

[c] The right artery with hemodynamically significant disease (> 70%).

[d] Neither artery with hemodynamically significant disease (> 70%).

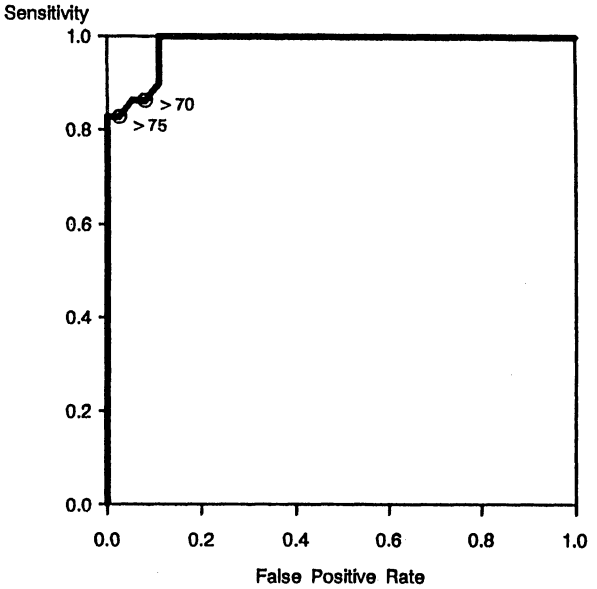

Figure 1. Empirical ROC curve for Reader 1. With a criterion of >70, the estimated sensitivity and specificity are 86% and 92%, respectively. With a criterion of >75, the estimated sensitivity and specificity are 83% and 97%.

Consider the estimator for the variance of the ROC area for clustered data given in equation (4). The expectation of S_{10} is the expectation of the square of the difference between the sum of the X-components within a cluster and its mean

$$\mathrm{E}[S_{10}] \approx \frac{I_{10}}{M}\mathrm{E}[V_{10}(X_\cdot) - m\mathrm{E}(V_{10}(X))]^2.$$

For uncorrelated data, $\mathrm{E}[S_{10}] \approx \{I_{10}/M\} \times m\,\mathrm{var}(V_{10}(X))$. For correlated data, $\mathrm{E}[S_{10}] \approx \{I_{10}/M\} \times \{m\,\mathrm{var}(V_{10}(X)) + m(m-1)\mathrm{cov}(V_{10}(X_i), V_{10}(X_j))\}$, where $i \neq j$. Here, we define m as the average number of affected units among patients with disease. Similarly, we can define n as the average number of unaffected units among patients with unaffected units. For uncorrelated data, $\mathrm{E}[S_{01}] \approx \{I_{01}/N\} \times n\,\mathrm{var}(V_{01}(Y))$. For correlated data, $\mathrm{E}[S_{01}] \approx \{I_{01}/N\} \times \{n\,\mathrm{var}(V_{01}(Y)) + n(n-1)\mathrm{cov}(V_{01}(Y_i), V_{01}(Y_j))\}$, where $i \neq j$. The portion of the variance due to the correlation between affected and unaffected units within the same cluster (i.e., S_{11}) is relatively small, so we ignore it here. Then, d is given by

$$d = \frac{\mathrm{var}_{\mathrm{correlated}}}{\mathrm{var}_{\mathrm{uncorrelated}}} = \frac{m\,\mathrm{var}(V)(1 + (m-1)r_x)N^2 I_{10} + n\,\mathrm{var}(V)(1 + (n-1)r_y)M^2 I_{01}}{m\,\mathrm{var}(V)N^2 I_{10} + n\,\mathrm{var}(V)M^2 I_{01}}$$

where r_x is the correlation in X-components between affected units within a cluster, and r_y is the correlation in Y-components between unaffected units within a cluster. In designing a study, the distinct variances of the test results of the affected and unaffected units would seldom be known, thus we set $\mathrm{var}(V_{10}[X]) = \mathrm{var}(V_{01}[Y]) = \mathrm{var}(V)$.

Let $P = I_{10}/I$, the prevalence of disease. Let $f = m/s$, the average proportion of affected units among patients with disease, s being the average number of total units per cluster. Note that $n = N/I_{01}$; $M = P \times I \times f \times s$; and $N = (I \times s) - M = I \times s \times (1 - Pf)$. Then, d can be rewritten as

$$d = (1 - Pf)(1 + [fs - 1]r_x) + Pf(1 + [Is(1 - Pf)/I_{01} - 1]r_y). \tag{7}$$

Note that if $r_x = r_y = 0$, then $d = 1$.

I_{01} cannot be determined precisely from P, f, I, and s unless $f = 1$. The minimum value that I_{01} can take is: $I-$ integer part of $[I \times P \times (fs-1)/(s-1)]$. The maximum value that I_{01} can take is as follows: if $s(1-f) \geq 1$, then the maximum is I; if $s(1-f) < 1$, then the maximum equals $s(1-f) \times I \times P + (1-P)I$. Note that if $f = 1$, then $I_{01} = I - I_{10}$.

If f is unknown in designing a study, then a conservative estimate of the required sample size can be obtained by setting $f = 1$ in equation (7). Also, if we assume that the correlation between

structural components is the same for affected and unaffected units, i.e., $r_x = r_y = r$, then d simplifies to

$$d = 1 + (s - 1)r. \tag{8}$$

In Table 4, the design effect for studying the area under a single ROC curve is given. The empirical design effect from the Monte Carlo simulation study is compared with the design effect estimated from equations (7) and (8). P and f were determined from the simulated data in order to compute d from equation (7). The correlation r was obtained from a table by Hanley and McNeil (1983), where the correlation between estimated areas is given as a function of the area under the curve and the correlation in the test results.

For example, in the first row of Table 4 the empirical design effect is $0.037^2/0.036^2 = 1.056$ (from Table 1). In comparison, from equation (7), $d = 1.041$ (note that the minimum and maximum value of I_{01} is 69) and from equation (8), $d = 1.09$. In general, estimates based on equation (8) will provide a reasonable upper bound for the required sample size.

Consider as an example a study where it has been determined that a sample size of 100 diseased patients and 75 controls (i.e., 175 units from 175 patients) is required to achieve a desired precision in estimating the area under the ROC curve. For a study design with clustered data, more units will be required, but fewer patients. If we expect two units per patient and if the intracluster correlation between test results is 0.10, then $d = 1.09$ (from equation (8), using $r = 0.09$) and the upper bound for the required sample size is 109 diseased units and 82 control units (i.e., 191 total units from between 96 and 109 patients, depending on the value of f). If there are three units per patient, then $d = 1.18$ and the upper bound for the required sample size is 118 diseased units and 89 control units (i.e., 207 total units from between 69 and 118 patients).

6. Concluding Remarks

The main purpose of this paper was to present a simple method of estimating and comparing ROC curve areas from clustered data. However, the method proposed here has a much broader application than just diagnostic radiology. In particular, Bamber (1975) pointed out that the nonparametric estimate of the area under the ROC curve is connected to the two-sample Wilcoxon or Mann–Whitney test statistic. This is evident because the Wilcoxon statistic and the area under the ROC curve both measure the probability of correctly ranking randomly chosen subjects from two populations. Thus, in situations where the Wilcoxon statistic would be appropriate, except for the fact that the data are clustered, the methods proposed here for estimating a single ROC curve area and its variance can be applied. A test of the hypothesis that the area under the curve equals 0.5 is analogous to a test of the hypothesis that two populations have equivalent locations.

Table 4
Empirical versus estimated design effect

Correlation between				
Disease status	Test results	Empirical	Equation (7)	Upper bound equation (8)
Two units/cluster				
0.4	0.1	1.056	1.041	1.09
0.4	0.4	1.174	1.170	1.37
0.4	0.8	1.297	1.354	1.77
($P = 0.68$, $f = 0.73$, $s = 2$)				
0.8	0.1	1.114	1.059	1.09
0.8	0.4	1.297	1.242	1.37
0.8	0.8	1.563	1.503	1.77
($P = 0.60$, $f = 0.83$, $s = 2$)				
1–3 units/cluster				
0.4	0.1	1.066	1.058–1.086	1.153
0.4	0.4	1.203	1.239–1.352	1.629
0.4	0.8	1.503	1.497–1.732	2.309
($P = 0.74$, $f = 0.67$, $s = 2.7$)				

From Hanley and McNeil (1983), for $\theta = 0.7$ and correlation between test results of 0.1, 0.4, and 0.8, r equals 0.09, 0.37, and 0.77, respectively.

ACKNOWLEDGEMENTS

The author thanks Dr Mark Schluchter for constructive comments on the presentation of this paper. This work was supported in part by NIH grant 2RO1HL43812-04.

RÉSUMÉ

Les méthodes habituelles d'estimation de la précision du diagnostic de tests supposent l'indépendance des résultats du test dans l'échantillon. Néanmoins, de multiples résultats de tests provenant du même patient sont monnaie courante, comme les artères carotides gauche ou droite. Estimation et inférence sur la précision du diagnostic de tests doivent tenir compte de cette corrélation intragroupe. La méthode des composantes structurales de DeLong, DeLong et Clarke-Pearson (1988, *Biometrics* **44,** 837–844) est étendue à l'estimation de la surface sous la courbe ROC à des données regroupées. Cette extension inclut les concepts d'effet expérimental et de taille efficace d'échantillon utilisés par Rao et Scott (1992, *Biometrics* **48,** 577–585) pour des données binaires groupées. Les résultats d'étude de simulation de Monte Carlo indiquent que la taille pour des tests statistiques supposant l'indépendance est augmentée en présence de corrélation intragroupe. Par ailleurs, la méthode proposée traite de manière appropriée une grande variété de corrélations intragroupe, à savoir corrélation entre statuts réels de maladie et/ou entre résultats de tests. De plus, la méthode peut être appliquée à des résultats de tests, à la fois continus et discrets. Nous discutons d'une stratégie pour estimer les tailles nécessaires d'un échantillon pour des études ultérieures utilisant des données groupées.

REFERENCES

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* **12,** 387–415.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44,** 837–844.

Dorfman, D. D. and Alf, E., Jr. (1968). Maximum-likelihood estimation of parameters of signal detection theory—A direct solution. *Psychometrika* **33,** 117–124.

Dorfman, D. D. and Alf, E., Jr. (1969). Maximum-likelihood estimation of parameters of signal detection theory and determination of confidence intervals—Rating-method data. *Journal of Mathematical Psychology* **6,** 487–496.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143,** 29–36.

Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **148,** 839–843.

Kish, L. (1965). *Survey Sampling.* New York: Wiley.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* **18,** 50–60.

Masaryk, A. M., Ross, J. S., DiCello, M. C., Modic, M. T., Paranandi, L., and Masaryk, T. J. (1991). 3DFT MR angiography of the carotid bifurcation: Potential and limitations as a screening examination. *Radiology* **179,** 797–804.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine* **VIII,** 283–298.

Metz, C. E. and Kronman, H. B. (1980). Statistical significance tests for binormal ROC curves. *Journal of Mathematical Psychology* **22,** 218–243.

Metz, C. E., Wang, P-L., and Kronman, H. B. (1984). A new approach for testing the significance of differences between ROC curves measured from correlated data. In *Information Processing in Medical Imaging,* F. Deconinck (ed), 432–445. The Netherlands: Nijhoff.

Obuchowski, N. A. (1994). Computing sample size for receiver operating characteristic studies. *Investigative Radiology* **29,** 238–243.

Rao, J. N. K. and Scott, A. J. (1992). A simple method for the analysis of clustered binary data. *Biometrics* **48,** 577–585.

Rockette, H. E., Obuchowski, N., Metz, C. E., and Gur, G. (1990). Statistical issues in ROC curve analysis. SPIE, Volume 1234. Medical Imaging IV: PACS System Design and Evaluation, 111–119.

Scott, A. J. and Wu, C. F. J. (1981). On the asymptotic distribution of ratio and regression estimators. *Journal of the American Statistical Association* **76,** 98–102.

Sen, P. K. (1960). On some convergence properties of U-statistics. *Calcutta Statistical Association Bulletin* **10,** 1–18.

Smith, P. J. and Hadgu, A. (1992). Sensitivity and specificity for correlated observations. *Statistics in Medicine* **11,** 1503–1509.

Wieand, S., Gail, M. H., James, B. R., and James, K. L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* **76,** 585–592.

Zhou, X. H. and Gatsonis, C. A. (1996) A simple method for comparing correlated ROC curves using incomplete data. *Statistics in Medicine* **15,** 1687–1693.

Received February 1996; *revised September* 1996; *accepted November* 1996.