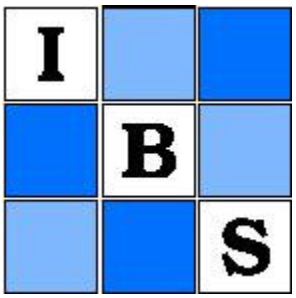

BIOMETRICS 44, 837–845
September 1988

Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach

Elizabeth R. DeLong

Quintiles, Inc., 1829 East Franklin Street,
Chapel Hill, North Carolina 27514, U.S.A.

David M. DeLong

SAS Institute, Cary, North Carolina 27511, U.S.A.

and

Daniel L. Clarke-Pearson

Division of Oncology, Department of OBGYN, Duke University Medical Center,
Durham, North Carolina 27710, U.S.A.

SUMMARY

Methods of evaluating and comparing the performance of diagnostic tests are of increasing importance as new tests are developed and marketed. When a test is based on an observed variable that lies on a continuous or graded scale, an assessment of the overall value of the test can be made through the use of a receiver operating characteristic (ROC) curve. The curve is constructed by varying the cutpoint used to determine which values of the observed variable will be considered abnormal and then plotting the resulting sensitivities against the corresponding false positive rates. When two or more empirical curves are constructed based on tests performed on the same individuals, statistical analysis on differences between curves must take into account the correlated nature of the data. This paper presents a nonparametric approach to the analysis of areas under correlated ROC curves, by using the theory on generalized *U*-statistics to generate an estimated covariance matrix.

1. Introduction

Methods of evaluating and comparing the performance of diagnostic tests or indices are of increasing importance as new tests or indices are developed or measured. When a test is based on an observed variable that lies on a continuous or graded scale, an assessment of the overall value of the test can be made through the use of a receiver operating characteristic (ROC) curve (Hanley and McNeil, 1982; Metz, 1978). The underlying population curve is theoretically given by varying the cutpoint used to determine the values of the observed variable to be considered abnormal and then plotting the resulting sensitivities against the corresponding false positive rates. If a test could perfectly discriminate, it would have a value above which the entire abnormal population would fall and below which all normal values would fall (or vice versa). The curve would then pass through the point (0, 1) on the unit grid. The closer an ROC curve comes to this ideal point, the better its discriminating ability. A test with no discriminating ability will produce a curve that follows the diagonal of the grid.

For statistical analysis, a recommended index of accuracy associated with an ROC curve is the area under the curve (Swets and Pickett, 1982). The area under the population ROC

Key words: Jackknifing; Mann–Whitney test; Receiver operating characteristic (ROC) curve; Structural components; *U*-statistics.

curve represents the probability that, when the variable is observed for a randomly selected individual from the abnormal population and a randomly selected individual from the normal population, the resulting values will be in the correct order (e.g., abnormal value higher than the normal value). Generally, parametric assumptions are applied on the distributions of the observed variable in the normal and the abnormal populations. Maximum likelihood programs for estimating the area under the curve and relevant parameters under a binormal model assumption have been widely employed (Dorfman and Alf, 1969; Metz, 1978; Swets and Pickett, 1982) in order to estimate this area, although these distributions cannot be uniquely determined from the ROC curve. The methodology has been extended (Metz, Wang, and Kronman, 1984) to a "bivariate binormal" model for testing differences between correlated sample ROC curves that arise, for example, when different diagnostic tests are performed on the same individuals.

This paper addresses the nonparametric comparison of areas under correlated ROC curves. When calculated by the trapezoidal rule, the area falling under the points comprising an empirical ROC curve has been shown to be equal to the Mann–Whitney U-statistic for comparing distributions of values from the two samples (Bamber, 1975). Although the trapezoidal rule systematically underestimates the true area (Hanley and McNeil, 1982; Swets and Pickett, 1982) when the number of distinct values taken on by a discrete-valued diagnostic variable is small (say, 5 or 6), it nonetheless produces a meaningful statistic that can be used with confidence when the variable takes on a larger number of values. Hanley and McNeil (1983) use some properties of this nonparametric statistic to compare areas under ROC curves arising from two measures applied to the same individuals. Their approach involves calculating for both the normal and the abnormal sample the correlation between the values of the original measures. The average of the two correlations is used along with the average of the areas under the two curves to arrive at an estimated correlation between the two areas. A table that applies when the average area is at least .70 is given. However, for measures that are not continuous or nearly so, their method relies on Gaussian modeling assumptions for estimating the variances of the two areas. In Section 2 we present an alternative methodology using a more completely nonparametric approach which exploits the properties of the Mann–Whitney statistic. Section 3 presents an example of three correlated ROC curves derived from data on ovarian cancer patients undergoing surgery for bowel obstruction. Three different prognostic indices are evaluated and compared.

2. Analysis of Areas Under Correlated ROC Curves

Suppose a sample of N individuals undergo a test for predicting an event of interest or determining presence or absence of a medical condition and that the test is based on a continuous-valued diagnostic variable. We will follow the convention that higher values of the test variable are assumed to be associated with the event of interest, e.g., positive disease status. Also suppose it can be determined by means independent of the test that m of these individuals truly undergo the event or have the condition. Let this group be denoted by C_1 and let the group of n $(= N - m)$ individuals who do not have the condition be denoted by C_2. Let X_i, $i = 1, 2, \ldots, m$ and Y_j, $j = 1, 2, \ldots, n$ be the values of the variable on which the diagnostic test is based for members of C_1 and C_2, respectively. These outcome values can be used to construct an empirical ROC curve for assessing the diagnostic performance of the test. For any real number z, let

$$\text{sens}(z) = \frac{1}{m} \sum_{i=1}^{m} I(X_i \geq z)$$

where $I(A) = 1$ if A is true and 0 otherwise. Also let

$$\text{spec}(z) = \frac{1}{n} \sum_{j=1}^{n} I(Y_j < z).$$

Then sens(z) is the empirical sensitivity of a test that is derived by dichotomizing the variable into positive or negative results on the basis of the cutpoint z and spec(z) is the corresponding empirical specificity. Now, as z varies over the possible values of the variable, the empirical ROC curve is a plot of sens(z) versus [1 − spec(z)]. Clearly, when z is larger than the largest possible value, the curve passes through (0, 0) and it monotonically increases to the point (1, 1) as z decreases to the smallest possible value. To be informative, the entire curve should lie above the 45° line where sens(z) = 1 − spec(z). Selection of an optimal cutpoint depends on a cost function of sensitivity and specificity.

It has been shown that the area under an empirical ROC curve, when calculated by the trapezoidal rule, is equal to the Mann–Whitney two-sample statistic applied to the two samples $\{X_i\}$ and $\{Y_j\}$. Because the Mann–Whitney statistic is a generalized U-statistic, statistical analysis regarding the performance of diagnostic tests can be performed by exploiting the general theory for U-statistics.

The Mann–Whitney statistic estimates the probability, θ, that a randomly selected observation from the population represented by C_2 will be less than or equal to a randomly selected observation from the population represented by C_1. It can be computed as the average over a kernel, ψ, as

$$\hat{\theta} = \frac{1}{mn} \sum_{j=1}^{n} \sum_{i=1}^{m} \psi(X_i, Y_j),$$

where

$$\psi(X, Y) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X. \\ 0 & Y > X \end{cases}$$

In terms of probabilities, $\mathrm{E}(\hat{\theta}) = \theta = \Pr(Y < X) + \frac{1}{2}\Pr(X = Y)$. For continuous distributions, $\Pr(Y = X) = 0$.

Asymptotic normality and an expression for the variance of the Mann–Whitney statistic can be derived from theory developed for generalized U-statistics by Hoeffding (1948). Define

$$\begin{aligned} \xi_{10} &= \mathrm{E}[\psi(X_i, Y_j)\psi(X_i, Y_k)] - \theta^2, \quad j \neq k; \\ \xi_{01} &= \mathrm{E}[\psi(X_i, Y_j)\psi(X_k, Y_j)] - \theta^2, \quad i \neq k; \\ \xi_{11} &= \mathrm{E}[\psi(X_i, Y_j)\psi(X_i, Y_j)] - \theta^2. \end{aligned} \tag{1}$$

Then

$$\mathrm{var}(\hat{\theta}) = \frac{(n-1)\xi_{10} + (m-1)\xi_{01}}{mn} + \frac{\xi_{11}}{mn}. \tag{2}$$

Bamber (1975) provides a method of estimating the variance in the context of testing the significance of a single ROC curve. Bamber introduces a quantity B_{XXY}, which is the probability that two randomly chosen elements of the population C_1 will both be greater than or less than a randomly chosen element of C_2, minus the complementary probability that the observation from C_2 will be between the two from C_1. A similar quantity B_{YYX} is also defined and the variance of $\hat{\theta}$ is given in terms of B_{XXY} and B_{YYX}. Var($\hat{\theta}$) is then

estimated by empirically estimating B_{YYX} and B_{XXY}. Formula (2) can be shown to be equivalent to Bamber's formula (4), which derives from work of Noether (1967) and applies when X and Y are not necessarily continuous.

Hoeffding's theory extends to a vector of U-statistics. Let $\hat{\boldsymbol{\theta}} = (\hat{\theta}^1, \hat{\theta}^2, \ldots, \hat{\theta}^k)$ be a vector of statistics, representing the areas under the ROC curves derived from the readings $\{X_i^r\}$, $\{Y_j^r\}$ $(i = 1, \ldots, m; j = 1, \ldots, n; 1 \leq r \leq k)$ of k different diagnostic measures. Then, similar to (1) above, define

$$\begin{aligned}
\xi_{10}^{rs} &= \mathrm{E}[\psi(X_i^r, Y_j^r)\psi(X_i^s, Y_k^s)] - \theta^r\theta^s, \quad j \neq k; \\
\xi_{01}^{rs} &= \mathrm{E}[\psi(X_i^r, Y_j^r)\psi(X_k^s, Y_j^s)] - \theta^r\theta^s, \quad i \neq k; \\
\xi_{11}^{rs} &= \mathrm{E}[\psi(X_i^r, Y_j^r)\psi(X_i^s, Y_j^s)] - \theta^r\theta^s.
\end{aligned} \tag{3}$$

The covariance of the rth and sth statistic can then be written as

$$\mathrm{cov}(\hat{\theta}^r, \hat{\theta}^s) = \frac{(n-1)\xi_{10}^{rs} + (m-1)\xi_{01}^{rs}}{mn} + \frac{\xi_{11}^{rs}}{mn}. \tag{4}$$

Sen (1960) has provided a method of structural components to provide consistent estimates of the elements of the variance–covariance matrix of a vector of U-statistics. This approach turns out to be equivalent to jackknifing, but is conceptually simpler when dealing with U-statistics. We will exploit this methodology to compare the areas under two or more ROC curves. For the rth statistic, $\hat{\theta}^r$, the X-components and Y-components are defined, respectively, as

$$V_{10}^r(X_i) = \frac{1}{n}\sum_{j=1}^{n} \psi(X_i^r, Y_j^r) \quad (i = 1, 2, \ldots, m)$$

and

$$V_{01}^r(Y_j) = \frac{1}{m}\sum_{i=1}^{m} \psi(X_i^r, Y_j^r) \quad (j = 1, 2, \ldots, n).$$

Also define the $k \times k$ matrix $\mathbf{S}_{10}$ such that the (r, s)th element is

$$s_{10}^{r,s} = \frac{1}{m-1}\sum_{i=1}^{m} [V_{10}^r(X_i) - \hat{\theta}^r][V_{10}^s(X_i) - \hat{\theta}^s]$$

and similarly $\mathbf{S}_{01}$, which has (r, s)th element

$$s_{01}^{r,s} = \frac{1}{n-1}\sum_{j=1}^{n} [V_{01}^r(Y_j) - \hat{\theta}^r][V_{01}^s(Y_j) - \hat{\theta}^s].$$

The estimated covariance matrix for the vector of parameter estimates, $\hat{\boldsymbol{\theta}} = (\hat{\theta}^1, \hat{\theta}^2, \ldots, \hat{\theta}^k)$, is thus

$$\mathbf{S} = \frac{1}{m}\mathbf{S}_{10} + \frac{1}{n}\mathbf{S}_{01}.$$

Let g be a real-valued function of $\hat{\boldsymbol{\theta}}$ that has bounded second derivatives in a neighborhood of $\boldsymbol{\theta}$. Combining results from Sen (1960) and Arveson (1969, Theorem 16), it follows that if $\lim_{N\to\infty} m/n$ is bounded and nonzero, then $N^{1/2}[g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})]$ is asymptotically normally distributed with mean 0 and variance σ_g^2, where

$$\sigma_g^2 = \lim_{N\to\infty} N \sum_{j=1}^{k}\sum_{i=1}^{k} \frac{\partial g}{\partial \theta^i}\frac{\partial g}{\partial \theta^j}\left(\frac{1}{m}\xi_{10}^{i,j} + \frac{1}{n}\xi_{01}^{i,j}\right).$$

Further,

$$s_g^2 = N \sum_{j=1}^{k} \sum_{i=1}^{k} \frac{\partial g}{\partial \theta^i} \frac{\partial g}{\partial \theta^j} \left(\frac{1}{m} s_{10}^{i,j} + \frac{1}{n} s_{01}^{i,j} \right)$$

is a consistent estimate of σ_g^2.

When g is simply a linear function, the theory reduces considerably, because the partial derivatives are the constants that comprise the linear function. Thus, for any contrast $\mathbf{L}\boldsymbol{\theta}'$, where $\mathbf{L}$ is a row vector of coefficients,

$$\frac{\mathbf{L}\hat{\boldsymbol{\theta}}' - \mathbf{L}\boldsymbol{\theta}'}{\left[\mathbf{L} \left(\frac{1}{m} \mathbf{S}_{10} + \frac{1}{n} \mathbf{S}_{01} \right) \mathbf{L}' \right]^{1/2}}$$

has a standard normal distribution. A confidence interval for $\mathbf{L}\boldsymbol{\theta}'$ naturally follows.

By a modest generalization of these results, we can also apply any set of linear contrasts to a vector of areas under correlated ROC curves and perform a test of significance on $\mathbf{L}\hat{\boldsymbol{\theta}}'$. The test then takes the form

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\mathbf{L}' \left[\mathbf{L} \left(\frac{1}{m} \mathbf{S}_{10} + \frac{1}{n} \mathbf{S}_{01} \right) \mathbf{L}' \right]^{-1} \mathbf{L}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})', \tag{5}$$

which has a chi-square distribution with degrees of freedom equal to the rank of $\mathbf{LSL}'$. A confidence region can also be constructed.

A computer program written in the SAS language is available from the authors for computing components, covariance matrices, and contrasts. However, as indicated in the next section, the components can be computed easily by hand or by a simple computer program. The components can then be input to any program which computes sums of squares and cross-products in order to obtain the covariance matrix $\mathbf{S}$.

3. Example

When to perform surgical correction of intestinal obstruction in patients known to have ovarian carcinoma is an unresolved problem. The dilemma centers around determining those patients for whom surgery presents a benefit. Castelado et al. (1981), and other authors have proposed that patients who survive longer than 2 months postoperatively can be declared to have "benefited" from the surgery. Using this criterion, Krebs and Goplerud (1983) devised a preoperative scoring system for use as a screening test in determining a patient's risk for failing to benefit from surgery. The scoring algorithm is presented in Table 1. According to this scoring system, patients with low scores should be good candidates for surgery and those with higher scores should be considered at risk for failing to benefit from surgery.

The following example evaluates the discriminating ability of the proposed screening algorithm on 49 consecutive ovarian cancer patients undergoing correction of intestinal obstruction at Duke University Medical Center. Of the 49 patients, 12 survived more than 2 months postoperatively and could be considered surgical successes; the remaining 37 are considered failures. The Krebs–Goplerud score (*K–G*) is compared against two other preoperatively measured indices: total protein (*TP*) and albumin (*ALB*), both of which are positively associated with the patient's nutritional status. Because *ALB* is one component of *TP*, these two measures are highly correlated, with a Kendall's tau-*b* value of .65. Increasing levels of *ALB* and *TP* are associated with better nutritional status, whereas increasing levels of *K–G* are associated with poorer prognosis. Thus, to simplify computa-

Table 1
Krebs–Goplerud scoring system for prognostic parameters in ovarian carcinoma complicated by bowel obstruction

Parameter	Assigned risk score
Age (yr)	
<45	0
45–65	1
>65	2
Nutritional status (deprivation)	
None or minimal	0
Moderate	1
Severe	2
Tumor status	
No palpable intra-abdominal masses	0
Palpable intra-abdominal masses	1
Liver involvement or distant metastases	2
Ascites	
None or mild (asymptomatic, abdomen not distended)	0
Moderate (abdomen distended)	1
Severe (symptomatic, requires frequent paracentesis)	2
Previous chemotherapy	
None, or no adequate trial	0
Failed single-drug therapy	1
Failed combination-drug therapy	2
Previous radiation therapy	
None	0
Radiation therapy to pelvis	1
Radiation therapy to whole abdomen	2

tions, we transformed by subtracting *K–G* from 12, the maximum possible value, so that all indices would prognosticate in the same direction.

Figure 1 displays the empirical ROC curves for the three indices. From this figure, it appears that *K–G* offers little improvement over either *ALB* or *TP*. The estimated areas under the curves for *K–G*, *ALB*, and *TP* are .69, .72, and .65, respectively. To analyze and compare these areas, the covariance matrix for the vector of areas is needed. The method of structural components easily produces this matrix. For each of the variables of interest, (*K–G*, *ALB*, *TP*), we can denote by X^r ($r = 1, 2, 3$) the values associated with success and by Y^r ($r = 1, 2, 3$) the values associated with surgical failures. Then, $\theta^r = \Pr(Y^r < X^r) + \frac{1}{2}\Pr(Y^r = X^r)$ and we compute the components individually for each of the three variables. If the data are first sorted by the variable of interest, it is a simple matter to calculate for each X the number of Y's less than X (NYL_X) and the number of Y's equal to X ($NYEQ_X$). The component for X is then $NYL_X + \frac{1}{2}NYEQ_X$. Likewise, for each Y we calculate the number of X's greater than Y (NXG_Y) and the number of X's equal to Y ($NXEQ_Y$). The component for Y is $NXG_Y + \frac{1}{2}NXEQ_Y$.

For this example, there are 12 X's and three variables of interest, so the X-components form a 12×3 matrix, $\mathbf{V}_{10}$. The 37 Y's yield a component matrix of dimension 37×3, $\mathbf{V}_{01}$. The 3×3 matrices $\mathbf{S}_{10}$ and $\mathbf{S}_{01}$ are then computed as

$$\mathbf{S}_{10} = \frac{1}{11}(\mathbf{V}'_{10}\mathbf{V}_{10} - 12\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}})$$

and

$$\mathbf{S}_{01} = \frac{1}{36}(\mathbf{V}'_{01}\mathbf{V}_{01} - 37\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}}).$$

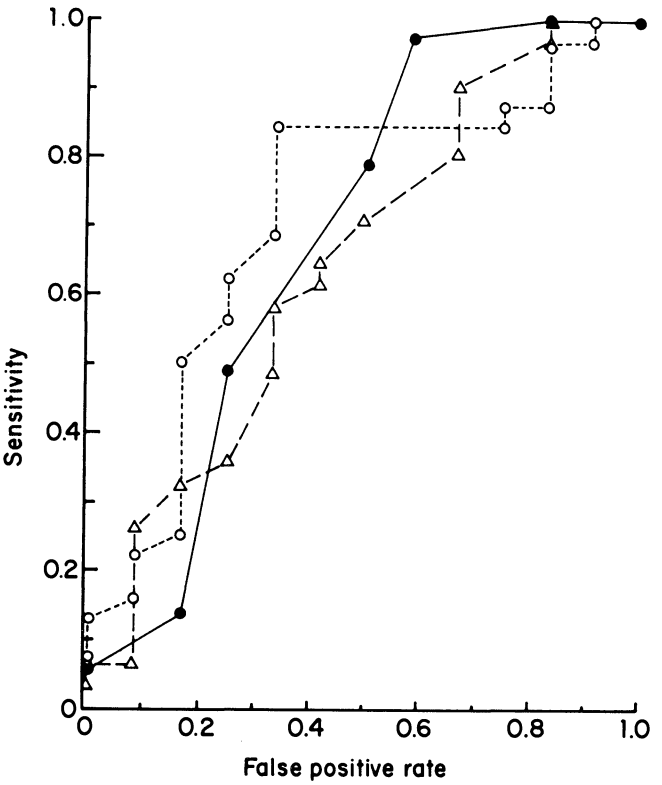

Figure 1. Receiver operating characteristic curves for Krebs–Goplerud score (●), total protein (△), and albumin (○).

It is clear that $\mathbf{S}_{10}$ and $\mathbf{S}_{01}$ are the covariance matrices of $\mathbf{V}_{10}$ and $\mathbf{V}_{01}$, respectively. They can readily be obtained from any computer program that computes covariance matrices. The covariance matrix for the vector of areas is then

$$\mathbf{S} = \frac{1}{12}\mathbf{S}_{10} + \frac{1}{37}\mathbf{S}_{01}.$$

Table 2
Estimated covariance matrix between areas under the three ROC curves

	Covariance		
	K–G score	Albumin	Total protein
K–G score	.0110	.0033	.0028
Albumin		.0086	.0076
Total protein			.0100

Table 3
Correlation coefficients of pairs of areas calculated from estimated covariance matrix (ECM) *and also from method of Hanley and McNeil* (HM)

	Correlation (ECM)	Kendall's tau-*b* Survivors	Kendall's tau-*b* Nonsurvivors	Correlation (HM)
K–G, ALB	.34	.20	.18	.17
K–G, TP	.27	−.01	.21	.10
ALB, TP	.82	.61	.66	.61

This matrix is displayed in Table 2. In Table 3, the resulting correlation coefficients are presented, along with Kendall's tau-*b* values for the group that benefited from surgery and for the remaining group, and finally the estimated correlations derived from the table in the paper by Hanley and McNeil (1983). For this set of data, our estimates tend to be larger.

Now, to compare *K–G* to the average of *ALB* and *TP*, we use the contrast $\mathbf{L} = (1, -.5, -.5)$. Evaluated at $\hat{\boldsymbol{\theta}}$, the value of the contrast is .004. The standard deviation of this estimate is

$$(\mathbf{LSL}')^{1/2} = .116.$$

A two-sided 95% confidence interval for this contrast is thus (−.223, .231), indicating negligible improvement by *K–G* over *ALB* and *TP*.

To determine whether the Krebs–Goplerud score is better than at least one of the other indices, *ALB* and *TP*, we use the contrast

$$\mathbf{L} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}.$$

Then based on (5), the χ^2 statistic with 2 degrees of freedom can be computed as 1.51 with a *P*-value of .47. Based on this sample of 49 patients, there appears to be no advantage in using the Krebs–Goplerud score over other routinely collected nutritional parameters, although power in this situation is likely to be very small because of the small sample size.

4. Discussion

ROC curves are frequently being applied to the evaluation of diagnostic or prognostic tests and indices. In order to make comparisons between two or more such indices derived from the same test units or subjects, the implicit correlation between the curves should be taken into account. This paper has presented a totally nonparametric approach to the comparison of the areas under two or more ROC curves by using the theory developed for generalized *U*-statistics. A covariance matrix can be estimated using the method of structural components and the resulting test statistic has asymptotically a chi-square distribution. The covariance matrix may also be used to construct confidence regions.

Acknowledgements

This work was supported in part by the Veterans' Administration Region 2 Health Services Research and Development Field Program.

Résumé

L'importance des méthodes d'évaluation et de comparaison de la performance des tests diagnostiques croît dans le même temps que de nouveaux tests se développent et sont lancés sur le marché. Quand un test est fondé sur une variable observée continue ou qui prend ses valeurs sur une échelle graduée, on peut faire une estimation globale de la valeur du test en utilisant la courbe caractéristique (ROC) du receveur. La courbe est construite en faisant varier la coupure utilisée pour déterminer quelles valeurs de la variable observée sont à considérer comme anormales, et ensuite en faisant la graphe des sensibilités résultantes contre les ratios correspondants faussement positifs. On doit tenir compte de la nature corrélée des données dans l'analyse statistique des différences entre courbes quand deux ou plusieurs courbes empiriques sont construites à partir de tests basés sur les mêmes individus. On présente dans ce papier une approche non paramétrique de l'analyse des aires sous des courbes ROC corrélées, en utilisant la théorie sur la statistique *U* généralisée, pour engendrer une matrice de covariance estimée.

References

Arveson, J. N. (1969). Jackknifing *U*-statistics. *Annals of Mathematical Statistics* **40,** 2076–2100.

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.

Castelado, T. W., Petrilli, E. S., Ballon, S. C., and Lagasse, L. D. (1981). Intestinal operations in patients with ovarian carcinoma. *American Journal of Obstetrics and Gynecology* **139,** 80–84.

Dorfman, D. D. and Alf, E. (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 area under two ROC curves derived from the same cases. *Radiology* **148,** 839–843.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* **19,** 293–325.

Krebs, H. B. and Goplerud, D. R. (1983). Surgical management of bowel obstruction in advanced ovarian carcinoma. *Obstetrics and Gynecology* **61,** 327–330.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine* **8,** 283–298.

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* VIII, F. Deconick (ed.), 432–445. The Hague: Martinus Nijhof.

Noether, G. E. (1967). *Elements of Nonparametric Statistics*. New York: Wiley.

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

Swets, J. A. and Pickett, R. M. (1982). *Evaluation of Diagnostic Systems*: *Methods from Signal Detection Theory*. New York: Academic Press.

Received April 1987; *revised October* 1987 *and January* 1988.