Appl. Statist. (2001)
50, *Part* 3, *pp.* 337–344

The average area under correlated receiver operating characteristic curves: a nonparametric approach based on generalized two-sample Wilcoxon statistics

Mei-Ling Ting Lee and Bernard A. Rosner

Brigham and Women's Hospital and Harvard Medical School, Boston, USA

[Received November 2000. Revised February 2001]

Summary. It is well known that, when sample observations are independent, the area under the receiver operating characteristic (ROC) curve corresponds to the Wilcoxon statistics if the area is calculated by the trapezoidal rule. Correlated ROC curves arise often in medical research and have been studied by various parametric methods. On the basis of the Mann–Whitney U-statistics for clustered data proposed by Rosner and Grove, we construct an average ROC curve and derive nonparametric methods to estimate the area under the average curve for correlated ROC curves obtained from multiple readers. For the more complicated case where, in addition to multiple readers examining results on the same set of individuals, two or more diagnostic tests are involved, we derive analytic methods to compare the areas under correlated average ROC curves for these diagnostic tests. We demonstrate our methods in an example and compare our results with those obtained by other methods. The nonparametric average ROC curve and the analytic methods that we propose are easy to explain and simple to implement.

Keywords: Generalized two-sample Wilcoxon statistics; Multiple readers; Sensitivity; Specificity; U-statistics for clustered data

1. Introduction

Correlated receiver operating characteristic (ROC) curves arise often in medical research. Franken *et al*. (1992) compared the detection accuracy of radiologists interpreting video images of neonatal examinations using plain film and digital radiography monitors with a picture archiving and communication system (PACS). The case sample consists of 100 chest or abdominal radiographs from the neonatal intensive care unit in which diagnosis was confirmed. Each observer examined each patient once using the plain film and once using digital workstation monitors with the PACS. The detection accuracy was measured by the area under the ROC curves. When ROC curves are based on two diagnostic tests performed on the same set of individuals, the correlated nature of the data must be taken into account in the analysis. For a single diagnostic test, DeLong *et al*. (1988) discussed an approach to the comparison of ROC curves that are correlated because they are from the same individuals. On the basis of a binormal model of the ROC curves, Dorfman *et al*. (1992) computed jackknife pseudovalues of areas under curves and performed significance tests using a standard linear mixed design analysis-of-variance model. Toledano and Gatsonis (1995) used a

Address for correspondence: Mei-Ling Ting Lee, Department of Medicine, Channing Laboratory, 181 Longwood Avenue, Boston, MA 02115-5804, USA.
E-mail: stmei@channing.harvard.edu

 0035–9254/01/50337

regression model and the technique of generalized estimating equations (GEEs) to account for the correlation between the observations. Obuchowski (1997) considered an estimation of the ROC curve area for clustered data but did not discuss how to construct an average ROC curve for the case where multiple readers examine results from two or more diagnostic tests performed on the same set of individuals. A review of these previous methods for ROC studies is given by Beam (1998).

In this paper, we take the nonparametric approach and provide a simple and intuitive method for meta-analysis of correlated ROC curves for the general case where multiple readers give scores using more than one diagnostic test performed on the same set of individuals. We review generalized two-sample Wilcoxon statistics. On the basis of the generalized Wilcoxon statistics, we construct an average ROC curve for multiple readers using the combined measures of sensitivity and specificity from correlated ROC curves. We present analytic methods to draw meta-analytic summaries in comparing correlated average ROC curves for these diagnostic tests. We demonstrate the methods proposed by using data collected from the study by Franken *et al.* (1992) and compare the results with those obtained by using the jackknife methods and by GEE regression models.

2. Receiver operating characteristic curves and U-statistics

Hanley and McNeil (1982) discussed the relationship between the area under an ROC curve and the two-sample Wilcoxon (Mann–Whitney's U-) statistics. They showed that, because the Wilcoxon statistic is based on pairwise comparisons of scores from diseased *versus* non-diseased subjects, it corresponds to the area under the ROC curve if the area is calculated by the trapezoidal rule. When a diagnostic test is based on a variable that is measured on a continuous or graded scale, an assessment of the overall value of the test can be made through the use of the ROC curve. The ROC curve is defined as the graph of the true positive ratio TPR against the false positive ratio FPR based on different cut points. The observed operating points for the diagnostic test are the set of points (FPR, TPR), where

$$\text{TPR} = P(\text{test positive} \,|\, \text{disease}) = \text{sensitivity}$$

and

$$\text{FPR} = P(\text{test positive} \,|\, \text{non-disease}) = 1 - \text{specificity}.$$

3. Generalized Mann–Whitney U-statistics for correlated data

Let vector $\mathbf{X}^{(i)} = (X_1^{(i)}, X_2^{(i)}, \ldots, X_{J_i}^{(i)})$ denote J_i observations obtained from non-diseased subject i, for $i = 1, \ldots, m$. Let $\mathbf{Y}^{(k)} = (Y_1^{(k)}, Y_2^{(k)}, \ldots, Y_{L_k}^{(k)})$ denote L_k observations obtained from diseased subject k, for $k = 1, \ldots, n$. The total number of subjects $N = m + n$.

Denote $\Phi(\mathbf{X}^{(i)}, \mathbf{Y}^{(k)}) = \Sigma_{j=1}^{J_i} \Sigma_{l=1}^{L_k} \mathbf{1}(X_j^{(i)} - Y_l^{(k)})$, where $\mathbf{1}(x - y) = 1$ if $x < y$, $\mathbf{1}(x - y) = \frac{1}{2}$ if $x = y$ and $\mathbf{1}(x - y) = 0$ otherwise. Obuchowski (1997) and Rosner and Grove (1999) considered the generalized Mann–Whitney statistic for clustered data. Let

$$W_c = \sum_{i=1}^{m} \sum_{k=1}^{n} \Phi(\mathbf{X}^{(i)}, \mathbf{Y}^{(k)}) = \sum_{i=1}^{m} \sum_{k=1}^{n} \left\{ \sum_{j}^{J_i} \sum_{l}^{L_k} \mathbf{1}(X_j^{(i)} - Y_l^{(k)}) \right\}$$

and $\theta^{(i,k)} = P(X_j^{(i)} < Y_l^{(k)})$ for any j, l. Denote

$$\theta = \frac{1}{mn}E(W_c) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{n} E\{\Phi(\mathbf{X}^{(i)}, \mathbf{Y}^{(k)})\} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{n} J_i L_k \theta^{(i,k)}.$$

Obuchowski (1997) gave an asymptotic variance for W_c. Rosner and Grove (1999) derived an explicit formula for the exact variance of W_c, which can be expressed in terms of some clustering parameters. Dehling and Lee (2000) have derived explicit formulae for asymptotic distributions of generalized U-statistics for clustered data.

4. Construction of an average receiver operating characteristic curve for multiple readers

Using the generalized Mann–Whitney U-statistics as reviewed in the previous section, we present in this section a method to construct an average ROC curve based on combined measures of sensitivity and specificity from correlated ROC curves obtained from multiple readers.

Assume that, for any given patient, a score will be given by each of J readers. We consider the special case where $J_i = J$ for all $i = 1, \ldots, m$ and $L_k = L$ for all $k = 1, \ldots, n$. In this section we consider the special case that $L = J$.

For any given diagnostic test, let $X_j^{(i)}$ denote the score given by the jth reader for a non-diseased subject i, where $i = 1, \ldots, m$ and $j = 1, \ldots, J$. Similarly, let $Y_l^{(k)}$ denote the score given by the lth reader for a diseased subject k, where $l = 1, \ldots, J$ and $k = 1, \ldots, n$.

If high values of score z indicate positive disease status, then we define combined sensitivity(z) as

$$\sum_{k=1}^{n}\sum_{l=1}^{J} \mathbf{1}(Y_l^{(k)} > z)/nJ$$

and combined specificity(z) as

$$\sum_{i=1}^{m}\sum_{j=1}^{J} \mathbf{1}(X_j^{(i)} < z)/mJ.$$

By reversing the inequality sign, similar definitions can be made if low values of the score indicate positive disease status.

The average empirical ROC curve can then be defined as the plot of combined sensitivity(z) *versus* 1− combined specificity(z), where z belongs to the set of cut points used in individual ROC curves. The area under this average ROC curve can be estimated by the generalized Mann–Whitney statistic as discussed in Section 2. We can estimate the standard error of the area under the average ROC curve by using either the exact variance formula given by Rosner and Grove (1999) or the asymptotic variance considered by Obuchowski (1997).

5. Comparisons of areas under correlated average receiver operating characteristic curves

In this section, we assume that, in addition to J readers, there are T different diagnostic methods. Our interest is to derive a test to compare the average ROC curves for these different diagnostic methods. Given any diagnostic test t, $t = 1, \ldots, T$, let $X_j^{(i)t}$ denote the score given by reader j, for observations taken from non-diseased subject i, where $i = 1, \ldots, m$ and $j = 1, \ldots, J$. The vector of scores given by J readers is denoted by

$\mathbf{X}^{(i)t} = (X_1^{(i)t}, \ldots, X_J^{(i)t})$. Similarly, let $Y_l^{(k)t}$ denote the score given by reader l for observations from diseased subject k, with $k = 1, \ldots, n$ and $l = 1, \ldots, J$. And the vector of scores given by J readers is denoted by $\mathbf{Y}^{(k)t} = (Y_1^{(k)t}, \ldots, Y_J^{(k)t})$. The vector of estimated generalized Wilcoxon statistics corresponding to the set of T diagnostic tests under study for comparison is denoted by $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^T)'$. Taking into account the covariance between correlated average ROC areas, we extend the methods of DeLong *et al.* (1988) and Obuchowski (1997) and derive methods for inferences on comparing areas under correlated average ROC curves.

Given any diagnostic test $t = 1, \ldots, T$, define the X-component and the Y-component of the statistic θ^t by

$$\hat{\Theta}_{(i,.)}(\mathbf{X}^{(i)t}) = \frac{1}{n}\sum_{k=1}^{n}\Phi(\mathbf{X}^{(i)t}, \mathbf{Y}^{(k)t}),$$

$$\hat{\Theta}_{(.,k)}(\mathbf{Y}^{(k)t}) = \frac{1}{m}\sum_{i=1}^{m}\Phi(\mathbf{X}^{(i)t}, \mathbf{Y}^{(k)t})$$

where $\Phi(\mathbf{X}^{(i)t}, \mathbf{Y}^{(k)t}) = \Sigma_{j=1}^{J_i}\Sigma_{l=1}^{L_k}\mathbf{1}(X_j^{(i)t} - Y_l^{(k)t})$; $\mathbf{1}(x-y) = 1$ if $x < y$, $\mathbf{1}(x-y) = \frac{1}{2}$ if $x = y$ and $\mathbf{1}(x-y) = 0$ otherwise.

Let Σ_{10} denote a $T \times T$ matrix with $\sigma_{10}^{(t_1,t_2)}$ in the (t_1, t_2) cell and Σ_{01} denote a matrix with $\sigma_{01}^{(t_1,t_2)}$ in the (t_1, t_2) cell, where

$$\sigma_{10}^{(t_1,t_2)} = \frac{1}{m-1}\sum_i\{\hat{\Theta}_{(i,.)}^{t_1}(\mathbf{X}^{(i)t_1}) - \hat{\theta}^{t_1}\}\{\hat{\Theta}_{(i,.)}^{t_2}(\mathbf{X}^{(i)t_2}) - \hat{\theta}^{t_2}\},$$

$$\sigma_{01}^{(t_1,t_2)} = \frac{1}{n-1}\sum_k\{\hat{\Theta}_{(.,k)}^{t_1}(\mathbf{Y}^{(k)t_1}) - \hat{\theta}^{t_1}\}\{\hat{\Theta}_{(.,k)}^{t_2}(\mathbf{Y}^{(k)t_2}) - \hat{\theta}^{t_2}\}.$$

Here

$$\hat{\theta}^{t_1} = \frac{1}{m}\sum_{i=1}^{m}\hat{\Theta}_{(i,.)}(\mathbf{X}^{(i)t}) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{n}\Phi(\mathbf{X}^{(i)t_1}, \mathbf{Y}^{(k)t_1})$$

and

$$\hat{\theta}^{t_2} = \frac{1}{n}\sum_{k=1}^{n}\hat{\Theta}_{(.,k)}(\mathbf{Y}^{(k)t}) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{n}\Phi(\mathbf{X}^{(i)t_2}, \mathbf{Y}^{(k)t_2}).$$

Then, the estimated asymptotic covariance matrix for the vector $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^T)'$ is

$$\Sigma = \frac{1}{m}\Sigma_{10} + \frac{1}{n}\Sigma_{01}.$$

Let B denote a vector of contrasts; then the statistic

$$B\hat{\boldsymbol{\theta}}' \Big/ \left\{B\left(\frac{1}{m}\Sigma_{10} + \frac{1}{n}\Sigma_{01}\right)B'\right\}^{1/2}$$

is distributed as a standard normal distribution. Using these results, we can apply any set of linear contrasts to a vector of areas under correlated ROC curves and perform a test of significance on $B\boldsymbol{\theta}'$.

6. Application

We demonstrate the methods proposed by using data obtained from the study conducted by

Franken *et al*. (1992) comparing the diagnostic accuracy of the PACS and plain film for evaluating clinical neonatal radiographs. According to the study design, four radiologists with considerable experience in interpreting neonatal examinations analysed the radiographs. During two initial reading sessions, half the studies were viewed on digital radiograph monitors and half on plain film. Observers indicated whether each patient had normal or abnormal findings and their degree of confidence in this judgment. 6 weeks later, observers viewed cases on the alternative presentation system. The images were presented in random order. Each of the four participating readers interpreted both the PACS and the plain film version of each radiographic study. Observers were asked to determine whether the findings

Table 1. Summary results of 100 radiographs examined by four readers using the PACS and plain film

Status	*Numbers with the following scores:*					*Total*
	1	*2*	*3*	*4*	*5*	
(*a*) *PACS*						
Reader 1						
Negative	1	5	4	17	6	33
Positive	38	15	5	5	4	67
Total	39	20	9	22	10	100
Reader 2						
Negative	0	3	7	14	9	33
Positive	0	49	11	4	3	67
Total	0	52	18	18	12	100
Reader 3						
Negative	1	2	5	18	7	33
Positive	36	10	8	12	1	67
Total	37	12	13	30	8	100
Reader 4						
Negative	2	4	2	4	21	33
Positive	35	13	2	6	11	67
Total	37	17	4	10	32	100
(*b*) *Plain film*						
Reader 1						
Negative	1	3	7	14	8	33
Positive	34	14	9	7	3	67
Total	35	17	16	21	11	100
Reader 2						
Negative	0	2	10	2	19	33
Positive	0	41	16	4	6	67
Total	0	43	26	6	25	100
Reader 3						
Negative	0	5	4	21	3	33
Positive	31	16	9	8	3	67
Total	31	21	13	29	6	100
Reader 4						
Negative	1	4	4	3	21	33
Positive	36	10	4	4	13	67
Total	37	14	8	7	34	100

of each case were normal or abnormal and to indicate the degree of confidence in this decision by using a five-point confidence scale: 1, definitely abnormal; 2, probably abnormal; 3, possibly abnormal; 4, probably normal; 5, definitely normal. The sample included 100 cases, of which 33 were normal (true negatives) and 67 were abnormal (true positives). Table 1 lists the scores for the PACS and plain films given by each of the four readers. It is clear that scores given by reader 2 are consistently different from those of the other three readers.

For the PACS and plain film separately, we use the generalized Wilcoxon statistics to compute average areas under the ROC curves combining four readers. We estimate the exact standard error SE of the two average areas by using the method of Rosner and Grove (1999) and the asymptotic standard error ASE by using the method of Obuchowski (1997). As a result, the area under the average ROC curve for the PACS is $\hat{\theta}_1 = 0.840$ with $\mathrm{SE}(\hat{\theta}_1) = 0.050$ and $\mathrm{ASE}(\hat{\theta}_1) = 0.030$; the corresponding area under the average ROC curve for plain film is $\hat{\theta}_2 = 0.831$ with $\mathrm{SE}(\hat{\theta}_2) = 0.050$ and $\mathrm{ASE}(\hat{\theta}_2) = 0.030$. Table 2 displays the observed operating points from four readers for both the PACS and the plain film system. Also listed in Table 2 are comparisons of results obtained by using the nonparametric methods proposed *versus* those of the jackknife methods by Dorfman *et al.* (1992) and the GEE method by Toledano and Gatsonis (1995). Table 2 includes estimated areas under the ROC curve for each reader and the area under the average ROC curve. Corresponding ROC curves for each of the four readers and the average ROC curve are displayed in Fig. 1.

In this paper the areas under the curves are nonparametric estimates calculated by the trapezoidal rule, whereas for parametric models a smooth curve was fitted to the discrete set of operating points. Hence, the estimated areas for individual ROC curves are relatively smaller than those obtained by Dorfman *et al.* (1992) and Toledano and Gatsonis (1995). As expected, there were some differences between the exact and asymptotic standard errors in the example considered, with the exact standard errors being larger.

Using methods discussed in Section 5, we test the hypothesis that, after combining the

Table 2. Operating points (FPR, TPR) and areas under the curves (with standard errors in parentheses) of ROC curves for each reader and for the average curves from four readers†

Reader	*Operating points (FPR, TPR)*				*Estimated areas under the curves for the following methods:*		
	Point 1	*Point 2*	*Point 3*	*Point 4*	*Nonparametric*	*Jackknife*	*GEE*
(*a*) *PACS*							
1	(0.03, 0.57)	(0.18, 0.79)	(0.30, 0.87)	(0.82, 0.94)	0.853 (0.059)	0.860 (0.040)	0.86 (0.04)
2	(0.00, 0.00)	(0.09, 0.73)	(0.30, 0.90)	(0.73, 0.96)	0.865 (0.057)	0.895 (0.034)	0.89 (0.03)
3	(0.03, 0.54)	(0.09, 0.69)	(0.24, 0.81)	(0.79, 0.99)	0.857 (0.059)	0.876 (0.036)	0.84 (0.04)
4	(0.06, 0.52)	(0.18, 0.72)	(0.24, 0.75)	(0.36, 0.84)	0.815 (0.060)	0.841 (0.044)	0.84 (0.03)
Average	(0.03, 0.41)	(0.14, 0.73)	(0.27, 0.83)	(0.68, 0.93)	0.840 (0.050) exact 0.840 (0.030) asymptotic	0.868 (0.025)	0.87 (0.03)
(*b*) *Plain film*							
1	(0.03, 0.51)	(0.12, 0.72)	(0.33, 0.85)	(0.76, 0.96)	0.850 (0.060)	0.860 (0.039)	0.86 (0.04)
2	(0.00, 0.00)	(0.06, 0.61)	(0.36, 0.85)	(0.42, 0.91)	0.844 (0.058)	0.872 (0.037)	0.86 (0.04)
3	(0.00, 0.46)	(0.15, 0.70)	(0.27, 0.84)	(0.91, 0.96)	0.840 (0.060)	0.846 (0.04)	0.85 (0.04)
4	(0.03, 0.54)	(0.15, 0.69)	(0.27, 0.75)	(0.36, 0.81)	0.814 (0.059)	0.823 (0.047)	0.85 (0.03)
Average	(0.02, 0.38)	(0.12, 0.68)	(0.31, 0.82)	(0.61, 0.91)	0.831 (0.050) exact 0.831 (0.030) asymptotic	0.850 (0.026)	0.85 (0.03)

†Comparison results (estimated differences of PACS − plain film): nonparametric method, 0.009 ± 0.018; jackknife method, 0.018 ± 0.020; GEE method, 0.022 ± 0.020.

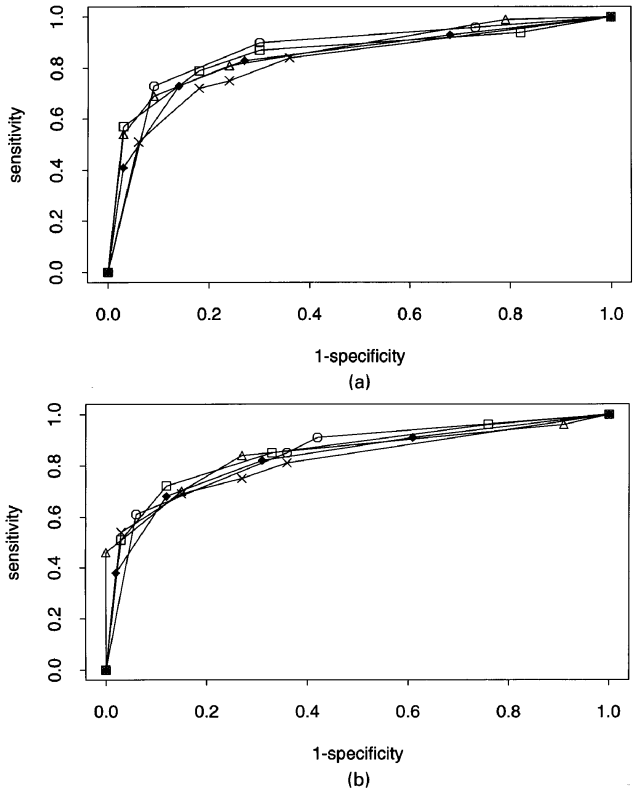

Fig. 1. ROC curves for the four readers (□, reader 1; ○, reader 2; △, reader 3; X, reader 4; ◆, combined): (a) PACS; (b) plain film

results from all four readers, there is no difference between the average areas under the ROC curves for the PACS and for plain film. To test whether there is a difference between the overall results of plain film *versus* the PACS, we consider the hypothesis that H_0: $\theta^1 = \theta^2$. Using the contrast $B = (1, -1)$, we obtain a test statistic value $\chi^2 = 0.2427$, which is not significant for a χ^2-distribution with 1 degree of freedom. Thus, we conclude that there is no significant difference between the two methods. This result is consistent with results obtained by Toledano and Gatsonis (1995), who considered a GEE regression model to account for correlations.

7. Concluding remarks

One advantage of this method is the avoidance of the binormal assumption that is required in some other approaches discussed in the literature. Another advantage is the ability to obtain exact *versus* asymptotic standard errors of the area under the curve estimates. This paper extends the nonparametric methods considered by DeLong *et al.* (1988) and by Obuchowski (1997). The nonparametric average ROC curve and the analytic methods that we propose are easy to explain and simple to implement. Because of the nonparametric nature of our methods, the average ROC curve can be used as an exploratory tool to draw meta-analytic summaries in comparing diagnostic tests when multiple readers are involved.

Acknowledgement

This research is supported in part by National Institutes of Health grants HL40619-09 and EY12269-02.

References

Beam, C. A. (1998) Analysis of clustered data in receiver operating characteristic studies. *Statist. Meth. Med. Res.*, **7**, 324–336.

Dehling, H. G. and Lee, M.-L. T. (2000) Asymptotic distribution of generalized U-statistics for clustered data. *Technical Report*. Channing Laboratory, Boston.

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., Berbaum, K. S. and Metz, C. E. (1992) Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jacknife method. *Invest. Radiol.*, **27**, 723–731.

Franken, Jr, E. A., Berbaum, K. S., Marley, S. M., Smith, W. L., Sato, Y., Kao, S. C. S. and Milam, S. G. (1992) Evaluation of a digital workstation for interpreting neonatal examinations: a receiver operating characteristic study. *Invest. Radiol.*, **27**, 732–737.

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

Obuchowski, N. A. (1997) Nonparametric analysis of clustered ROC curve data. *Biometrics* **53**, 567–578.

Rosner, B. and Grove, D. (1999) Use of the Mann-Whitney U-statistics for clustered data. *Statist. Med.*, **18**, 1387–1400.

Toledano, A. and Gatsonis, C. (1995) Regression analysis of correlated ROC data. *Acad. Radiol.*, **2**, S30–S36.