Journal of Applied Mathematics and Physics, 2015, 3, 828-833
Published Online July 2015 in SciRes. http://www.scirp.org/journal/jamp
http://dx.doi.org/10.4236/jamp.2015.37102

Optimal Weights in Nonparametric Analysis of Clustered ROC Curve Data

Yougui Wu

Department of Epidemiology and Biostatistics, College of Public Health, University of South Florida, Tampa, Florida, USA
Email: ywu@health.usf.edu

Received 8 May 2015; accepted 23 June 2015; published 30 June 2015

Abstract

In diagnostic trials, clustered data are obtained when several subunits of the same patient are observed. Within-cluster correlations need to be taken into account when analyzing such clustered data. A nonparametric method has been proposed by Obuchowski (1997) to estimate the Receiver Operating Characteristic curve area (AUC) for such clustered data. However, Obuchowski's estimator gives equal weight to all pairwise rankings within and between cluster. In this paper, we modify Obuchowski's estimate by allowing weights for the pairwise rankings vary across clusters. We consider the optimal weights for estimating one AUC as well as two AUCs' difference. Our results in this paper show that the optimal weights depends on not only the within-patient correlation but also the proportion of patients that have both unaffected and affected units. More importantly, we show that the loss of efficiency using equal weight instead of our optimal weights can be severe when there is a large within-cluster correlation and the proportion of patients that have both unaffected and affected units is small.

Keywords

Diagnostic Test, Optimal Weight, Asymptotic Relative Efficiency, Receiver Operating Characteristic Curve, Area under a ROC Curve

1. Introduction

In diagnostic trials, clustered data are obtained when several subunits of the same patient are observed. For example, in a study by Masaryk *et al*. (1991) [2], two radiologists evaluated 65 carotid arteries (left and right) in 36 patients using three-dimensional Magnetic Resonance Angiography(MRA), a potential screeening tool for athe- rosclerosis of the carorid arteries. These patients also underwent intra-arterial digital subtraction angiography (DSA), which is considered the gold standard for characterizing the degree of stenosis. The goals of the study were to evaluate the performance of MRA according to each reader, and to compare the performance for the two radiologists.

In the above example, each patient(cluster) contributes a number of unaffected and affected units. Correlation exists for outcomes between two unaffected units, between two affected units, and between an unaffected and an affected unit from the same cluster, and between the outcomes of the two diagnostic tests from the same cluster.

How to cite this paper: Wu, Y.G. (2015) Optimal Weights in Nonparametric Analysis of Clustered ROC Curve Data. *Journal of Applied Mathematics and Physics*, **3**, 828-833. http://dx.doi.org/10.4236/jamp.2015.37102

All these correlations need to be taken into account when analyzing such clustered data.

An ROC curve is a plot of a diagnostic test's sensitivity versus 1-specificity. The curve is constructed by changing the cutpoint that defines a positive diagnostic test result. The area under the ROC curve (AUC) summarizes the test's overall diagnostic ability and is typically used as a global measure of the accuracy of the diagnostic test.

In the clustered data case, Obuchowski (1997) [1] proposed a nonparametric AUC estimator, and derived an asymptotic variance estimate for the AUC estimator, taking into account of within-cluster correlations. However, Obuchowski's AUC estimator gives equal weight to all pairwise rankings within and between clusters. Clusters can be different in terms of cluster size, the number of unaffected units, and the number of affected units. In the presence of various within-cluster correlations, these differences would affect the contribution of a cluster to the overall variance of the AUC estimator and hence weights should vary across clusters.

In this paper, we modify Obuchowski's estimator by allowing the weight assigned to each pairwise ranking to vary across clusters, and derive the optimal weights that minimize the variance of the AUC estimator. Our results in this paper show that the optimal weights depends not only on the within-cluster correlation but also the proportion of clusters that have both unaffected and affected units. More importantly, we show that the gain of efficiency in comparison with two simple weighting schemes can be doubled when there is a large within-cluster correlation and the proportion of clusters that have both unaffected and affected units is small.

The rest of this paper is organized as follows. In Section 2, the optimal weights for one AUC are derived and the estimators of the optimal weights are discussed. The relative asymptotic efficiencies in comparing our optimal estimator with two simple weighting schemes are studied. A data example is presented in Section 3 and conclusions are provided in Section 4.

2. Optimal Weights for Estimating One Auc

2.1. Optimal Weights Derivation

Assume that there are n clusters, of which n_{10} clusters contain only unaffected units, n_{11} clusters contain both unaffected and affected units, and n_{01} clusters contain only affected units. The total number of clusters with at least one unaffected unit is given by $n_{1+} = n_{10} + n_{11}$, and the total number of clusters with at least one affected unit is given by $n_{+1} = n_{01} + n_{11}$. Without loss of generality, we assume that clusters $1, \cdots, n_{10}$ contain only unaffected units, clusters $n_{10}+1, \cdots, n_{1+}$ contain both unaffected and affected units, and clusters $n_{1+}+1, \cdots, n$ contain only affected units. Let X_{jk} denote the diagnostic test result of the kth unaffected unit in the jth cluster $(k = 1, 2, \cdot, r_j), (j = 1, 2, \cdots, n_{1+})$. Similarly, let Y_{jk} denote the diagnostic test result of the kth affected unit in the jth cluster $(k = 1, 2, \cdot, s_j), (j = n_{10}+1, \cdots, n)$.

Let $F(t)$ and $G(t)$ be the distribution functions of X_{jk} and Y_{jk}, respectively. Assume that if the value of X_{jk} or Y_{jk} exceeds a predetermined cut-off point c the diagnostic test will be considered positive. Then the area under the ROC curve of the diagnostic test is $\theta = \int_0^\infty F(t)dG(t)$. Obuchowski (1997) [1] proposed a non-parametric estimate for θ, given by

$$\hat{\theta} = \frac{1}{RS}\sum_{j=1}^{n_{1+}}\sum_{j'=n_{10}+1}^{n}\sum_{k=1}^{r_j}\sum_{k'=1}^{s_{j'}} I(X_{jk} \le Y_{j'k'}), \tag{1}$$

where $R = \sum_{j=1}^{n_{1+}} r_j$ and $S = \sum_{j=n_{10}+1}^{n} s_j$. This estimate gives equal weight to all pairwise ranking.

Note that $F(t)$ can be estimated by

$$\hat{F}(t) = \sum_{j=1}^{n_{1+}} w_{1j}\left\{\frac{1}{r_j}\sum_{k=1}^{r_j} I(X_{jk} \le t)\right\}, \tag{2}$$

where $(w_{1j}, j = 1, 2, \cdots, n_{1+})$ is a set of weights assigned to the clusters with at least one unaffected unit satisfying $w_{1j} > 0, j = 1, \cdots, n_{1+}$ and $\sum_{j=1}^{n_{1+}} w_{1j} = 1$. Similarly, $G(t)$ can be estimated by

$$\hat{G}(t) = \sum_{j=n_{10}+1}^{n} w_{2j}\left\{\frac{1}{s_j}\sum_{k=1}^{s_j} I(Y_{jk} \le t)\right\}, \tag{3}$$

where $(w_{2j}, j = n_{10}+1, \cdots, n)$ is a set of weights assigned to the clusters with at least one affected unit satisfying $w_{2j} > 0, j = n_{10}+1, \cdots, n$ and $\sum_{j=n_{10}+1}^{n} w_{2j} = 1$. Similar to Emir *et al*. (2000) [3], two simple weighting schemes can be considered: (1) assigning equal weights to observations, *i.e.*, $w_{1j} = r_j / \sum_{j'=1}^{n_{1+}} r_{j'}, w_{2j} = s_j / \sum_{j'=n_{10}+1}^{n} s_{j'}$, when within-cluster correlation is low, and (2) assigning equal weights to clusters, *i.e.*, $w_{1j} = 1/n_{1+}, w_{2j} = 1/n_{+1}$, when within-cluster correlation is high.

We propose to estimate θ by

$$\hat{\theta} = \int_0^{\infty} \hat{F}(t) d\hat{G}(t) = \sum_{j=1}^{n_{1+}} \sum_{j'=n_{10}+1}^{n} \frac{w_{1j} w_{2j'}}{r_j s_{j'}} \sum_{k=1}^{r_j} \sum_{k'=1}^{s_{j'}} I(X_{jk} \le Y_{j'k'}). \tag{4}$$

Notice that when $w_{1j} = r_j / \sum_{j'=1}^{n_{1+}} r_{j'}$ and $w_{2j} = s_j / \sum_{j'=1}^{n_{1+}} s_{j'}$, our estimator is the same as that in Obuchowski (1997) [1].

To derive our optimal weight, we utilize the following result which can be found in the Appendix of Emir, *et al.* (2000) [3]:

$$\hat{\theta} - \theta = \sum_{j=1}^{n} (\epsilon_j + \xi_j) + o(n^{-1/2}), \tag{5}$$

where

$$\xi_j = \frac{\Delta_{0j} w_{1j}}{r_j} \sum_{k=1}^{r_j} \int_0^{\infty} \left\{ I(X_{jk} \le t) - F(t) \right\} dG(t),$$

$$\epsilon_j = \frac{\Delta_{1j} w_{2j}}{s_j} \sum_{k=1}^{s_j} \int_0^{\infty} F(t) d\left\{ I(Y_{jk} \le t) - G(t) \right\},$$

and $\Delta_{0j} = 1$ if the *j*th cluster contains at least one unaffected unit and =0 otherwise and $\Delta_{1j} = 1$ if the *j*th cluster contains at least one affected unit and =0 otherwise. Hence, the variance of $\hat{\theta}$ is approximately

$$Var(\hat{\theta}) = Var\{\sum_{j=1}^{n} (\epsilon_j + \xi_j)\}. \tag{6}$$

Note that

$$\xi_j = \Delta_{0j} w_{1j} \left\{ 1 - \theta - \frac{1}{r_j} \sum_{k=1}^{r_j} G(X_{jk}) \right\},$$

and

$$\epsilon_j = \frac{\Delta_{1j} w_{2j}}{s_j} \sum_{k=1}^{s_j} \left\{ F(Y_{jk}) - \theta \right\}.$$

Defining the transformation

$$U_{jk} = G(X_{jk}), V_{jk} = F(Y_{jk}), \tag{7}$$

we can express the variance of $\hat{\theta}$ in (6) in terms of U_{jk} and V_{jk} as

$$\begin{aligned} Var(\hat{\theta}) &= \sum_{j=1}^{n_{1+}} Var\left\{ \frac{1}{r_j} \sum_{k=1}^{r_j} U_{jk} \right\} w_{1j}^2 + \sum_{j=n_{10}+1}^{n} Var\left(\left\{ \frac{1}{s_j} \sum_{k=1}^{s_j} V_{jk} \right\} w_{2j}^2 \right. \\ &-2 \sum_{j=n_{10}+1}^{n_{1+}} w_{1j} w_{2j} Cov\left\{ \frac{1}{r_j} \sum_{k=1}^{r_j} U_{jk}, \frac{1}{s_j} \sum_{k=1}^{s_j} V_{jk} \right\} \\ &= \sum_{j=1}^{n_{1+}} a_j w_{1j}^2 - 2 \sum_{j=n_{10}+1}^{n_{1+}} b_j w_{1j} w_{2j} + \sum_{j=n_{10}+1}^{n} c_j w_{2j}^2, \end{aligned} \tag{8}$$

where

$$a_j = Var\left\{\frac{1}{r_j}\sum_{k=1}^{r_j} U_{jk}\right\} = \frac{\sum_{k,k'=1}^{r_j}\sigma_{kk'}^{uu}}{r_j^2},$$

$$b_j = Cov\left\{\frac{1}{r_j}\sum_{k=1}^{r_j} U_{jk}, \frac{1}{s_j}\sum_{k=1}^{s_j} V_{jk}\right\} = \frac{\sum_{k=1k'=1}^{r_j s_j}\sigma_{kk'}^{uv}}{r_j s_j},$$

$$c_j = Var\left\{\frac{1}{s_j}\sum_{k=1}^{s_j} V_{jk}\right\} = \frac{\sum_{k,k'=1}^{s_j}\sigma_{kk'}^{vv}}{s_j^2},$$

$$\sigma_{kk'}^{uu} = Cov(U_{jk}, U_{jk'}), k \neq k', \sigma_{kk}^{uu} = Var(U_{jk}) = \sigma_u^2,$$

$$\sigma_{kk'}^{uv} = Cov(U_{jk}, V_{jk'}),$$

and

$$\sigma_{kk'}^{vv} = Cov(V_{jk}, V_{jk'}), k \neq k', \sigma_{kk}^{vv} = Var(V_{jk}) = \sigma_v^2.$$

The optimal weights can be obtained by minimizing (8) with respect to $w_{1j}, j = 1,\cdots,n_{1+}$ and $w_{2j}, j = n_{10}+1,\cdots,n$ with constraints $w_{1j} > 0, j = 1,\cdots,n_{+1}, w_{2j} > 0, j = n_{10}+1,\cdots,n$, $\sum_{j=1}^{n_{1+}} w_{1j} = 1$, and $\sum_{j=n_{10}+1}^{n} w_{2j} = 1$. Applying Langrage Multipler Method, we have

$$w_{1j} = \begin{cases} a_j^{-1}\tilde{e}_1^t\left(H + \sum_{j=n_{10}+1}^{n_{1+}} G_j^{-1}\right)^{-1}\tilde{1} & j = 1,\cdots,n_{10} \\ \tilde{e}_1^t G_j^{-1}\left(H + \sum_{j=n_{10}+1}^{n_{1+}} G_j^{-1}\right)^{-1}\tilde{1} & j = n_{10}+1,\cdots,n_{1+} \end{cases} \quad (9)$$

and

$$w_{2j} = \begin{cases} \tilde{e}_2^t G_j^{-1}\left(H + \sum_{j=n_{10}+1}^{n_{1+}} G_j^{-1}\right)^{-1}\tilde{1} & j = n_{10}+1,\cdots,n_{1+} \\ c_j^{-1}\tilde{e}_2^t\left(H + \sum_{j=n_{10}+1}^{n_{1+}} G_j^{-1}\right)^{-1}\tilde{1} & j = n_{1+}+1,,\cdots,n \end{cases} \quad (10)$$

where $\tilde{e}_1 = (1,0)^t, \tilde{e}_2 = (0,1)^t, \tilde{1} = (1,1)^t$,

$$H = \begin{pmatrix} \sum_{j=1}^{n_{10}}\frac{1}{a_j} & 0 \\ 0 & \sum_{j=n_{1+}+1}^{n}\frac{1}{c_j} \end{pmatrix},$$

and

$$G_j = \begin{pmatrix} a_j & -b_j \\ -b_j & c_j \end{pmatrix}.$$

2.2. Asymptotic Variance Comparison

Let $\hat{\delta}_{op}$ be the estimated optimal weight, $\hat{\delta}_1$ be the estimator of δ using simple weighting Scheme 1: $w_{1j}^* = r_j / \sum_{j'=1}^{n_{1+}} r_{j'}, w_{2j} = s_j / \sum_{j'=n_{10}+1}^{n} s_{j'}$, and $\hat{\delta}_2$ be the estimator of δ using simple weighting Scheme 2: $w_j^* = 1/n_{1+}, w_{2j} = 1/n_{+1}$.

Along the same line of the proofs for (??), (??) and (??), we can show that $\sqrt{n}(\hat{\delta}_{op} - \delta)$ is approximately normal $N(0, \sigma_{op}^{2*})$, and $\sqrt{n}(\hat{\delta}_i - \delta)$ is approximately normal $N(0, \sigma_i^{2*})$, $i = 1, 2$, with

$$\sigma_{op}^{*2} = d_1\left[\tilde{e}_1^t (D^* + \Sigma^*)^{-1}\tilde{1}\right]^2 + \tilde{1}^t (D^* + \Sigma^*)^{-1}\Sigma^*(D^* + \Sigma^*)^{-1} + d_2\left[\tilde{e}_2^t (D^* + \Sigma^*)^{-1}\tilde{1}\right]^2, \tag{11}$$

$$\sigma_1^{*2} = \frac{\tau_{1+}}{(Er)^2} E\sum_{k,k'=1}^{r} \sigma_{kk'}^{u^*u^*} - 2\frac{\tau_{11}}{\tau_{1+}\tau_{+1}}\frac{E(rs)}{ErEs} E\sum_{k}^{r}\sum_{k'}^{s}\sigma_{kk'}^{u^*v^*} + \frac{\tau_{+1}}{(Es)^2} E\sum_{k,k'=1}^{s}\sigma_{kk'}^{v^*v^*}, \tag{12}$$

and

$$\sigma_2^{*2} = \frac{1}{\tau_{1+}} E\frac{\sum_{k,k'=1}^{r}\sigma_{kk'}^{u^*u^*}}{r^2} - 2\frac{\tau_{11}}{\tau_{1+}\tau_{+1}} E\frac{\sum_{k}^{r}\sum_{k'}^{s}\sigma_{kk'}^{u^*v^*}}{rs} + \frac{1}{\tau_{+1}} E\frac{\sum_{k,k'=1}^{s}\sigma_{kk'}^{v^*v^*}}{s^2}, \tag{13}$$

where

$$D^* = \begin{pmatrix} d_1^* & 0 \\ 0 & d_2^* \end{pmatrix}, \Sigma^* = \begin{pmatrix} \sigma_{11}^* & \sigma_{12}^* \\ \sigma_{21}^* & \sigma_{22}^* \end{pmatrix},$$

$$d_1^* = \lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n_{10}}\frac{1}{a_j} = \tau_{10} E\frac{r^2}{\sum_{k,k'=1}^{r}\sigma_{kk'}^{u^*u^*}},$$

$$d_2^* = \lim_{n\to\infty}\frac{1}{n}\sum_{j=n_{1+}+1}^{n}\frac{1}{c_j} = \tau_{01} E\frac{r^2}{\sum_{k,k'=1}^{r}\sigma_{kk'}^{v^*v^*}},$$

$$\sigma_{11}^* = \lim_{n\to\infty}\frac{1}{n}\sum_{j=n_{01}+1}^{n_{+1}}\frac{c_j^*}{a_j^* c_j^* - b_j^{*2}} = \tau_{11} E\frac{r^2\sum_{k,k'=1}^{s}\sigma_{kk'}^{v^*v^*}}{\sum_{k,k'=1}^{r}\sigma_{kk'}^{u^*u^*}\sum_{k,k'=1}^{s}\sigma_{kk'}^{v^*v^*} - \left(\sum_{k=1k'=1}^{r}\sum^{s}\sigma_{kk'}^{u^*v^*}\right)^2},$$

$$\sigma_{12}^* = \sigma_{21}^* = \lim_{n\to\infty}\frac{1}{n}\sum_{j=n_{10}+1}^{n_{+1}}\frac{b_j^*}{a_j^* c_j^* - b_j^{*2}} = \tau_{11} E\frac{rs\sum_{k=1k'=1}^{r}\sum^{s}\sigma_{kk'}^{u^*v^*}}{\sum_{k,k'=1}^{r}\sigma_{kk'}^{u^*u^*}\sum_{k,k'=1}^{s}\sigma_{kk'}^{v^*v^*} - \left(\sum_{k=1k'=1}^{r}\sum^{s}\sigma_{kk'}^{u^*v^*}\right)^2},$$

and

$$\sigma_{22}^* = \lim_{n\to\infty}\frac{1}{n}\sum_{j=n_{01}+1}^{n_{1+}}\frac{a_j^*}{a_j^* c_j^* - b_j^{*2}} = \tau_{11} E\frac{s^2\sum_{k,k'=1}^{r}\sigma_{kk'}^{u^*u^*}}{\sum_{k,k'=1}^{r}\sigma_{kk'}^{uu}\sum_{k,k'=1}^{s}\sigma_{kk'}^{v^*v^*} - \left(\sum_{k=1k'=1}^{r}\sum^{s}\sigma_{kk'}^{u^*v^*}\right)^2}.$$

Let $\mathrm{RE}_1^* = \sigma_{op}^{*2}/\sigma_1^{*2}$ be the asymptotic relative efficiency for comparing $\hat{\delta}_1$ with $\hat{\delta}_{op}$, and $\mathrm{RE}_2^* = \sigma_{op}^{*2}/\sigma_2^{*2}$ be the asymptotic relative efficiency for comparing $\hat{\delta}_2$ with $\hat{\delta}_{op}$. Similar to the case of a single AUC, for the special case where $\sigma_{u^*}^2 = \sigma_{v^*}^2$, and Corr $(U_{jk}^*, U_{jk'}^*) = \rho_{00}^*$, $\mathrm{Corr}(U_{jk}^*, V_j^*) = \rho_{01}^*$, we have that both RE_1^* and RE_2^* increases dramatically as ρ_{01}^* increases and ψ decreases, and increases slowly as ρ_{00}^* decreases (**Figure 1**).

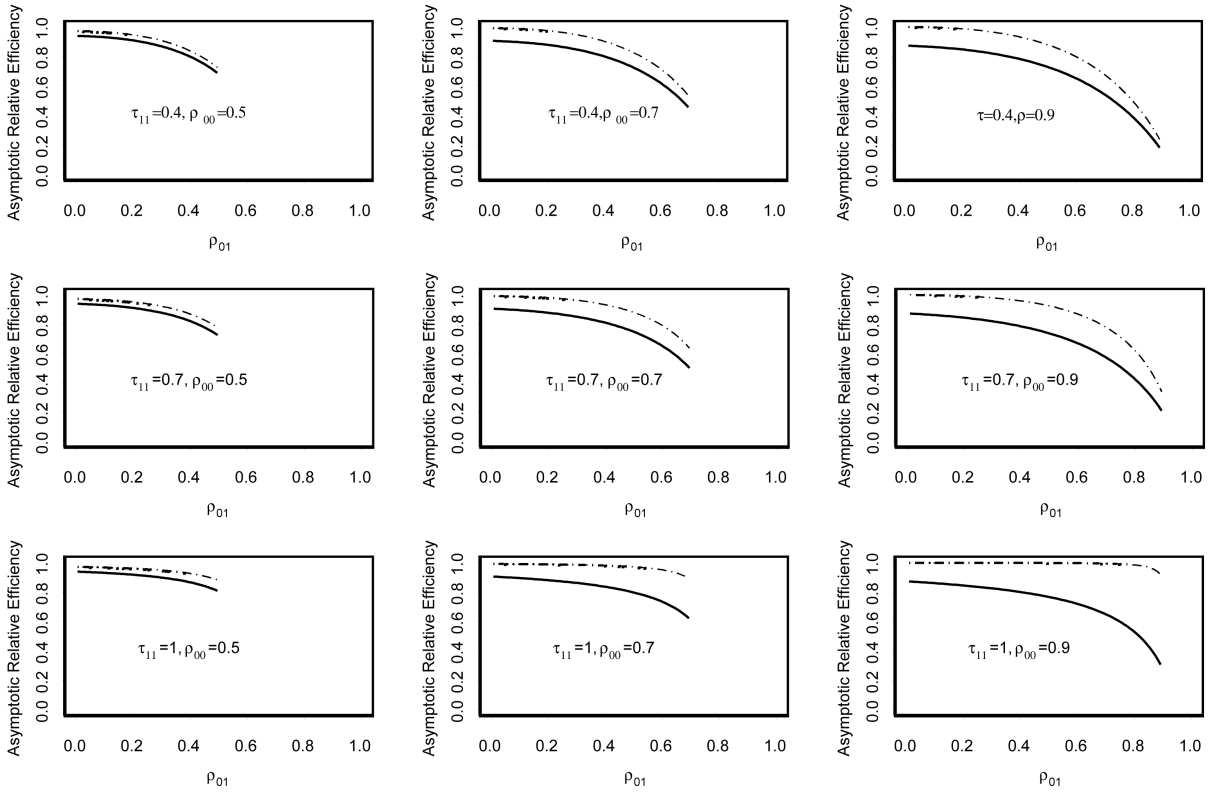

Figure 1. The effect of ρ_{01}, ρ_{00} and τ_{11} on the asymptotic relative efficiencies, RE_1 (solid line) and RE_2 (broken line).

3. Conclusions

We have proposed an optimal nonparametric estimator for one AUC, which modifies Obuchowski's estimate by allowing different weights for the pairwise rankings within and between cluster. Optimal weights for one AUC has been derived by minimizing the variance of the estimate of one AUC(two AUCs' difference). Asymptotic performance of the AUC estimate using our optimal weights has been studied in contrast with the two weighting schemes.

We have shown that when there is a moderate within-cluster unaffected-affected units correlation and the proportion of clusters that contain both unaffected and affected units is small, using either of the two weighting schemes, corresponding to Obuchowski's estimator or the estimator with equal cluster weights, can lead to dramatic efficiency loss. For this situation, the optimal weights are recommended.

References

[1] 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. http://dx.doi.org/10.1148/radiology.179.3.2027995

[2] Obuchowski, N.A. (1997) Nonparametric Analysis of Clustered ROC Curve Data. *Biometrics*, **53**, 567-578. http://dx.doi.org/10.2307/2533958

[3] Emir, B., Wieand, S., Jung, S. and Ying, Z. (2000) Comparison of Diagnostic Markers with Repeated Measurements: A Non-Parametric ROC Curve Approach. *Statistics in Medicine*, **19**, 511-523. http://dx.doi.org/10.1002/(SICI)1097-0258(20000229)19:4<511::AID-SIM353>3.0.CO;2-3