Published in final edited form as:

Lifetime Data Anal. 2013 April ; 19(2): 242–256. doi:10.1007/s10985-012-9240-6.

Assessing Discrimination of Risk Prediction Rules in a Clustered Data Setting*

Bernard Rosner,

Channing Division of Network Medicine, Brigham and Women's Hospital/Harvard Medical School,
181 Longwood Avenue, Boston, MA, 02115, U.S.A.

Weiliang Qiu, and

Channing Division of Network Medicine, Brigham and Women's Hospital/Harvard Medical School,
181 Longwood Avenue, Boston, MA, 02115, U.S.A.

Mei-Ling T. Lee

Department of Epidemiology and Biostatistics, University of Maryland, College Park, MD 20742,
U.S.A.

Abstract

The AUC (area under ROC curve) is a commonly used metric to assess discrimination of risk prediction rules; however, standard errors of AUC are usually based on the Mann-Whitney U test that assumes independence of sampling units. For ophthalmologic applications, it is desirable to assess risk prediction rules based on eye-specific outcome variables which are generally highly, but not perfectly correlated in fellow eyes (eg. progression of individual eyes to age-related macular degeneration (AMD)). In this article, we use the extended Mann-Whitney U test (Rosner et al, 2009) for the case where subunits within a cluster may have different progression status and assess discrimination of different prediction rules in this setting. Both data analyses based on progression of AMD and simulation studies show reasonable accuracy of this extended Mann-Whitney U test to assess discrimination of eye-specific risk prediction rules.

Keywords

risk prediction; ROC curves; clustered data; GEE

1 Introduction

Age-related macular degeneration (AMD) is the leading cause of irreversible visual impairment and blindness in the United States and other developed countries throughout the world (Seddon et al., 2003). Evidence is accumulating regarding modifiable factors that may decrease the risk of progression to the advanced forms of AMD. The Progression of Age-Related Macular Degeneration Study is a longitudinal study designed to measure multiple risk factors for the onset and progression of AMD. A total of 261 individuals were included in the analyses. Details about the study are provided in Seddon et al. (2003). The average age of the subjects was 72.3 years ($sd = 6.1$, $range = [60, 87]$). Approximately 61% were female. Subjects were followed for 4 to 6 years. We are interested in assessing whether the prediction of progression to the advanced form of AMD would be improved if we include total fat intake in the prediction model in addition to other AMD risk factors.

*This work was supported by the National Institutes of Health Grant EY12269 from the National Eye Institute.

A common criterion for assessing discrimination of risk prediction rules is the improvement in the area under the receiver-operating-characteristic curve (AUC) (e.g. Obuchowski, 2003; Zou et al., 2007). By using the equivalence between the Mann-Whitney U Statistic and AUC, (Hanley and McNeil, 1982), one can assess discrimination of risk prediction rules via the Mann-Whitney U test, which is commonly used in nonparametric two-group comparisons when the normality of the underlying distribution is questionable. However, an assumption of the Mann-Whitney U test is that sampling units (eg. eyes) are independent. For the AMD data set, progression status for the right and left eyes for a subject are correlated.

Moreover, previous works on non-parametric two-sample comparisons mainly focus on the shift alternative in the original data space (Hodges and Lehmann, 1956; Kendall and Stuart, 1969). However, (a) the meaning of the shift might be different for each underlying null distribution, and (b) the underlying null distribution is usually unknown. Rosner and Glynn (2009) proposed a shift alternative in a transformed data space based on the probit transformation and applied their method to assess and compare discrimination of risk prediction rules for a case-control AMD dataset where a subject is a case if either eye has AMD and a control if neither eye has AMD. However, their work requires the assumption that the observations are independent, which is not applicable to clustered data, where the eye is the unit of analysis.

Rosner et al. (2006) extended the Wilcoxon rank sum test, which is equivalent to the Mann-Whitney U test, for clustered data for two group comparisons where group membership is defined at the subunit level. In our example, there may be patients where one eye has progressed, while the fellow eye has not progressed.

In this article, we define the extended Mann-Whitney U statistic, which is equivalent to Rosner et al.'s (2006) extended Wilcoxon rank sum statistic, and derive its variance under the transformed shift alternative to assess discrimination of risk prediction rules in a clustered data setting with group membership defined at the subunit level. We also derive the variance of the difference of two extended Mann-Whitney U statistics to compare discrimination of different risk prediction rules when applied to the same data set.

The structure of the remaining parts of this article is as follows. In Section 2, we define the extended Mann-Whitney U statistic and derive its variance. In Section 3, we derive the variance of the difference of two extended Mann-Whitney U statistics for two prediction rules when applied to the same data set. In Section 4, multiple imputation approaches are described to incorporate uncertainty in the estimation of regression coefficients used in the prediction rule. In Section 5, we apply the extended Mann-Whitney U statistic to the AMD data set described in Section 1 to assess and compare discrimination of different risk prediction models for a prospective study of age-related macular degeneration (AMD), an important eye disease among the elderly resulting in substantial losses of vision. In Section 6, we conduct a simulation study to evaluate the performance of the extended Mann-Whitney U statistic in terms of bias of the variance estimate and the power of the extended Mann-Whitney U test. Section 7 is a discussion. Technique details are shown in the Web Appendices.

2 Extended Mann-Whitney U statistic for clustered data

Suppose that there are N independent clusters, where the subunit is the unit of analysis and the i -th cluster has g_i subunits, $i = 1, \dots, N$. We can use the following generalized estimating equations (GEE) model with exchangeable correlation structure to fit the AMD data since each eye has a progression score and these two scores for a subject are correlated:

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_k x_{ijk}, i=1, \dots, N, j=1, 2, \quad (1)$$

where $x_{ij\ell}$ is the value of the ℓ -th risk factor for the j -th eye of the i -th subject, $\ell=1, \dots, k$, $p_{ij} = \Pr(\text{progression for the } j\text{-th eye of the } i\text{-th subject} | x_{ij1}, \dots, x_{ijk})$, N is the number of subjects, and k is the number of risk factors. In our example, the initial AMD grade is an eye-specific risk factor, while other risk factors are person-specific.

Denote Z_{ij} as the prediction score for the j -th subunit of the i -th cluster based on the GEE Model (1) given by

$$Z_{ij} = \hat{\beta}_0 + \hat{\beta}_1 x_{ij1} + \cdots + \hat{\beta}_k x_{ijk},$$

where $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ are estimated via GEE Model (1). We are interested in testing if the distribution of prediction scores among subunits that progress is the same as that among subunits that do not progress.

To incorporate the information that subunits within a cluster are highly correlated, and that subunits within a cluster might have different progression status, we propose the following extended Mann-Whitney U statistic:

$$\hat{\eta}_c = \frac{\sum_{i=1}^N \sum_{j=1}^{g_i} \sum_{k=1}^N \sum_{\ell=1}^{g_k} U(Z_{ij} - Z_{k\ell}) (1 - \delta_{ij}) \delta_{k\ell}}{\sum_{i=1}^N \sum_{j=1}^{g_i} \sum_{k=1}^N \sum_{\ell=1}^{g_k} (1 - \delta_{ij}) \delta_{k\ell}}, \quad (2)$$

where $U(Z_{ij} - Z_{k\ell}) = 1$ if $Z_{ij} < Z_{k\ell}$, $= 1/2$ if $Z_{ij} = Z_{k\ell}$, and $= 0$ otherwise, for $i \neq k$, or $j \neq \ell$, and $\delta_{ij} = 1$ if the j -th subunit of the i -th cluster has progressed, $= 0$ if it has not progressed.

The statistic $\hat{\eta}_c$ is the proportion of pairs of subunits where the subunit that did not progress has a lower score than the subunit that did progress. If the proportion is equal to $1/2$, then there is no difference between the location parameter for subunits that progress versus subunits that do not progress. If the proportion is much greater than or smaller than $1/2$, then there exist evidence that the location parameters for the two groups of subunits are different.

We assume that (1) clusters are independent of each other; (2) δ_{ij} are fixed; (3)

$\Pr(Z_{ij} = Z_{k\ell}) = 0$ for $i \neq k$ or $j \neq \ell$. Furthermore, by definition the probit transformation $H = \Phi^{-1}$ will transform the response variable to a normal distribution. Let

$$H_{ij} \equiv H(F(Z_{ij})), \quad (3)$$

where F is the cumulative distribution of Z_{ij} . In data analysis, F will be estimated by the empirical cumulative distribution $\hat{F}_n$. We assume that

$$H_{ij} \sim \begin{cases} N(0, 1) & \text{if } \delta_{ij} = 0, \\ N(\mu, 1), \mu \neq 0 & \text{if } \delta_{ij} = 1, \end{cases}$$

and that after the transformation H , the bivariate random vector

$$\begin{pmatrix} H_{i_1 j_1} - H_{k_1 \ell_1} \\ H_{i_2 j_2} - H_{k_2 \ell_2} \end{pmatrix}$$

is bivariate normally distributed for any $i_1, k_1, i_2, k_2, j_1, \ell_1, j_2$, and ℓ_2 , with covariance matrix

$$\text{Cov} \begin{pmatrix} H_{ij} \\ H_{i\ell} \end{pmatrix} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, j \neq \ell, \text{ and } \text{Cov} \begin{pmatrix} H_{ij} \\ H_{k\ell} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, i \neq k.$$

We are interested in testing the hypotheses:

$$\begin{aligned} H_0: H(X) = H(Y) = H(Y_c) \quad \text{versus} \\ H_a: H(Y) = H(Y_c) + \mu, \mu \neq 0, \end{aligned}$$

where X is the prediction score of a randomly selected non-progressing subunit, Y is the prediction score of a randomly selected progressing subunit, and Y_c is the counterfactual random variable obtained if each subunit that had progressed actually had not progressed. We refer to H_a as a *probit-shift alternative*. The probit-shift alternative is useful in calculating $\text{Var}(\widehat{\eta}_c)$ under H_a in closed-form, which we will need to (a) obtain confidence limits for η_c ; (b) compare η_c between two risk prediction rules assessed on the same subjects; and (c) compute power for future studies.

We can obtain the expected value of the extended Mann-Whitney U statistic $\widehat{\eta}_c$ under the alternative hypothesis:

$$E(\widehat{\eta}_c | H_a) = \frac{\theta_c \sum_{i=1}^N c_i d_i + \theta [C \cdot D - \sum_{i=1}^N c_i d_i]}{C \cdot D}, \quad (4)$$

where

$$\theta_c = \Phi\left(\frac{\mu}{\sqrt{2(1-\rho)}}\right), \theta = \Phi\left(\frac{\mu}{\sqrt{2}}\right),$$

and $c_k = \sum_{\ell=1}^{g_k} \delta_{k\ell}$ is the number of progressing subunits for the k -th subject, $d_i = \sum_{j=1}^{g_i} (1 - \delta_{ij})$ is the number of non-progressing subunits for the i -th subject, $C = \sum_{k=1}^N c_k$, $D = \sum_{i=1}^N d_i$. Note that for clustered data, there are two θ s, whereas in non-clustered data there is only a single θ . θ_c is used for comparison of nonprogressing and progressing subunits (e.g. eyes) within the same cluster (e.g. subject), and θ for comparison of nonprogressing and progressing subunits (e.g. eyes) in different clusters (e.g. different subjects). Under the null hypothesis, i.e., $\mu = 0$, $\theta_c = \theta = 1/2$, and $E(\widehat{\eta}_c | H_0) = 1/2$.

To derive the variance of $\widehat{\eta}_c$, we rewrite $\widehat{\eta}_c$ as

$$\widehat{\eta}_c = \frac{A+B}{C \cdot D} \quad (5)$$

where

$$\begin{aligned}
 A &= \sum_{i=1}^N \sum_{j=1}^{g_i} \sum_{\ell=1}^{g_i} U(Z_{ij} - Z_{i\ell}) (1 - \delta_{ij}) \delta_{i\ell} \\
 B &= \sum_{i=1}^N \sum_{k=1, k \neq i}^N \sum_{j=1}^{g_i} \sum_{\ell=1}^{g_k} U(Z_{ij} - Z_{k\ell}) (1 - \delta_{ij}) \delta_{k\ell}
 \end{aligned}$$

Then

$$\text{Var}(\widehat{\eta}_c | H_a) = \frac{1}{C^2 \cdot D^2} [\text{Var}(A | H_a) + \text{Var}(B | H_a) + 2\text{Cov}(A, B | H_a)]. \quad (6)$$

We can use second order moments of the bivariate normal distribution to derive a closed-form expression for $\text{Var}(\widehat{\eta}_c | H_a)$. The complete derivation of $\text{Var}(\widehat{\eta}_c | H_a)$ is given in Web Appendix A.

3 Variance of the difference of two extended Mann-Whitney U statistics applied to the same data set

Suppose we have two prediction rules (indexed by $t = 1, 2$) applied to the same data set. We can compare the area under ROC curves (AUCs) of two prediction rules by testing the hypothesis $H_0: \eta_c^{(1)} = \eta_c^{(2)}$ versus $H_a: \eta_c^{(1)} \neq \eta_c^{(2)}$, where $\eta_c^{(t)}$ = AUC for the t -th prediction rule, $t = 1, 2$. We estimate $\eta_c^{(t)}$ by:

$$\widehat{\eta}_c^{(t)} = \frac{A^{(t)} + B^{(t)}}{C \cdot D}, \quad t = 1, 2,$$

and

$$A^{(t)} = \sum_{i=k=1}^N \sum_{j=1}^{g_i} \sum_{\ell=1}^{g_i} (1 - \delta_{ij}) \delta_{i\ell} U(Z_{ij}^{(t)} - Z_{i\ell}^{(t)})$$

and

$$B^{(t)} = \sum_{i=1}^N \sum_{k=1, k \neq i}^N \sum_{j=1}^{g_i} \sum_{\ell=1}^{g_k} (1 - \delta_{ij}) \delta_{k\ell} U(Z_{ij}^{(t)} - Z_{k\ell}^{(t)}).$$

Since the two prediction rules are applied to the same data set, the two extended Mann-Whitney U statistics are correlated. To calculate the variance of $\widehat{\eta}_{c1} - \widehat{\eta}_{c2}$, we assume the following correlation structure:

$$\begin{aligned}
 \rho &= \text{Cov}(H_{ij}^{(1)}, H_{ij}^{(2)}) = \text{Corr}(H_{ij}^{(1)}, H_{ij}^{(2)}), \\
 \rho_{11} &= \text{Cov}(H_{ij_1}^{(1)}, H_{ij_2}^{(1)}) = \text{Corr}(H_{ij_1}^{(1)}, H_{ij_2}^{(1)}), \\
 \rho_{22} &= \text{Cov}(H_{ij_1}^{(2)}, H_{ij_2}^{(2)}) = \text{Corr}(H_{ij_1}^{(2)}, H_{ij_2}^{(2)}), \\
 \rho_{12} &= \text{Cov}(H_{ij_1}^{(1)}, H_{ij_2}^{(2)}) = \text{Corr}(H_{ij_1}^{(1)}, H_{ij_2}^{(2)}).
 \end{aligned}$$

We denote

$$\theta_{c_t} = \Phi\left(\frac{\mu_t}{\sqrt{2(1-\rho_{tt})}}\right), \theta_t = \Phi\left(\frac{\mu_t}{\sqrt{2}}\right), t=1, 2.$$

We wish to calculate

$$\text{Var}(\widehat{\eta}_c^{(1)} - \widehat{\eta}_c^{(2)}) = \text{Var}(\widehat{\eta}_c^{(1)}) + \text{Var}(\widehat{\eta}_c^{(2)}) - 2\text{Cov}(\widehat{\eta}_c^{(1)}, \widehat{\eta}_c^{(2)}). \quad (7)$$

We already derived the formula for $\text{Var}(\widehat{\eta}_c^{(t)})$ in Equation (6). Hence, we only need to derive $\text{Cov}(\widehat{\eta}_c^{(1)}, \widehat{\eta}_c^{(2)})$. We can decompose it into 4 parts:

$$\text{Cov}(\widehat{\eta}_c^{(1)}, \widehat{\eta}_c^{(2)}) = \frac{1}{C^2 D^2} [\text{Cov}(A^{(1)}, A^{(2)}) + \text{Cov}(A^{(1)}, B^{(2)}) + \text{Cov}(B^{(1)}, A^{(2)}) + \text{Cov}(B^{(1)}, B^{(2)})]$$

The derivation of $\text{Var}(\widehat{\eta}_c^{(1)} - \widehat{\eta}_c^{(2)} | H_a)$ in Equation 7 is given in Web Appendix B.

Hence, a large sample test statistic to test the hypotheses $H_0: \eta_c^{(1)} = \eta_c^{(2)}$ versus $H_a: \eta_c^{(1)} \neq \eta_c^{(2)}$ is given by $Z_{12} = (\widehat{\eta}_c^{(1)} - \widehat{\eta}_c^{(2)}) / [\widehat{\text{Var}}(\widehat{\eta}_c^{(1)} - \widehat{\eta}_c^{(2)})]^{1/2} \sim N(0, 1)$ under H_0 .

Finally, we can calculate the power of the test $H_0: \Delta = 0$ vs $H_a: \Delta = \Delta_0$, where $\Delta_0 = \eta_c^{(1)} - \eta_c^{(2)}$ using the formula

$$\text{power} = 1 - \Phi\left(\frac{\sigma_0 Z_{\alpha/2} - \Delta_0}{\sigma_1}\right) + \Phi\left(\frac{-\sigma_0 Z_{\alpha/2} - \Delta_0}{\sigma_1}\right) \quad (8)$$

where $Z_{\alpha/2}$ is the upper $100\alpha/2\%$ percentile of the standard normal distribution $N(0, 1)$, $\sigma_0 = \sqrt{\text{Var}(\widehat{\Delta} | H_0)}$ and $\sigma_1 = \sqrt{\text{Var}(\widehat{\Delta} | H_a)}$.

4 Multiple imputation

When constructing scores, we used estimated regression coefficients from GEE. Hence, we did not account for the variation of these estimates in the calculation of the variance of $\widehat{\eta}_c$.

One remedy is to use multiple imputation. Specifically, we randomly generate m random vectors $\boldsymbol{\beta}$ from the multivariate normal distribution $N(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}})$, where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}$ are the estimated regression coefficients and their variance-covariance matrix.

Then we can obtain m estimates of η_c . Denote these as $\widehat{\eta}_{c,i}$, $i = 1, \dots, m$. We then use

$$\widehat{\eta}_c^* = \frac{1}{m} \sum_{i=1}^m \widehat{\eta}_{c,i}$$

as a final estimate of η_c , which will reflect both between- and within-imputation variance.

We next calculate the variance $\text{Var}(\widehat{\eta}_c^*)$ by the following formula (Rubin, 1987)

$$\text{Var}(\widehat{\eta}_c^*) = \frac{1}{m} \sum_{i=1}^m \text{Var}(\widehat{\eta}_{c,i}) + \left(\frac{1 + \frac{1}{m}}{m-1} \right) \sum_{i=1}^m (\widehat{\eta}_{c,i} - \widehat{\eta}_c^*)^2,$$

where $\text{Var}(\widehat{\eta}_{c,i})$ is calculated using the usual formula.

We will calculate the p-value for a parameter estimate based on multiple imputation by using the method mentioned in Rubin (1987) (c.f. Web Appendix C).

We can obtain the adjusted estimate $\widehat{\eta}_c^{(1)} - \widehat{\eta}_c^{(2)}$ and its variance and confidence interval using the same approach.

5 Example

In this section, we apply the proposed method to the AMD data set (with $n = 261$ subjects) mentioned in Section 1. The data set is from the prospective longitudinal study of AMD: the Progression of Age-Related Macular Degeneration Study designed to measure multiple risk factors for the onset and progression of AMD. One modifiable factor is dietary total fat intake. The results of Seddon et al. (2003) show that higher total fat intake increased the risk of progression to the advanced forms of AMD, with a relative risk (RR) of 2.90 (95% confidence interval, 1.15-7.32) for the highest fat-intake quartile relative to the lowest fat-intake quartile, after controlling for other factors (P trend = 0.01).

The increased risk of the complications of AMD are indicated by drusen, which are yellow deposits under the retina (National Eye Institute, 2011). According to Seddon et al. (2003), eyes with extensive small drusen (≥ 15 drusen; with size of drusen $< 63 \mu m$), nonextensive intermediate drusen (< 20 drusen; with size of drusen $\geq 63 \mu m$ but $< 125 \mu m$), or pigment abnormalities associated with AMD were assigned a grade of 2. Eyes with extensive intermediate or large drusen (size of drusen $\geq 125 \mu m$) were assigned a grade of 3. Eyes with geographic atrophy received a grade of 4. If there was evidence of retinal pigment epithelial detachment or choroidal neovascular membrane, a grade of 5 was assigned. Eyes received a grade of 1 if none of these signs was present. All eyes in our analysis had a grade of 2 and 4 at baseline. Each subject contributed 2 eyes to the analysis. Advanced AMD is defined as grades 4 or 5.

Progression to advanced AMD in an eye over 4 to 6 years was defined either as progression from a grade of less than 4 at baseline to grades 4 or 5 at any follow-up visit, or progression from grade 4 at baseline to grade 5 at any follow-up visit.

If a set of risk factors has no predictive ability to distinguish progressing versus non-progressing eyes, the value of the extended Mann-Whitney U statistic

$$\widehat{\eta}_c = \frac{\sum_{i=1}^N \sum_{j=1}^{g_i} \sum_{k=1}^N \sum_{\ell=1}^{g_k} U(Z_{ij} - Z_{k\ell}) (1 - \delta_{ij}) \delta_{k\ell}}{\sum_{i=1}^N \sum_{j=1}^{g_i} \sum_{k=1}^N \sum_{\ell=1}^{g_k} (1 - \delta_{ij}) \delta_{k\ell}},$$

will be close to the expected null value $E(\widehat{\eta}_c | H_0) = 1/2$. In general, we will use $\widehat{\eta}_c$ to assess and compare the discrimination of different risk prediction rules.

If we use the same dataset to both construct a prediction rule and calculate prediction accuracy, the prediction accuracy might be over-estimated. Hence, we randomly split the

261 subjects into 2 groups. One group with 131 subjects is used as the training set to obtain the estimated regression coefficients $\widehat{\beta}$ in the GEE model. The other group with 130 subjects is used as the testing set to obtain the estimates $\widehat{\eta}_c^{(1)}$, $\widehat{\eta}_c^{(2)}$, and related estimates and p-values. When constructing prediction scores for subjects in testing set, we used the regression coefficients estimated from the training set. To account for the variability of these estimated regression coefficients, we used the multiple imputation approach mentioned in Section 4. Hence, for each imputation, we obtain different estimates of η_c in the testing set and use multiple imputation approaches to obtain an overall estimate of η_c that reflects both between and within imputation variations.

Table 1 shows the parameter estimates of two prediction rules based on risk factors: BMI, age, gender, cardiovascular disease (cvd), systolic blood pressure (sbp), current and past smoking (smkcur and smkpst), ln(caloric intake) (lncalor), beta carotene intake (adjbcaro), alcohol intake (alco), initial eye grade (inieye3 and inieye4), protein intake (rprot1, rprot2, and rprot3), and total fat intake (rtotfat1, rtotfat2, and rtotfat3).

The regression coefficients and their standard errors, p-values, and 95% confidence intervals (CIs) were based on the training set (with $n = 131$ subjects). The estimates $\widehat{\eta}_c^{(1)}$ and $\widehat{\eta}_c^{(2)}$ and their standard errors and p-values were estimated based on the testing set (with $n = 130$ subjects).

In the training set, there are 131 subjects (262 eyes). The number of eyes that progressed was 51. There are 9 subjects where both eyes progressed, 17 subjects where the right eye progressed and the left eye did not progress, 16 subjects where the right eye did not progress and the left eye progressed, and 88 subjects where both eyes did not progress. The odds ratio for progression between fellow eyes is $9 * 88 / (17 * 16) = 2.91$. The correlation between risk scores for fellow eyes was 0.54 in the training set and 0.51 in the testing set.

Table 1 shows that there were significant effects of total fat intake in model 1. The relative risks and 95% CIs of rtotfat1 (second quartile), rtotfat2 (third quartile), and rtotfat3 (fourth quartile) relative to rtotfat0 (first quartile) are 1.7 (95% CI: [0.4, 6.7], p-value = 0.48), 4.3 (95% CI: [0.9, 20.1], p-value = 0.06), and 22.4 (95% CI: [4.2, 118.6], p-value < 0.01). However, there was no significant difference between the AUCs for the two prediction rules (multiple-imputation-based estimates: $\widehat{\mu}^{(1)}=0.406$, $\widehat{\mu}^{(2)}=0.381$, $\widehat{\theta}^{(1)}=0.613$, $\widehat{\theta}_c^{(1)}=0.718$, $\widehat{\theta}^{(2)}=0.606$, $\widehat{\theta}_c^{(2)}=0.650$, $AUC_1 = 0.677$ versus $AUC_2 = 0.672$, p-value= 0.93).

6 A simulation study

We conducted a simulation study to evaluate the performance of the extended Mann-Whitney U statistic on comparing two prediction rules. The values of the model parameters were set to be the estimated parameters for the AMD data set (see Section 5). We assume that all subjects have two subunits ($g = 2$).

Denote

$$\mathbf{H}_i = \left(H_{i1}^{(1)}, \dots, H_{ig}^{(1)}, H_{i1}^{(2)}, \dots, H_{ig}^{(2)} \right)^T, i=1, \dots, N,$$

as the transformed values for the g subunits of the i -th cluster obtained from two prediction rules, where $\left(H_{i1}^{(t)}, \dots, H_{ig}^{(t)} \right)$ are from the t -th prediction rule. We assume $\mathbf{H}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where

$$\mu_{ki} = \begin{pmatrix} \mu_{ki1} \\ \mu_{ki2} \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}$$

and $\mu_{kij} = \mu_k \delta_{ij}$ is the mean probit for the j -th subunit in the i -th cluster for the k -th prediction rule, $k = 1, 2$, $i = 1, \dots, N$, $j = 1, 2$,

$$\Sigma_{11} = \begin{pmatrix} 1 & & \rho_{11} \\ & \ddots & \\ \rho_{11} & & 1 \end{pmatrix}, \Sigma_{22} = \begin{pmatrix} 1 & & \rho_{22} \\ & \ddots & \\ \rho_{22} & & 1 \end{pmatrix}, \Sigma_{12} = \begin{pmatrix} \rho & & \rho_{12} \\ & \ddots & \\ \rho_{12} & & \rho \end{pmatrix}.$$

where $\delta_{ij} = 1$ if the j -th eye ($j = 1$ means left eye; $j = 2$ means right eye) of the i -th subject has progressed, and $= 0$ if it has not progressed.

To simulate the data, we set $\mu_1 = 0.80$, $\rho = 0.93$, $\rho_{11} = 0.59$, $\rho_{22} = 0.56$, and $\rho_{12} = 0.52$, as were estimated using the whole AMD data set ($n = 261$ subjects). We considered 4 sample sizes: $N = 50, 100, 200$, and 500 .

We generated 4,000 simulated data sets for each scenario. Since δ_{ij} is assumed to be fixed (c.f. Assumption 2 in Section 2), for each of 4,000 simulated data sets in a scenario, we used the same set of δ_{ij} , $i = 1, \dots, N$, $j = 1, 2$, where δ_{i1} (indicating left eye's progression status) and δ_{i2} (indicating right eye's progression status), $i = 1, \dots, N$, are generated from multinomial distributions with parameters $p_{11} = 0.115$ (proportion where both eyes progressed), $p_{10} = 0.142$ (proportion where only the left eye progressed), $p_{01} = 0.130$ (proportion where only the right eye progressed), and $p_{00} = 0.613$ (proportion where both eyes did not progress), respectively, where

$$\begin{aligned} p_{11} &= Pr(\delta_{i1}=1 \ \& \ \delta_{i2}=1), \\ p_{10} &= Pr(\delta_{i1}=1 \ \& \ \delta_{i2}=0), \\ p_{01} &= Pr(\delta_{i1}=0 \ \& \ \delta_{i2}=1), \\ p_{00} &= Pr(\delta_{i1}=0 \ \& \ \delta_{i2}=0), \\ 1 &= p_{11} + p_{10} + p_{01} + p_{00}. \end{aligned}$$

We are interested in testing the null hypothesis $H_0 : \Delta = 0$ versus the alternative hypothesis $H_a : \Delta = \Delta_0$, where

$$\Delta = \eta_c^{(1)} - \eta_c^{(2)}, \quad (9)$$

and

$$\eta_c^{(t)} = \left\{ \theta_c^{(t)} \sum_{i=1}^N c_i d_i + \theta^{(t)} \left[C \cdot D - \sum_{i=1}^N c_i d_i \right] \right\} / (C \cdot D), t=1, 2.$$

We consider three values (0, 0.025, and 0.05) for Δ_0 . We can obtain μ_2 by solving equation (9) using numerical methods, given μ_1 , ρ_{11} , ρ_{22} , Δ , and δ_{ij} , since there is no explicit closed-form expression to express μ_2 as a function of these variables.

We evaluate the performance of the extended Mann-Whitney U statistic in terms of percent bias $\left(100 \sum_{b=1}^B (\widehat{\Delta}_b - \Delta) / \Delta\right)$ where $\Delta = 0$, and V-statistic $\left(1 / (B - 1) \sum_{b=1}^B (\widehat{\Delta}_b - \widehat{\Delta})^2 / \text{Var}(\widehat{\Delta} | H_0)\right)$ of parameter estimates $\widehat{\Delta}_b$, and coverage of the confidence interval $[\widehat{\Delta}_{L,b}, \widehat{\Delta}_{U,b}]$ of the parameter Δ , Type I error rate and power of the hypothesis test $H_0 : \Delta = 0$ versus $H_a : \Delta = \Delta_0$.

The power of the test is computed using Equation (8) and the Type I error rate is calculated by setting $\Delta_0 = 0$.

The simulation results are shown in Table 2. We can see that the estimate $\widehat{\Delta}$ is unbiased as the percent bias is close to zero. The sample variance tends to be slightly smaller than the theoretical variance when $nSubj = 50$ as the V-statistic is smaller than one. The coverage is close to the nominal level 95%, except when $nSubj = 50$ where the procedure slightly over-covers (consistent with the V-statistic). The theoretical and empirical powers are also in close agreement.

7 Discussion

In this article, we provide methods for assessing discrimination of risk prediction rules for disease progression as characterized by AUC, where the unit of analysis is the subunit (e.g. the eye) within a cluster (e.g., the person) and there is correlation between risk scores for multiple subunits (eyes) in the same cluster (person). The methods are applicable to both balanced (equal cluster size) and unbalanced (unequal cluster size) clustered data. The data analysis and the simulation study show that the proposed test performs well in assessing AUC of risk prediction rules and comparing AUC of competing risk prediction rules estimated from the same subjects.

We assumed that after probit transformation the bivariate random vector

$[H_{i_1 j_1} - H_{k_1 \ell_1}, H_{i_2 j_2} - H_{k_2 \ell_2}]$ is bivariate normally distributed. For the AMD data set, we test the bivariate normality of (H_{ij}, H_{kl}) using the Shapiro-Wilk test for the training set and the p-value = 0.22 indicates we do not reject the bivariate normality assumption. It will be a future research topic to check the robustness of our method to the violation of the bivariate normality assumption.

Another assumption of the extended Mann-Whitney U test is that replicates within a cluster are exchangeable. This is appropriate for the AMD data in which the progression status of the left eye and right eye for a subject are ascertained at one point in time. An interesting extension would be to consider a non-exchangeable within-cluster correlation structure, as might be applicable for longitudinal data.

Obuchowski and McClish (1997) have also considered nonparametric analysis of clustered ROC curve data. Inference is based on the statistic $\widehat{\theta}_c$ which is equivalent to $\widehat{\eta}_c$ in Equation (2). A U statistic approach is used to calculate $\text{Var}(\widehat{\theta}_c)$ based on the quantities

$$V_{10}(Z_{ij}) = \frac{1}{D} \sum_{k=1}^N \sum_{\ell=1}^{c_k} U(Z_{ij} - Z_{k\ell})(1 - \delta_{k\ell}),$$

$$V_{01}(Z_{k\ell}) = \frac{1}{C} \sum_{i=1}^N \sum_{j=1}^{d_i} U(Z_{ij} - Z_{k\ell})\delta_{ij}.$$

An inherent assumption of this approach is that under H_a , $E[U(Z_{ij} - Z_{kl})]$ is the same when $i = k$ or when $i \neq k$. Under the probit shift alternative in the current paper, there are two distinct parameters θ_c and θ when $i = k$ or when $i \neq k$ (see Equation 4). Hence, variances and covariances need to be computed separately for the components A and B corresponding to θ_c and θ , respectively. Obuchowski and McClish (1997) also provide upper bounds on sample size estimates based on variance components in the observed data. In the current manuscript, we provide an explicit power formula (see Equation 8) as a function of μ , the location parameter of the probit shift under H_a , and $\rho = \text{Corr}(H_{i1}, H_{i2}) = \text{correlation}$ between probit scores between two subunits in the same cluster. This power formula has shown to be accurate in our simulation studies. An advantage of this approach is that power calculations can be performed for various levels of μ and ρ , whereas in Obuchowski and McClish (1997) they are based on a specific dataset. An interesting design question is for a fixed total number of subunits $= C + D$, what is the optimal allocation between having many clusters with small cluster size vs having fewer clusters with large cluster size so as to minimize $\text{Var}[\hat{\eta}_c^{(1)} - \hat{\eta}_c^{(2)}]$.

Li and Zhou (2008) provide a unified approach to nonparametric comparison of ROC curves for clustered data. The asymptotic joint distribution of ROC curves defined by two different markers accounting for both between marker and within-cluster (subject) variation is

provided. The difference between AUC's corresponding to Δ is estimated from $\int_0^1 D(\rho) d\rho$, where

$$\begin{aligned} D(\rho) &= \text{ROC}^{(1)}(\rho) - \text{ROC}^{(2)}(\rho) \\ \text{ROC}^{(v)}(\rho) &= 1 - G^{(v)}\left\{\left(F^{(v)}\right)^{-1}(1 - \rho)\right\} \end{aligned}$$

$F^{(v)}$ and $G^{(v)}$ are the cumulative distribution function of $Z^{(v)}$ for cases (progressors) and controls (non-progressors), respectively. Unfortunately, there is in general no closed-form expression available for $\text{Var}(\widehat{D}(\rho))$ and Monte-Carlo simulation and numerical integration is used instead. A nice feature of this approach is the avoidance of nonparametric density estimate, which can be computationally challenging in this setting. However, in the current paper, we use the rank-preserving property of the probit transformation to obtain a closed-form expression for $\text{Var}[\hat{\eta}_c^{(1)} - \hat{\eta}_c^{(2)}]$ in Equation (7) which can be used for any continuous or ordinal scale.

Furthermore, another distinction between our paper and Obuchowski and McClish (1997) and Li and Zhou (2008) is that our scores are derived from a regression model as a function of one or more risk factors rather than based on a single diagnostic marker as is common in the ROC literature. Since the scores are based on estimated regression coefficients, the variability in estimation needs to be taken into account which we accomplished using multiple imputation methods.

Finally, Toledano and Gatsonis (1996) propose an ordinal regression model to estimate effects of covariates on the area under the ROC, and compare AUC between competing risk prediction rules, where parameters in the marginal ordinal regression model are estimated using GEE approaches. Our example is based on a continuous risk score, where there are few patients with exactly the same risk score. It is an open question how the Toledano and Gatsonis's (1996) approach would work with a sparse number of subjects in individual risk categories.

Recently, new measures of performance of prediction models have been proposed, such as net reclassification improvement (NRI) and integrated discrimination improvement (IDI) proposed by Pencina et al. (2008). These new measures offer incremental information over the AUC. Pencina et al. (2008) pointed out in their Discussion that they still believe that improvement in the AUC should remain the first criterion, while NRI and IDI should also be taken into consideration. To our knowledge, these new measures have not been extended to handle clustered data yet. Hence, one possible future research area is to extend these new measures for clustered data.

The program that implements the methods in this paper was written in the R language and is available from the authors upon request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

- Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Diagnostic Radiology*. 1982; 143:29–36.
- Hodges JL Jr. Lehmann EL. The efficiency of some nonparametric competitors of the t test. *Annals of Mathematical Statistics*. 1956; 27:324–335.
- Kendall, MG.; Stuart, A. *The Advanced Theory of Statistics*. Hafner; New York: 1969.
- Li G, Zhou K. A unified approach to nonparametric comparison of receiver operating characteristic curves for longitudinal and clustered data. *Journal of the American Statistical Association*. 2008; 103(482):705–713. [PubMed: 20209021]
- National Eye Institute. [March, 2011] Age-related macular degeneration. 2011. from <http://www.nei.nih.gov/health/maculardegen/armdfacts.asp>
- Obuchowski NA. Receiver operating characteristic curves and their use in radiology. *Radiology*. 2003; 229:3–8. [PubMed: 14519861]
- Obuchowski NA, McClish DK. Sample size determination for diagnostic accuracy studies involving binormal roc curve indices. *Statistics in Medicine*. 1997; 16(13):1529–1542. [PubMed: 9249923]
- Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr. Vasan RS. Evaluating the added predictive ability of a new marker: From area under the roc curve to reclassification and beyond. *Statistics in Medicine*. 2008; 27:157–172. [PubMed: 17569110]
- Rosner B, Glynn RJ. Power and sample size estimation for the wilcoxon rank sum test with application to comparisons of c statistics from alternative prediction models. *Biometrics*. 2009; 65:188–97. [PubMed: 18510654]
- Rosner B, Glynn RJ, Lee MT. Extension of the rank sum test for clustered data: Two-group comparisons with group membership defined at the subunit level. *Biometrics*. 2006; 62:1251–1259. [PubMed: 17156300]
- Rubin, DB. *Multiple imputation for non-response in surveys*. Wiley; New York: 1987.
- Seddon JM, Cote J, Rosner B. Progression of age-related macular degeneration. *Archives of Ophthalmology*. 2003; 121:1728–1737. [PubMed: 14662593]
- Toledano AY, Gatsonis CA. GEEs for ordinal categorical data: arbitrary patterns of missing responses and missingness in a key covariate. *Biometrics*. 1996; 55:488–496. [PubMed: 11318205]
- Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*. 2007; 115:654–657. [PubMed: 17283280]

Table 1

Comparison of the two prediction rules for the AMD data set based on GEE (Number of subjects = 131, Number of eyes = 262, Number of eyes which progressed = 51)

training set	Prediction rule 1				Prediction rule 2			
	(model with total fat intake)				(model without total fat intake)			
variable ^a	Estimate	s.e.	pval	RR	95% CI	Estimate	s.e.	pval
(Intercept)	-7.95	1.58	0.00			-6.15	1.42	0.00
bmi2529	1.72	0.65	0.01			1.72	0.55	0.00
bmi30+	1.86	0.69	0.01			1.43	0.59	0.02
male6069	-0.61	0.80	0.44			-0.15	0.82	0.86
male7079	1.42	0.86	0.10			1.41	0.79	0.07
male80+	0.64	1.18	0.59			0.50	1.22	0.68
feml7079	1.74	0.86	0.04			1.21	0.77	0.11
feml80+	0.82	1.16	0.48			0.27	0.97	0.78
cvd	0.97	0.54	0.07			0.73	0.48	0.13
sysbp _c	-0.02	0.01	0.15			-0.02	0.01	0.14
smkeur	1.58	0.77	0.04			1.46	0.69	0.04
smkpst	0.14	0.50	0.78			0.13	0.50	0.79
ln calor _c	-3.95	1.18	0.00			-1.37	1.00	0.17
adjbcaro _c	0.43	0.27	0.11			0.28	0.31	0.37
alco _c	0.01	0.01	0.42			0.00	0.02	0.92
inিয়ে3	3.15	0.67	0.00			3.19	0.64	0.00
inিয়ে4	2.06	0.67	0.00			2.27	0.73	0.00
rprot1	0.33	0.51	0.52			0.12	0.51	0.81
rprot2	0.53	0.75	0.48			0.54	0.78	0.49
rprot3	-0.51	1.01	0.61			-0.50	0.92	0.58
rtofat1	0.51	0.71	0.48	1.7	[0.4, 6.7]			
rtofat2	1.47	0.78	0.06	4.3	[0.9, 20.1]			
rtofat3	3.11	0.85	0.00	22.4	[4.2, 118.6]			

Prediction rule 1				Prediction rule 2				
(model with total fat intake)				(model without total fat intake)				
training set	Estimate	s.e.	pval	RR	95% CI	Estimate	s.e.	pval
variable ^a	Estimate	s.e.	pval	RR	95% CI	Estimate	s.e.	pval
testing set	$\hat{\eta}_c^{(1)} = 0.677(sd = 0.050, df = 48.14)$					$\hat{\eta}_c^{(2)} = 0.672(sd = 0.049, df = 38.13)$		
$\hat{\eta}_c^{(1)} - \hat{\eta}_c^{(2)} = 0.005(sd = 0.057, df = 7.57), p\text{-value} = 0.93$								
$corr(Z_1, Z_2)^b$				0.51				

^a bmi2529 = BMI between 25 and 29.9; bmi30+ = BMI greater or equal to 30; male6069 = male aged between 60 and 69; male7079 = male aged between 70 and 79; male80+ = male aged greater than or equal to 80; feml7079 = female aged between 70 and 79; feml80+ = female aged greater than or equal to 80; cvd = cardiovascular disease; sysbp_c = mean-centered systolic blood pressure (mean = 140.17); smkcur = current smoker; smkpast = past smoker; lncalor_c = mean-centered log caloric intake (mean = 7.25); adjbcaro_c = mean-centered calorie-adjusted beta carotene intake (μg/d, values are expressed as geometric mean after sex-specific energy adjustment) (mean = 8.11); alco_c = mean-centered alcohol intake (g/d), (mean = 6.77); inieye3 = 1 if initial eye grade = 3, = 0 otherwise; inieye4 = 1 if initial eye grade = 4, = 0 otherwise; rprot1 = protein intake (second quartile); rprot2 = protein intake (third quartile); rprot3 = protein intake (fourth quartile); rtofat1 = total fat intake (second quartile); rtofat2 = total fat intake (third quartile); rtofat3 = total fat intake (fourth quartile). The quantiles of total fat intake are sex-specific.

^b estimated correlation between the scores of the two eyes for the same subject after adjusting for the above covariates, based on training set.

Table 2

Simulation results ($\rho = 0.93, \rho_{11} = 0.59, \rho_{22} = 0.56, \rho_{12} = 0.52$)

$\Delta = 0$					
nSubj	pBias ^a	coverage	V-stat ^b	Type I error rate	emp. Type I error rate ^c
50	-	96.1	0.84	0.05	0.04
100	-	95.8	0.91	0.05	0.04
200	-	95.1	0.93	0.05	0.05
500	-	94.9	0.98	0.05	0.05
$\Delta = 0.025$					
nSubj	pBias	coverage	V-stat	power ^d	emp. power ^e
50	2.23	96.4	0.85	0.15	0.13
100	0.62	95.6	0.95	0.21	0.20
200	0.05	94.9	0.98	0.45	0.43
500	0.21	95.2	0.99	0.84	0.84
$\Delta = 0.05$					
nSubj	pBias	coverage	V-stat	power ^d	emp. power ^e
50	-0.05	96.4	0.87	0.46	0.45
100	0.27	95.0	1.01	0.72	0.72
200	-0.31	94.9	0.97	0.94	0.94
500	0.42	94.6	1.01	1.00	1.00

^a pBias is the percent bias $\left(100 \sum_{b=1}^B (\widehat{\Delta}_b - \Delta) / \Delta\right)$ where $\Delta = 0$ and B is the total number of simulated data sets.

^b V-stat is the ratio of the empirical variance to the average theoretical variance.

^c emp. Type I error rate = empirical Type I error rate = the proportion of simulated data sets for which the test statistic $Z = |\widehat{\Delta}| / sd(\widehat{\Delta} | H_0)$ is greater than the critical value $Z_{\alpha/2}$ given $\Delta = 0$, where $Z_{\alpha/2}$ is the upper $100(\alpha/2)\%$ percentile of a standard normal distribution.

^d power = theoretical power from Equation (8).

^e emp. power = empirical power = the proportion of simulated data sets for which the test statistic $Z = |\widehat{\Delta}| / sd(\widehat{\Delta} | H_0)$ is greater than the critical value $Z_{\alpha/2}$ given $\Delta > 0$.