Journal of Data Science **3**(2005), 257-278

Testing Statistical Significance of the Area under a Receiving Operating Characteristics Curve for Repeated Measures Design with Bootstrapping

Honghu Liu, Gang Li,
William G. Cumberland and Tongtong Wu
University of California at Los Angeles

Abstract: Receiver operating characteristic (ROC) curve is an effective and widely used method for evaluating the discriminating power of a diagnostic test or statistical model. As a useful statistical method, a wealth of literature about its theories and computation methods has been established. The research on ROC curves, however, has focused mainly on cross-sectional design. Very little research on estimating ROC curves and their summary statistics, especially significance testing, has been conducted for repeated measures design. Due to the complexity of estimating the standard error of a ROC curve, there is no currently established statistical method for testing the significance of ROC curves under a repeated measures design. In this paper, we estimate the area of a ROC curve under a repeated measures design through generalized linear mixed model (GLMM) using the predicted probability of a disease or positivity of a condition and propose a bootstrap method to estimate the standard error of the area under a ROC curve for such designs. Statistical significance testing of the area under a ROC curve is then conducted using the bootstrapped standard error. The validity of bootstrap approach and the statistical testing of the area under the ROC curve was validated through simulation analyses. A special statistical software written in SAS/IML/MACRO v8 was also created for implementing the bootstrapping algorithm, conducting the calculations and statistical testing.

Key words: Area under ROC Curve, bootstrapping, generalized linear mixed model (GLMM), standard error, simulation.

1. Introduction

Receiver operating characteristic (ROC) curves display the relationship between sensitivity (true-positive rate) and 1-specificity (false-positive rate) across all possible threshold values that define a disease or positivity of a condition. ROC curves show the full picture of trade-off between true-positive rate and false-positive rate at different levels of positivity for the specific question we are studying. ROC originated from signal detection theory and psychophysics (Green and Swets, 1966) and later found its wide range of extensive applications in different fields such as biology, psychology and radiology and medicine (Metz, 1986; Pepe, 1998; Zou, 2001). It is a very useful diagnostic tool, particularly in medical and image research. Now, ROC curves have been widely used for evaluating the accuracy and discriminating power of a diagnostic test/bio-marker or statistical model/methods. For example, the accuracy of a medical diagnostic test can be typically described by a ROC curve through sensitivity and specificity. Summary measures of ROC curves, such as the area under a curve (AUC) or the projected length of a ROC curve (PLC) and the area swept out of a ROC curve (ASC), can summarize the inherent capacity of a test or a statistical model for distinguishing a diseased from a non-diseased subject across all possible levels of positive cut points into a single statistic. (Hanley and McNeil 1982; Begg, 1991; Lee and Hsiao 1996) These summarized statistics of ROC curves can be used to make inferences about the strength of the tests or statistical models. Among the different summary statistics of ROC curves, the area under a ROC curve (often referred as ROC statistics) is the most popular and widely used summary statistic for ROC curves. AUC statistics depict the probability that the value of the test result or biomarker of a randomly selected diseased subject will exceed that of a randomly selected non-diseased subject.

A considerable amount of literature has been established on the statistical methods for estimating ROC curves, calculating the area under a ROC curve and testing the significance of the area under a ROC curve (though less on this topic) for cross sectional data. (Hanley and McNeil 1983; Swets 1997; Wieand *et al.*, 1989; Tosteson and Begg 1988; Li *et al.*, 1999). Most of the major statistical software packages, such as SAS, STATA, and SPSS, have procedures that can directly generate a single ROC curve and calculate the area under a ROC curve for a cross-sectional design. However, limited research has been done on estimating and testing the area under a ROC curve for longitudinal repeated measures design, especially significance testing of the area under a ROC curve that is estimated from a repeated measures regression model (Liu and Wu, 2003). Model-based estimates of ROC curves are very useful since these methods provide the opportunities for evaluating the impact of surrounding covariates on the accuracy and potency of a test or statistical model for discrimination. However, because model-based ROC curves are constructed from correlated measurements, traditional methods for testing or comparing ROC curves are no longer valid. Using a bootstrapping re-sampling approach, this paper estimates the standard error of the area under a ROC curve estimated by generalized linear mixed model with the Wilcoxon non-parametric approach (Bamber 1975) under a repeated measures design and provides a method for conducting significance testing of the area under a ROC curve with repeated measures design.

It is worth noting that one has to be very careful when using bootstrap for non-i.i.d. settings. Efron's (1979) "resampling with replacement" method was originally proposed for the i.i.d case. For complex models with dependent data, bootstrap would work only if resampling is done *in an appropriate manner* consistent with the model. For example, in a simple linear regression model where the errors are "homoscedastic", the bootstrap could be done by *resampling the residuals with replacement*. However, in a correlation model where the errors are "heteroscedastic", *resampling the residuals with replacement* is no longer appropriate. Instead, one should resample the vectors (X_i, Y_i), $i = 1, \ldots, n$, where X_i and Y_i are the covariate and response variables for subject i; see Freedman (1981). For the repeated measures design, bootstrap would not work if one resamples the observations within a subject since the within-subject dependence structure would be obliterated. In this paper, we propose to resample the vectors as done in the correlation model (Freedman, 1981) so that the within-subject dependence is preserved. For a balanced design in which there are no missing values within each subject, the vectors from the n subjects are a random sample and thus sampling the vectors with replacement is expected to work, as shown by our simulation study. When there are missing values, we anticipate that the same resample procedure would still work under the "completely missing at random" assumption. A rigorous justification is beyond the scope of this paper since a theoretical proof of the consistency of bootstrap can be extremely technical even for the simple i.i.d. cases (see, e.g., Bickel and Freedman, 1981; Freedman, 1981; Li and Datta, 2001). We conducted a simulation study to investigate the performance of the proposed bootstrap approach for repeated measures design. We also performed an empirical study to illustrate the asymptotic normality of the estimated area under the ROC curve in repeated measures design. Details of the simulation studies are described in section 5. We have also created an *ad hoc* computer software that implements the algorithm, calculates and tests the area under a ROC curve for repeated measures designs.

2. ROC Curve and Its Standard Error for Cross Sectional Data

A ROC curve is a plot of sensitivity versus 1-specificity, where the sensitivity is defined as the probability that a test result is positive given the subject is truly diseased and specificity is defined as the probability that the test result is negative given the subject is truly non-diseased. Let $y \in (-\infty, \infty)$ denote the result of a test for a continuous outcome measure, D be an indicator of diseased/positive ($D = 1$) or non-diseased/negative ($D = 0$) status of a subject. Let c be a threshold value that any test result $y \geq c$ is considered to be diseased (positive), otherwise it is non-diseased (negative). Let $F_D(c) = Pr(y \leq c \,|\, D = 1)$

and $F_{\bar{D}}(c) = Pr(Y \leq c \,|\, D = 0)$ be the cumulative distribution function of y for diseased and non-diseased groups, then $1-F_D(c)$ and $F_{\bar{D}}(c)$ will be the sensitivity and specificity, respectively. Let $S_D(\cdot) = 1 - F_D(\cdot)$ and $S_{\bar{D}}(\cdot) = 1 - F_{\bar{D}}(\cdot)$ be the survival function of y for the two groups and $t \in (0, 1)$ be the false positive rate (1− specificity), then we can then write a ROC curve in a succinct form: $ROC(t) = S_D\{S_{\bar{D}}^{-1}(t)\}$, where varies from 0 to 1 as the corresponding threshold c varies from ∞ to $-\infty$. (Pepe 1997) Let y_D and $y_{\bar{D}}$ be the test results of a diseased/positive and non-diseased/negative subject, respectively. It has been shown that the area under a ROC curve equals the unconditional probability of correct ordering of a outcome measures between two populations, say positive and negative, (Hanley and McNeil 1982; Pepe, 2000) that is

$$\theta = \int ROC(t)dt = Pr(y_D > y_{\bar{D}}).$$

Regression modeling can evaluate the impact of covariates on the accuracy of a diagnosis test or a biomarker. When a test or biomarker is continuous with normal error, we can model it through a standard linear regression model:

$$y = \mu + \epsilon = X\beta + \epsilon$$

where $\mu = E(y)$ is the mean of y and β are the regression coefficients of X. But very often, these continuous outcomes can be dichotomized according to some critical value c. Based on the critical value, the outcome can be defined as positive if $y > c$ and negative if $y \leq c$. To model the impact of covariates on non-continuous measures, one can model functions of μ rather than μ itself; that is to use generalized linear model (GLM) framework with link functions. The basic form of GLM can be written as $\eta = g(\mu) = X\beta$, where η is a link function of μ that links $E(y)$ to the linear predictor $X\beta$. Through GLM, the impact of covariates on a diagnosis test is converted into the predicted value $\hat{y}$, which will be used to replace the observed y in constructing ROC curves and summary statistics of ROC such as the area under a ROC curve. The similar idea for calculating ROC curve extends naturally to repeated measures design, which will be discussed in details in section 3.

The precision of an estimate of the area under a ROC curve needs to be calculated to conduct a statistical significance test of the area under the ROC curve and to construct the confidence intervals of the area under a ROC curve. Due to the complexity of estimating the area under a ROC curve, it is very hard to get a closed-form solution of the area under a ROC curve. As a result, there is no available close-form solution of the standard error $\sqrt{var(\hat{\theta})}$ of the area under a ROC curve. In the situation where the observations used to estimate the ROC curve are independent, Dorfman (1969) and Wieand *et al.* (1989) derived a formula for calculating the standard error of the area under a ROC curve that is related to Wilcoxon non-parametric estimate. This method has been used for estimating the standard error of the area under a ROC curve and is implemented in different statistical software such as SPSS and EPISTAT. Unfortunately, model-based estimate of the area under a ROC curve is calculated from the predicted values that are no longer independent regardless of the independent or dependent status of the original observations, and the Dorfman method is no longer valid. The dependence among the predicted values complicates the calculation of the standard error of the area under a ROC curve and needs to be taken into account in the estimation.

3. Estimating ROC Curve under a Repeated Measures Design

Data collected from a repeated measures design have several advantages over data from a cross-sectional design. First, repeated measures design can reduce the possible bias from one snapshot of data collection from each subject and increase the reliability of the data collected. Second, repeated measures design costs less to collect the same amount of data (number of observations) with fewer number of subjects compared to cross-sectional design that each observation is a different subject (collecting additional data points from an existing subject likely costs less compared to collecting data from recruiting additional subjects.) Third, since each patient has multiple observations, repeated measures design provides the opportunity for us to analyze the intra-patient variation as well as the change over time of the entire cohort. Under a repeated measures design, the observations within a given subject will no longer be independent and intra-subject correlation and variation are introduced. Therefore, the impact of confounding covariates on the accuracy of a diagnostic test/bio-marker or a statistical model will come from both global fixed effect (e.g., patient's race/ethnicity) as well as individual patient random effects (e.g., change over time of a time varying covariate.) To model outcome variables that could be continuous or non-continuous, the random effects can be taken into account through the extension of generalized linear model (GLM) to the generalized linear mixed model (GLMM) (Bresloe and Clayton 1993), in which the linear predictor is composed of two parts of fixed and random effects. Define $\mu_i = E(y_i \,|\, \gamma_i)$ as the conditional mean of an outcome variable y_i and let $\eta_i = g(\mu)$ be the link function that connects μ_i with the linear predictor that consists of both fixed and random effects. We can then write GLMM as $\eta_i = X_i\beta + Z_i\gamma_i$ (for $i = 1, \ldots, n.$) with a conditional variance $Var(y_i \,|\, \gamma_i)$, where y_i is a $n \times 1$ vector of a test results for the i th subject, n_i is the number of outcome measures for the i th patient, and X_i is $n_i \times p$, which contains known covariates that are associated with the fixed effects. β, the fixed effect parameter vector, is $p \times 1$, and Z_i is $n_i \times k$ representing known covariates that are associated with the random part of the model. γ_i, the random effect parameter vector, is

$k \times 1$ and is distributed as $N(0, D_i)$.

Now assume that the outcome is actually binary (diseased/non-diseased) and let $p_{ij}, (i = 1, \ldots, ; j = 1, \ldots, n_i)$ be the probability of being diseased/positive for the ith subject at the jth time point; η_{ij} be logit link function for the ith subject at the j th time point between the mean and the linear predictor, then we can model the impact of covariates on the predicted probability of being diseased/positive through GLMM:

$$\eta_{ij} = x_{ij}\beta + z_{ij}\gamma_i$$

and

$$\eta_{ij} = g(p_{ij}) = \log(p_{ij}/(1 + p_{ij}))$$

or

$$\log(p_{ij}/(1 + p_{ij})) = x_{ij}\beta + z_{ij}\gamma_i$$

Let $\hat{\beta}$ and $\hat{\gamma}_i$ be the estimates through penalized quasi-likelihood (Breslow and Clayton, 1993) or restricted pseudo-likelihood (Wolfinger and O'Connell, 1993) and $\hat{p}_{ij}$ be the corresponding estimate of p_{ij}. Then we have

$$\hat{p}_{ij} = \frac{\exp(x_{ij}\hat{\beta} + z_{ij}\hat{\gamma}_i)}{1 + \exp(x_{ij}\hat{\beta} + z_{ij}\hat{\gamma}_i)}.$$

The estimated probability $\hat{p}_{ij}$ ($i = 1, \ldots, n$ and $j = 1, \ldots, n$), which is a function of all the covariates, will then serve as a biomarker for constructing the ROC curve for discriminating a diseased/positive subject from a non-diseased/negative subject longitudinally.

Let $roc_{x,z}(t)$ be the ROC value with false-positive rate t that is associated with the fixed effect predictors x and random effects predictors z. By definition, the area under a ROC curve θ is:

$$\theta = \int roc_{x,z}(t)dt,$$

where the integration limits run from 0 to 1. The area under a ROC curve θ can be calculated using the Wilcoxon non-parametric method by comparing the magnitude of the predicted probabilities of each discordant pair. In repeated measures design, each subject has more than one observation and the outcome values may vary from time to time. Therefore, the classification of a diseased/positive and a non-diseased/negative case needs to be sorted at the observational level rather than at the individual subject level. Let $\hat{p}_{ij(D)}$ ($i = 1, \ldots, n$ and $j = 1, \ldots, s_i$) be the predicted probability of a disease/positivity for the ith subject at the j th time point that had a diseased (positive) observed value, and let $\hat{p}_{k\ell(\bar{D})}(k = 1, \ldots, n$ and $\ell = 1, \ldots, t_k)$ be the predicted probability of a disease/positivity for the k th patient at the ℓ th time point that had a non-diseased (negative) observed value. Let $N_D = \sum_{i=1}^{n} s_i$ and $N_{\bar{D}} = \sum_{k=1}^{n} t_k$ be the total number of observations with positive or negative observed values. The total number of discordant pairs then equals $N = N_D * N_{\bar{D}}$. The area under a ROC curve can be calculated by comparing the predicted probabilities of each discordant pair that is defined at observation level (Liu and Wu 2003). Let $A(\cdot)$ be an indicator that:

$$\begin{aligned} A(\bar{p}_{ij(D)}) > \bar{p}_{k\ell(\bar{D})} &= 1 \quad \text{if} \;\; \hat{p}_{ij(D)} > \hat{p}_{ij(\bar{D})} \\ &= 0 \quad \text{if} \;\; \hat{p}_{ij(D)} < \hat{p}_{ij(\bar{D})} \end{aligned}$$

and

$$\begin{aligned} A(\bar{p}_{ij(D)} = \bar{p}_{k\ell(\bar{D})}) &= 1 \quad \text{if} \;\; \hat{p}_{ij(D)} = \hat{p}_{ij(\bar{D})}, \\ &= 0 \quad \text{otherwise} \end{aligned}$$

the estimate of area under a ROC curve θ is then estimated by the ratio:

$$\theta = \frac{\sum_i \sum_j \sum_k \sum_\ell [(A(\hat{p}_{ij(D)} > \hat{p}_{k\ell(\bar{D})})) + \frac{1}{2} A(\hat{p}_{ij(D)} = \hat{p}_{k\ell(\bar{D})})]}{N_D N_{\bar{D}}}.$$

To generate the actual ROC curve, a series of pairs of sensitivity and 1-specificity based on predictions from the generalized linear mixed models are calculated. In order to obtain a smooth curve, an increment of 0.005 in predicted probability for defining positivity is used. That is, 200 pairs of sensitivity and 1-specificity will be calculated to generate the ROC curve. Assume that cut points of positivity are $c(1), c(2), \ldots, c(200)$, then for any $c(i)$ with $(1 \leq i \leq 500)$, the sensitivity and specificity will be calculated and the ROC curve will be plotted.

4. Estimate Standard Errors of Area under ROC Curves through Bootstrapping

The variance of the area under a ROC curve $var(\theta) = var(\int roc_{x,z}(t)dt)$ is essential for statistical testing of the summary statistic of a ROC curve and it measures the reliability of an estimate of the area under a ROC curve. Given the nature of the complexity in estimating the area under a ROC curve, it is hard to get closed-form solution of the variance. If the area under a ROC curve is estimated by Wilcoxon non-parametric method and the observations are independent, one could estimate the variance of the area under a ROC curve by the following equation (Hanley and McNeil 1982):

$$var(\hat{\theta}) = \frac{\hat{\theta}(1-\hat{\theta}) + (n_D - 1)(q_1 - \hat{\theta}^2) + (n_{\bar{D}} - 1)(q_2 - \hat{\theta}^2)}{n_D n_{\bar{D}}},$$

where q_1 is the probability that two randomly chosen positive observations will both be ranked with greater suspicion than a randomly chosen negative observation. q_2 is the probability that one randomly chosen positive observations will be ranked with greater suspicion than two randomly chosen negative observations. n_D is the number of observations of the positive group and $n_{\bar{D}}$ is the number of observations of the negative group.

Unfortunately, due to the common coefficients from a regression model used for calculating predicted values, model-based estimates of predicted probabilities are always non-independent. When the predicted probabilities are estimated from a repeated measures model, this dependence will be even stronger since in this case not only the common coefficients but also the intra-subject correlation generate dependence among predicted probabilities. Therefore, the above formula is no longer valid. The dependence among the observations or discordant pairs cannot be ignored and needs to be taken into account in the calculation for estimating the variance of a ROC curve. To get an estimate of the standard error of the area under a ROC curve, we propose a bootstrapping method to estimate the variance of the area under a ROC curve that is calculated by the Wilcoxon non-parametric method described in section 3. (Efron 1979; Efron, Tibshirani 1986). Assume that there are a total of n subjects, the i th subject has n_i observations (for $1 \leq i \leq n$) and there are a total of $N = \sum_{i=1}^{n} n_i$ observations in the data set. In order to preserve the original intra-patient variation and data structure, the bootstrapping re-sampling algorithm is designed to sample data at subject level rather than at observation level. That is from the n subject pool, a random sample of n subjects is drawn with replacement. For any given subject drawn from the pool, say subject k, all n_k observations that belong to this subject in the original data set will be automatically included in the bootstrapped sample. Therefore, a bootstrapped sample will still consist of n subjects but with only s unique subjects, where $a \leq n$ (the equal sign holds if and only if when the bootstrapped sample is identical to the original data set.) A bootstrapped data set will consist of $\tilde{N}$ observations with $\tilde{N} = \sum_{i=1}^{n} s_i$ where s_i is the number of observations of the ith subject drawn. The $\tilde{N}$ usually does not equal to N, the original number of observations.

For each bootstrapping sample, the statistic of area under the ROC curve is estimated using the Wilcoxon non-parametric algorithm (Liu and Wu 2003). Suppose that r bootstrapping samples are generated, then r statistics of the area under the ROC curves $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_r$ will be estimated through the Wilcoxon non-parametric method. Based on the series of estimated area under the curve values, the standard error of the originally estimated area under the curve will be estimated as:

$$S.E._{\hat{\theta}} = \sqrt{\frac{\sum_{i=1}^{r}(\hat{\theta}_i - \bar{\hat{\theta}})^2}{r-1}}$$

where $\bar{\hat{\theta}} = \sum_{i=1}^{r} \hat{\theta}_i / r$ is the mean of the r estimated areas under the ROC curves. The stability and accuracy of $S.E._{\hat{\beta}}$ are determined by the estimates of $\hat{\theta}$'s and the number of repeated bootstrapping samples r. The larger the number of bootstrapping samples is, the better estimate the standard error is. The standard error $S.E._{\hat{\beta}}$ approaches stable when the number of replicate r is gets large. In Section 7 through an example, we show the stabilization of $S.E._{\hat{\beta}}$ as a function of number of replicate r.

5. Simulation Validation of Normality of ROC Statistics and Bootstrap Approach under Repeated Measure Design

To evaluate the validity of the statistical test of ROC statistics with normal approach and the bootstrapping algorithm under repeated measure design with non-independent observations, we conducted a simulation analysis to assess two issues: (a) how ROC statistics are distributed (b) how bootstrapped and true parameter estimates are close to each other. Through this simulation analysis, we empirically proved the validity of our approach in testing ROC statistics and using bootstrap approach with repeated measures data. The simulation analysis was performed in two steps. Step one was to show that the ROC statistics calculated from simulated data sets follows a normal distribution and to get the estimate of "true" standard error of ROC statistics of the simulated data sets; step two demonstrated that the standard error of ROC statistics calculated through bootstrap was a consistent estimate of the "true" one.

We used a balanced repeated measures design with 200 subjects and each subject had 4 repeated measures observations. For the i th subject ($i = 1, \ldots, 200$) and j th observation ($j = 1, \ldots, 4$), the repeated measures model can be written as:

$$Y_{ij} = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + Z_{ij1}\gamma_{i1} + A_{ij2}\gamma_{i2} + \epsilon_{ij}$$

where x_{i1} and x_{i2} were generated from a standard normal and Bernoulli distribution, respectively. We took $\beta_0 = 10, \beta_1 = 3, \beta_2 = -4$, and

$$\begin{pmatrix} \gamma_{11} \\ \gamma_{12} \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} 4 & 3.2 \\ 3.2 & 11.56 \end{pmatrix}\right),$$

$\epsilon_{ij} \sim N(0, 4)$, and $cov(\gamma, \epsilon) = 0$. With the above set-up, we have generated a simulated data set with 200 subjects and each subject had 4 repeated measure observations. Using this simulated data set, an estimate of area under the ROC curve was calculated and recorded. Then another data set with the same set up was generated and its corresponding area under the ROC curve was calculated again. This procedure was repeated 200 times and 200 area under ROC curves (or 200 ROC statistics) were calculated. The normality of the ROC statistics was tested through QQ-plot (Harter, H. L., 1984; Evans, M. *et al.*, 2000) and Shapiro-Wilk test. To get the QQ-plot, first the 200 ROC statistics were sorted and then a measure was calculated as function of number of non-missing observations and the order i:

$$\nu_i = [(i - 0.375)/(n + 0.25)] \quad \text{for} \quad i = 1, 2 \ldots, n$$

The i th ordered observation is plotted against the normal quantile $\Phi^{-1}(\nu_i)$, where $\Phi^{-1}(\nu_i)$ is the inverse standard cumulative normal distribution. If the data are normally distributed with mean μ and standard deviation σ, the points on the plot should lie approximately on a straight line with intercept μ and slope σ. The resulting Q-Q plot from the 200 simulated ROC statistics showed approximately a straight line as in Figure 1.

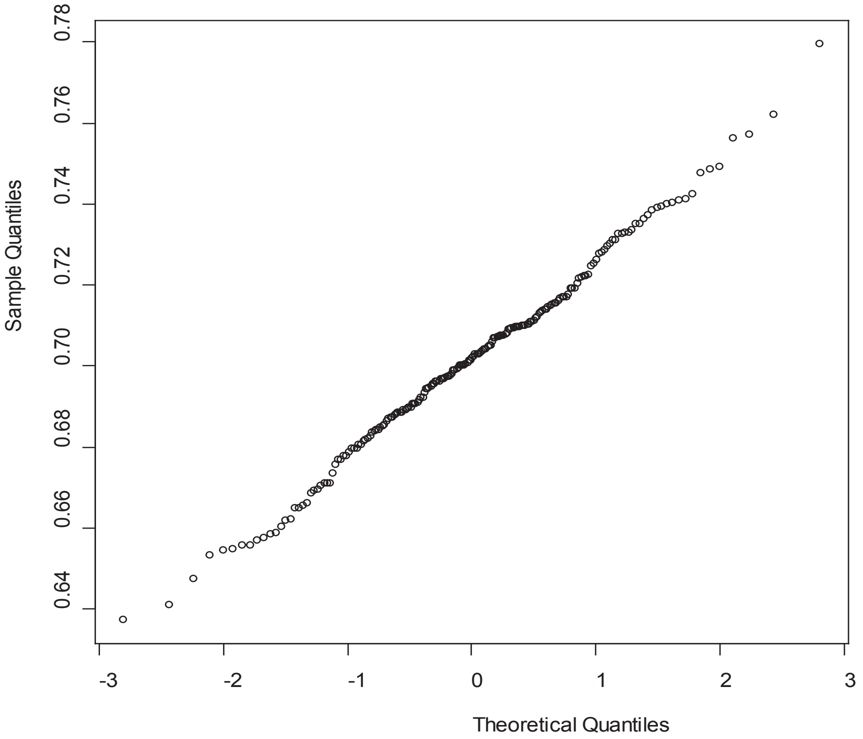

Figure 1: Normal QQ plot of ROC

The normality of the 200 simulated ROC statistics was also tested using Shapiro-Wilk test with the null hypothesis that z-statistics is distributed as $N(0,1)$. The test results yielded a statistic of 0.995 and a p value of 0.7906 for the null hypothesis that the ROC statistic is normally distributed.

With these 200 simulated ROC statistics, we have also estimated the 'true' standard error of ROC statistics, which was 0.0245381. This "true" standard error was used as the anchor to compare with the bootstrapped standard error in the next step.

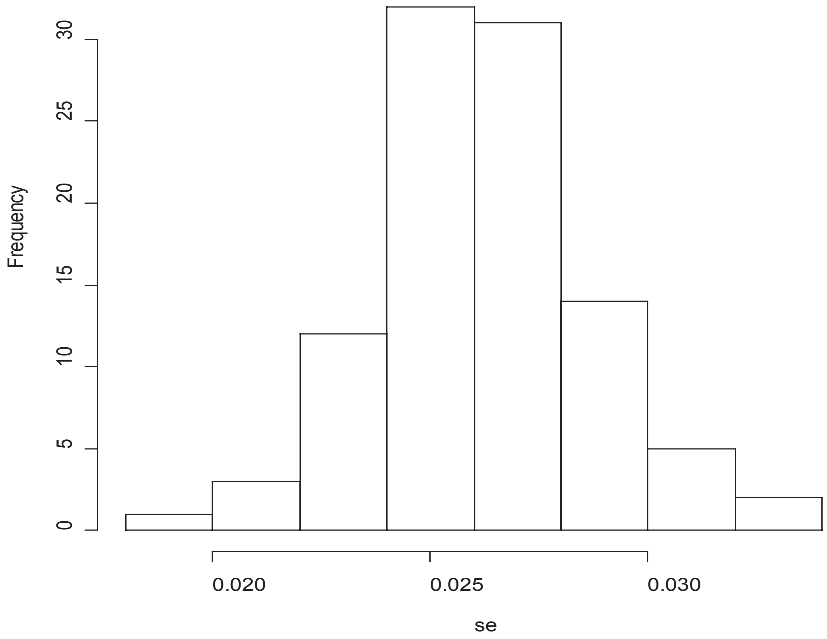

Figure 2: Histogram of se

To show that the bootstrapped standard deviation is a consistent estimator of the true one (0.0245381), we performed bootstrap with 200 replicates for a given simulated data set. The bootstrapping draw was with replacement at subject level. Once a subject was drawn, all 4 repeated measures of the subject were automatically included in the bootstrapped data set. Each bootstrapped sample had 200 subjects and each subject had 4 repeated measures observations. With the 200 bootstrapped samples from a simulated data set, 200 bootstrapped ROC statistics were calculated and an estimate of bootstrapped standard error of the ROC statistics was obtained. This whole procedure was repeated on 100 out of the

200 simulated data sets in step one. Thus, we have generated 100 bootstrapped standard errors of ROC statistics. Figure 2 is the histogram of these 100 standard errors of ROC statistics. We can see that they are distributed surrounding the "true" standard error of 0.0245381 with a standard error of 0.00252109. The mean of the 100 bootstrapped standard error is 0.0261701, which is very closed to the true one with a difference as small as 0.001632. We have also conducted the simulation analysis with a unbalanced design with number of repeated measures varying from 3 to 6 and obtained very similar results.

6. Statistical Significant Test and Confidence Intervals

To test the area under a ROC curve θ equal to a specific value c (e.g., $c = 0.5$), one can base the test on the asymptotic distribution of $\hat{\theta}$. According to large sample asymptotic distribution theory (Furguson 1967), when r, the number of replicates goes to infinity, we have the following statistic distributed approximately as a standard normal distribution with a mean of 0 and variance of 1:

$$z_{\hat{\beta}} = \frac{\hat{\theta} - c}{S.E >_{\hat{\beta}}} \sim N(0,1).$$

The null hypothesis $H_0 : \ \theta = c$ will be accepted if ${}_{\hat{\beta}} \leq z_{1-\alpha/2}$ or rejected if ${}_{\hat{\beta}} > z_{1-\alpha/2}$ at the type I error level of α. The $(1-k)\%$ (with $0 < k < 1$) level of confidence interval of the estimated area under the ROC curve $\hat{\theta}$ can be calculated by normal approximation:

$$\hat{\theta} \pm z_{1-k/2}\ S.E._{\hat{\beta}}.$$

%r_roc Software

In this section, we introduce a newly created *ad hoc* statistical software %rm_roc (standards for repeated measures ROC) that can automatically conduct bootstrapping, calculating the standard error of the area under a ROC curve and other statistics. The functionality of the software, the algorithm used, the input parameters, the output contents and the usage will be described. The function of this software is to conduct ROC curve analysis for repeated measures data. Taking input data with a repeated measures design, this software can automatically fit a ROC curve based on Wilcoxon non-parametric approach, calculate the

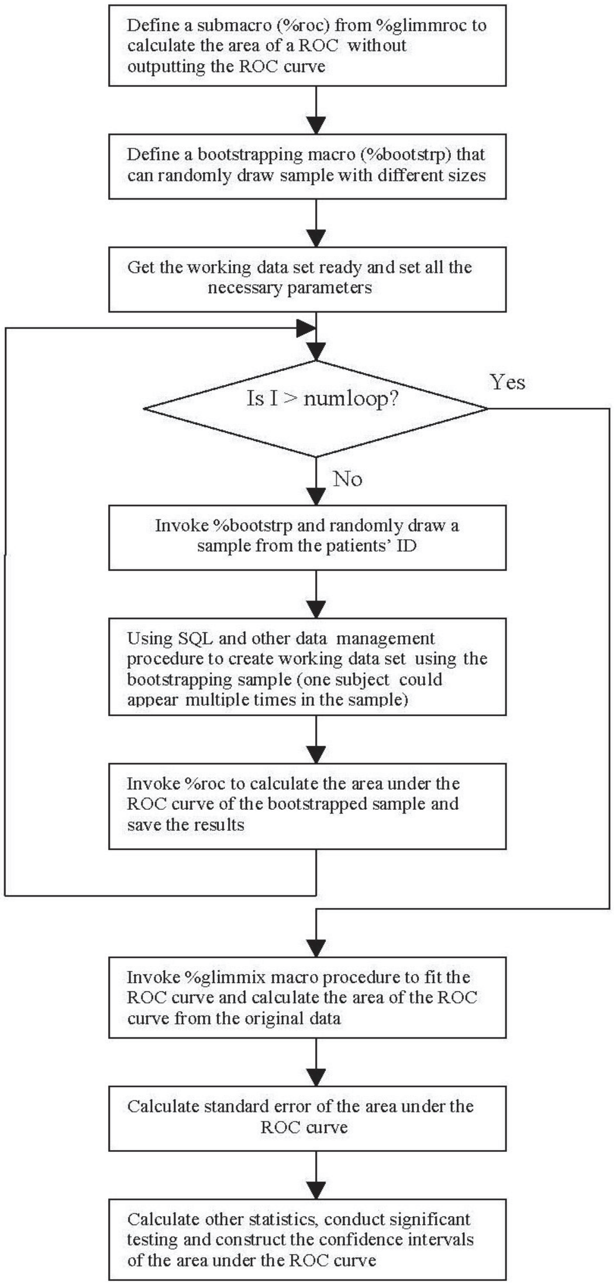
Define a submacro (%roc) from %glimmroc to calculate the area of a ROC without outputting the ROC curve
Define a bootstrapping macro (%bootstrp) that can randomly draw sample with different sizes
Get the working data set ready and set all the necessary parameters
Is I > numloop?
Yes
No
Invoke %bootstrp and randomly draw a sample from the patients' ID
Using SQL and other data management procedure to create working data set using the bootstrapping sample (one subject could appear multiple times in the sample)
Invoke %roc to calculate the area under the ROC curve of the bootstrapped sample and save the results
Invoke %glimmix macro procedure to fit the ROC curve and calculate the area of the ROC curve from the original data
Calculate standard error of the area under the ROC curve
Calculate other statistics, conduct significant testing and construct the confidence intervals of the area under the ROC curve

Figure 3: Foow chart of algorithm.

area under the ROC curve, conduct bootstrapping sampling and estimate the standard error of the area under the ROC curve, perform a significance test of the area under the curve and construct a confidence interval of the area under the curve.

%rm_roc is written in SAS/Macro (SAS Institute Inc. 2001) and SAS//IML (SAS Institute Inc. 2001). It is created based on two previously existing macro procedures %glimm from SAS (http://ftp.sas.com/techsup/download/stat/) and %glimmroc (Liu and Wu 2003). The bold steps of the algorithm are: (a) draw a bootstrapping sample from the original data pool. (b) with the bootstrapped data, fit a generalized linear mixed model and calculate the area under the ROC curve. (c) repeat (a) and (b) until the desired number of estimates of the area under the curves have all been calculated. (d) using the original data, fit the ROC curve and estimate the area under the ROC curve (e) using the estimated areas under the ROC curves from bootstrapping samples to estimate the standard error of the area under the curve of the original data. (f) conduct statistical significant testing and construct the confidence interval of the area under the curve (see Figure 3: Flowchart of the Algorithm.)

The macro software has been designed with an easy and user-friendly syntax so that those with basic SAS software literacy can understand and use the software. There are total of 12 parameters that need to be entered (some can be left as default) in the input statement. For those parameters that one wants to leave as the default value, these parameters should be left out from the list of input statement. The input parameters include:

- the dependent variable
- the list of the independent variables that are associated with the fixed effect
- the independent variables that are associated with the random effects
- the subject identifier, the type of the covariance structure for the random part
- the type of the covariance structure for the error part, weight option
- the test value of the area under the curve
- the number of subjects sampled for each bootstrapped sample
- the number of replicates, the level of confidence intervals
- the name of the data set you are going to use for the analysis.

The syntax of the input is:

```
%RM_ROC(y=,x_list=,z_list=,id=,c_s_d=,c_s_r=,
        weight=,c=,numsamp=,numloop=,ci=,dataset=);
```

The following explain each of the parameters in the syntax:

```
Y       the variable name of the binary outcome measure

x_list  contains the list of all independent variables for the fixed
        effects with space in between(e.g., age sex race)

z_list  contains the list of all independent variables for the random
        effects with space in between (e.g., time)

ID      the variable of patient ID which identifies observations within
        a patient

c_s_d   specify the covariance structure of matrix D (the random part
        and the default is simple format, e.g., diagonal matrix.)

c_s_r   specify the covariance structure of matrix R (the random error,
        the default is compound symmetry)

weight  weight variable (the default is 1, which means unweighted)

c       the constant that the area under the curve is tested against it.
        The default value is 0.5.

numsamp the number of subject sampled for each bootstrapped sample
        (numsamp=0 is the default, which will draw a sample with the
        original number of subject).

Numloop the number bootstrapping sample replicates.

CI      confidence interval level. The options are 90, 95  and 99.
        The default value is 95.

dataset the name of the data set on which one will run the software.
```

The output of the software includes the actual ROC curve, the estimate of the area under the ROC curve, the standard error of the area under the ROC curve, the z_statistic, the p_value of the test and the confidence interval of the area under the ROC curve.

The usage of the software is simple. One just needs to either copy the macro procedure into his SAS program or use "%include" statement at the beginning of a SAS program to load in the procedure to avoid the lengthy display of the code. Since this macro is built on top of the other existing macros, one also needs to load the macro procedure of %glimm and %glimmroc in their program.

The entire software package can be provided upon request from the author at hhliu@ucla.edu.

7. Application

In this section, we show how the method and the statistical software work in an example with a repeated measures design. Through this example, we will also show empirically how large the number of replicates r needs to be in order to obtain a stable estimate of the standard error of the area under a ROC curve. To illustrate this, the standard error of the area under the ROC curve for this example was estimated with different number of replicates r, ranging from 10 to 300 in increments of 10 to display the stability of the estimates as a function of the number of replicates.

The data in this example concerns measurement of HIV patient adherence to antiretroviral medications. Adherence to antiretroviral medication is critical in suppressing viral replication and preventing drug-resistant strains. Due to the complexity of measuring patient adherence behavior, different measurement tools and mechanisms have been developed, each with different inherent strengths and weaknesses. The Medication Event Monitoring system (MEMS) and Pill Count (PC) are two popular measuring methods (**** APREX 1998****; Grymonpre, *et al.* 1998). MEMS is a relatively new technique that utilizes a pill bottle cap containing a microchip and records each instance of bottle opening. PC is a measurement method that physically calculates the number of pills remaining in a patient's bottle or bottles at a visit by a person (normally a nurse). Evidence has shown that MEMS adherence is more objective and accurate than PC. Although not costly, PC is likely overestimates a patient's true adherence level due to reasons such as pill dumping. Even though MEMS is more objective, it is not always practical to implement. To evaluate how good PC can measure adherence over time (using MEMS as a gold standard for medication adherence), MEMS and PC data along with patient's baseline information of gender, age and lowest CD4 count, were collected for 140 HIV+ patients at every 4-week period (a "wave") for 48 weeks. A threshold of 85% is used to classify patient adherence behavior at each of the 12 waves as either "adherent" (patient took at least 85% of the prescribed doses for that wave), or "non-adherent" (the patient took less than

Table 1: Estimated standard errors from bootstrapping

Number of Replicates	Standard Error	Percent Change in S.E.	CPU Time (minutes)
10	0.037502	–	3.268
20	0.035045	−6.6	5.724
30	0.033183	−5.3	8.107
40	0.052684	58.8	10.815
50	0.048887	−7.2	13.188
60	0.054463	11.4	15.667
70	0.059754	9.7	18.252
80	0.062297	4.3	20.782
90	0.063003	1.1	23.375
100	0.064216	1.9	26.020
110	0.064493	0.43	28.515
120	0.062196	−3.6	31.067
130	0.060695	−1.6	33.693
140	0.059361	−2.2	36.140
150	0.061163	2.6	38.657
160	0.059860	−2.1	41.284
170	0.058438	−2.4	44.155
180	0.058688	0.44	46.694
190	0.057562	−1.9	49.297
200	0.056323	−2.2	51.931
210	0.055234	−1.9	54.328
220	0.054121	−2.0	57.419
230	0.053253	−1.6	59.965
240	0.053379	0.24	62.746
250	0.054409	2.17	64.092
260	0.053908	−0.92	66.603
270	0.053120	−1.46	69.164
280	0.052362	−1.43	71.641
290	0.053065	1.34	74.218
300	0.054544	2.79	76.799

85% of the prescribed medication for that wave). The 140 patients have an average of 8.76 repeated measures data points and a total of 1226 observations.

The data set is named ANAL and the variables are the following:

`bi_mems`	`medication adherence measured by MEMS (binary measure)`
`gender`	`gender of a patient`

age	age of a patient
lowestcd	lowest CD4 count at baseline
bi_p	medication adherence measured by PC (binary measure)
id	patient ID that identifies observations within a patient
wave	time of the 4-week period (ranges 1 to 12)

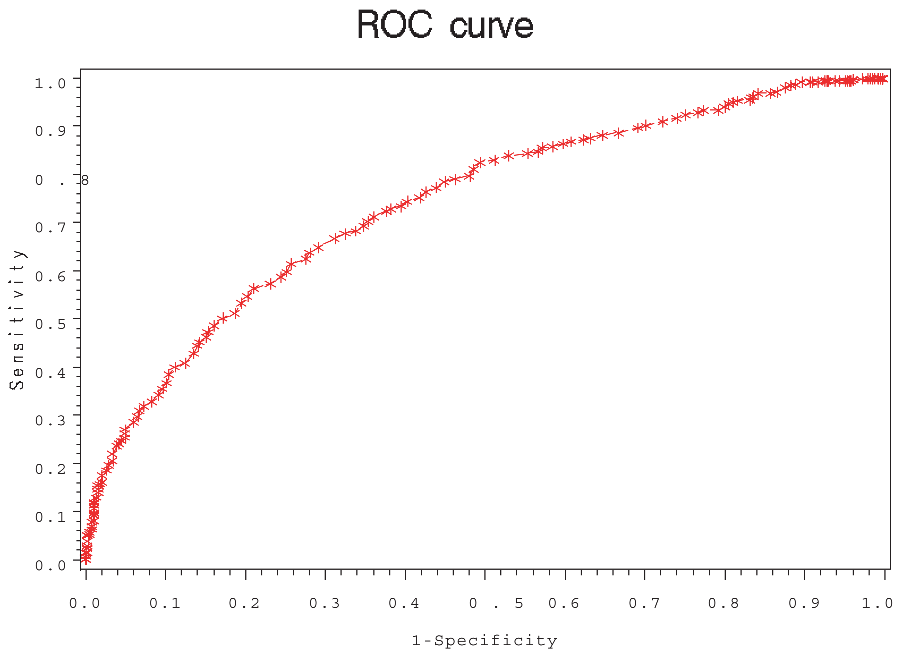

Figure 4: The fitted ROC curve.

For this data, we used 30 different sizes of the number of bootstrapping samples in estimating the standard error of the area under the curve, starting at 10 and increased by 10 up to 300. The following table shows the actual estimated standard errors, the percent changes of the estimated standard error, the CPU time used on a UNIX RS/6000 Cluster computer (see Table 1.) Based on the results, we can see that when the number of replicates increases, the estimates of the standard errors (ranging from 0.033183 to 0.064493) become stable. The percent changes of the estimates, ranging from 0.24 to 0.588, are quite large when the number of replicates is small. The CPU time used is a monotonous function of the number of bootstrapping sample replicates, ranging from 3.268 to 76.799 minutes. Overall, we can see that the estimated standard errors of the area under the curve become quite stable when the number of replicates is above 200. With the estimates of number of replicates above 200, the percent changes range from 0.24 to 2.79 and the estimates range from 0.052362 to 0.055234.

The fitted ROC curve is shown in Figure 4 below and the estimated area under the curve is 0.78. Since the area under a ROC curve is the unconditional probability of correct ordering, this statistic indicates that if we randomly select a patient who is adherent and randomly select another patient who is not adherent, then the probability of being adherent predicted by pill count whiling controlling patient characteristics for the patient who is adherent is greater than that of the patient who is not adherent will be 0.78. To test the significance of the area under the curve from 0.5, using the estimated standard error 0.055234 from the 210 replicates, we obtain a t-statistic of 5.02, which yields a significant p value of less than 0.0001 with a type I error of 0.05. This means that the discriminating power of the bio-maker of the predicted probability for adherence by pill count from the repeated measures model is significantly larger than that of chance alone. The 95% confidence interval of the estimated area under the curve is (0.66928, 0.88580).

8. Discussion

Repeated measures model-based estimate of ROC curves provides the opportunity to evaluate the impact of both fixed effect and the random effects on a test or bio-marker. Because of the common parameter estimates from the regression model, estimates of the area under a ROC curve are calculated from predicted values that are not independent. This is particularly true for repeated measures design since the repeated observations within a subject themselves are also correlated each other. Due to the involvement of both fixed and random effects in generalized linear mixed models, the statistical equation of the areas under a ROC curve is complicated and its standard error cannot be expressed in simple closed-form solutions. Bootstrapping from the original data will allow one to obtain estimates of the areas under ROC curves that reflect the sampling variation of the collected data. Sampling at the subject level rather than the observation level will enable us to preserve the within subject dependence and the intra-subject structure of the observed data. The simulation analyses in section 5 demonstrates that the area under ROC curves can be tested through normal theory and the bootstrap algorithm used in the paper for repeated measures design can generate valid estimates. These simulation results could bear important implications in other areas of research under repeated measures design. Although the proposed approach in this paper is computationally intensive, the dramatic advancement of computer technology with increasing high speed processing has made statistical computation less burdensome. Based on the example in this paper, a moderate large number of replicates (around 200) of bootstrapped samples will yield fairly stable estimates of the standard errors. The computer software %glimmroc, which has the special functionality that no other software contains, has a user-friendly syntax and can be easily applied to conduct a full ROC analysis, ranging from fitting the ROC curve to testing the significance and constructing the confidence intervals of the area under the curve, for data with a repeated measures design.

Acknowledgement

We wish to thank Victor Gonzalez for his technical assistance in preparing the manuscript.

References

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating graph. *J. Math. Psych.* **12**, 387-415.

Begg, C. B. (1991). Advances in statistical methodology for diagnostic medicine in the 1980's. *Statistics in Medicine* **10**, 1887-1895.

Bickel, P. J. and Freeman, D. A. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics* **9**, 1196-1217.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9-25.

Dorfman, D. D., Alf, E.(1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating-method data. *J. Math. Psych.* **6**, 487-496.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**, 1-26.

Efron, B. and Tibshirani, R. (1986). Bootstrap measures for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* **1**, 54-77.

Evans, M., Hastings, N., and Peacock, B. (2000). *Statistical Distributions*, 3rd ed. Wiley.

Ferguson, T. S. (1967). *Probability and Mathematical Statistics.* Academic Press.

Freedman, D. A. (1981). Bootstrapping regression models. *Annals of Statistics* **9**, 1218-1228.

Green, D. and Swets, J. (1966). *Signal Detection Theory and Psychophysics.* John Wiley and Sons.

Grymonpre, R. E., Didur, C. D., Montgomery, P. R., Sitar, D. S. (1998). Pill count, self-report, and pharmacy claims data to measure medication adherence in the elderly. *Ann. Pharmacother* **32**, 749-54.

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

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

Harter, H. L. (1984). Another Look at Plotting Positions. *Comm. Stat., Theory and Methods* **13**, 1613-1633.

Lee, W. C. and Hsiao, C. K. (1996). Alternative summary indices for the receiver operating characteristic curve. *Epidemiology* **7**, 605-611.

Li, G. and Datta, S. (2001). A bootstrap approach to nonparametric regression for right censored data. *Annals of Institute of Statistical Mathematics* **53**, 708-729

Li, G., Tiwari, R. C., Wells, M. T. (1999). Semiparametric inference for a quantile comparison function with applications to receiver operating characteristic curves. *Biometrika* **86**, 487-502.

Liu, H. H. and Wu, T.T. (2003). Estimating the area under a receiver operating characteristic (ROC) curve for repeated measures design. *Journal of Statistical Software* **12**, 1-18.

MEMS ViewTM, User's Guide (Version 2.61). (1998). APREX*, Intelligent Technology for Medication Management.

Metz, C. (1986). ROC methodology in radiologic imaging. *Investigative Radiology* **21**, 720-733.

Pepe, M. S. (1997). A regression modeling framework for receiver operating characteristics curves in medical diagnostic testing. *Biometrika* **84**, 595-608.

Pepe, M. S. (2000). In interpretation for the ROC curve and inference using GLM procedures. *Biometrics* **56**, 352-359.

Swets, J. A. (1997). ROC analysis applied to the evaluation of medical imaging techniques. *Invest Radiol.* **14**, 109-121.

Tosteson, A. N. A. and Begg, C. B. (1988). A general regression methodology for ROC curve estimation. *Med. Decis. Making* **8**, 204-215.

Wieand, H. S., Gail, M. H., Barray, R. J., James, K. L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* **67**, 585-592.

Wolfinger, R. and O'Connell, M. (1993), Generalized linear models: a pseudo-likelihood approach. *J. Statist. Comput. Simul.* **48**, 233-243.

Zou, K. H. (2001). Comparisons of correlated receiver operating characteristic curves derived from repeated diagnostic test data. *Acad. Radiol* **8**, 225-233.

Received May 6, 2004; accepted June 16, 2004.

Honghu Liu
UCLA Department of Medicine
Division of General Internal Medicine
and Health Services Research
911 Broxton Plaza
Los Angeles, CA 90095-1736
hhliu@ucla.edu

Gang Li
650 Charles E. Young Drive
Department of Biostatistics
School of Public Health, UCLA
Los Angelse, CA 90095-1772, USA

William G. Cumberland
650 Charles E. Young Drive
Department of Biostatistics
School of Public Health, UCLA
Los Angelse, CA 90095-1772, USA

Tongtong Wu
650 Charles E. Young Drive
Department of Biostatistics
School of Public Health, UCLA
Los Angelse, CA 90095-1772, USA