Comparing the Areas under Receiver Operating Characteristic Curves for Cluster Designs

William W. Davis and Katherine M. Flegal
National Cancer Institute and National Center for Health Statistics

Key words: cluster sample, statistical weight, mixed model, jackknife, balanced repeated replication

Abstract

We provide a model-based approach to test whether the difference between two receiver operating characteristic (ROC) curves is statistically significant when subjects are selected from a cluster sample design with unequal selection probabilities. The variance of the difference is expressed in terms of parameters of a mixed-model. The procedure is applied to detecting undiagnosed diabetes using data from the third National Health and Nutrition Examination Survey (NHANES III). The method is shown to yield similar conclusions to two approximate tests that utilize replication methods. Modeling the diagnostic measurements yields insight that is not captured by the two approximate hypothesis tests.

1. Introduction

When a diagnostic test is based on an observed variable that lies on a continuous scale, the receiver operating characteristic curve (ROC) can be used to assess the overall value of the new test relative to a standard diagnostic test. The ROC curve is given by varying the cut point used to determine whether the subject's measurement is considered normal. The ROC curve is then obtained by plotting the sensitivity versus one minus the specificity for each cut point. For statistical analysis, the area under the curve (AUC) is used as an index of accuracy [1]. The population ROC AUC is the probability that when the variable is observed for a randomly selected subject from the diseased population and a randomly selected subject from the normal population the resulting values will be in the correct order. Thus, if subjects from the diseased population tend to have high values on the test, the ROC AUC is the probability that a randomly selected subject from the diseased population will have a higher value than a randomly selected subject from the normal population. If a test can discriminate perfectly, it will have an AUC of 1.0 while a test with no diagnostic capability has an AUC of 0.5.

Maximum likelihood estimates of the AUC and model parameters have been developed for a number of parametric models [2-4]. The empirical ROC AUC has been shown to be equal to the Mann-Whitney U-statistic for comparing distributions of values from the two independent samples [5].

We assume that the data arose from a design with cluster sampling. Most national surveys are of this form with unequal selection probabilities used to guarantee desired domain accuracy and cluster sampling used to reduce survey cost. With this design, measurements on subjects within a cluster are correlated since individuals residing in a small area may be similar; thus, the intra-cluster correlation must be considered in the analysis. Obuchowski [6] provided an approach to estimation of the ROC AUC that for cluster sampling. However, this approach does not allow unequal selection probabilities, which is the case in national surveys.

In Section 2, we introduce notation for the ROC AUC for both independent and cluster samples, define a population model for the sample values obtained from a cluster sample, derive parameter estimates and a significance test of the difference. Section 3 provides a description of approximate variance estimation for complex survey data using two replication methods. In Section 4, we apply the model-based and replication methods to test for the equality of two alternative diagnostic procedures for undiagnosed diabetes.

2. ROC Estimation for Cluster Designs

2.1. ROC AUC from Independent Samples

The *N* subjects are classified into diseased and normal; for example, using a gold standard, which may involve comparison of another measurement on the subject with a threshold. We assume that n of the subjects are normal and *m* are diseased (with *N*=*m*+*n*) and *2* distinct test measurements are made on the subjects. For the two tests (*r*=1 and 2), we define the values for the diseased subjects as $X_{i(r)}$ for *i=1,..,m* and the values for the normal subjects as $Y_{j(r)}$ for *j=1,...,n*; also, $\underline{X}_i = (X_{i(1)}, X_{i(2)})$ and $\underline{Y}_j = (Y_{j(1)}, Y_{j(2)})$. The two components of $\underline{X}_i$ (or $\underline{Y}_j$) are typically correlated since they arise from measurements on the same subject. Procedures to handle this intra-subject correlation have been introduced in ROC curve testing and estimation [7].

The ROC curve is a plot of the set of pairs (1-specificity(z), sensitivity(z)) for all *z* where specificity (spec) and sensitivity (sens) are defined for the r^{th} diagnostic test as

$$spec(z) = n^{-1} \sum_{j=1}^{n} I(Y_{j(r)} < z),\ sens(z) = m^{-1} \sum_{i=1}^{m} I(X_{i(r)} \geq z)$$

where *I* is the indicator function that is *1* when the event is true and *0* otherwise. We define the Mann-Whitney (MW) statistic for the r^{th} diagnostic test by

$$\hat{\theta}_{(r)} = (mn)^{-1} \sum_{j=1}^{n} \sum_{i=1}^{m} \psi(X_{i(r)}, Y_{j(r)}) \qquad (1)$$

where $\Psi(X,Y) = I(X > Y)$. Then, the MW statistic is an unbiased estimated of the ROC AUC $\left(i.e., E\left(\hat{\theta}_{(r)}\right) = \theta_{(r)}\right)$.

The parameter $\theta_{(r)}$ is the probability that the rth test provides the correct order for persons who are randomly sampled from normal and diseased subjects. If $\theta_{(1)} > \theta_{(2)}$, the first diagnostic test is preferred. A test that the difference between the two diagnostic tests is zero (i.e., $\eta = \underline{L}'\underline{\theta} = 0$ where $\underline{L}' = (1, -1)$ and $\underline{\theta} = (\theta_{(1)}, \theta_{(2)})$) can be based on

$$\hat{\eta} = \underline{L}'\hat{\underline{\theta}} = \hat{\theta}_{(1)} - \hat{\theta}_{(2)} \tag{2}$$

with $\hat{\underline{\theta}}' = \left(\hat{\theta}_{(1)}, \hat{\theta}_{(2)}\right)$. For a sample from an infinite population, an asymptotically valid test of $\eta = 0$ is available [7].

2.2 ROC AUC from Cluster Samples

Now, we relax the independent sample assumption and assume that the sample arose from k clusters and allow the subject's selection probability to vary. Typically, estimates and tests based on data from a complex multi-stage design of this type are made using statistical weights [8]. The most important component of the weight is the inverse of the selection probability; however, other factors such as non-response and post stratification adjustment are often included. Korn and Graubard [9] recommend the use of statistical weights for descriptive analyses of populations. ROC AUC estimation falls into this category since it can be interpreted as the proportion of subjects satisfying a specified condition.

After sampling, all subjects are classified as normal or diseased. In the i^{th} cluster there are m_i and n_i diseased and normal subjects respectively with $m = \sum_{i=1}^{k} m_i$ and $n = \sum_{i=1}^{k} n_i$. For the r^{th} diagnostic test, the measurements on the subjects in the i^{th} cluster are labeled as $\underline{x}_{i(r)} = (x_{i1(r)},...x_{im_i(r)})$ and $\underline{y}_{i(r)} = (y_{i1(r)},...y_{in_i(r)})$ for $i=1,..,k$. The statistical weights for these observations are labeled as $\underline{w}_i^x = (w_{i1}^x,...w_{im_i}^x)$ and $\underline{w}_i^y = (w_{i1}^y,...w_{in_i}^y)$, and the statistical weights are normalized so that their sums over all clusters is one [10]; that is, $\sum_{i=1}^{k}\sum_{s=1}^{m_i} w_{is}^x = 1 = \sum_{i=1}^{k}\sum_{t=1}^{n_i} w_{it}^y$. The ROC is defined using

$$sens(z) = \sum_{i=1}^{k}\sum_{s=1}^{m_i} w_{is}^x I(x_{is(r)} \geq z), \quad spec(z) = \sum_{j=1}^{k}\sum_{t=1}^{n_j} w_{jt}^y I(y_{jt(r)} < z)$$

and the ROC AUC estimate is given by

$$\hat{\theta}_{(r)} = \sum_{i=1}^{k}\sum_{s=1}^{m_i}\sum_{j=1}^{k}\sum_{t=1}^{n_j} w_{is}^x w_{jt}^y \Psi(x_{is(r)}, y_{jt(r)}) \tag{3}$$

In the special case of a single cluster and equal selection probabilities, the weights reduce to $w_{is}^x = m^{-1}$ and $w_{jt}^y = n^{-1}$ so that the estimate (3) reduces to equation (1).

2.3 A Gaussian Population Model for the Sample Values

Here, we propose a population model for the sample values obtained from the two diagnostic tests and calculate the expected value of the ROC AUC estimate under the model, which utilizes random cluster effects [10]. We label the observed values in the i^{th} cluster as $\underline{z}_{i(r)} = (\underline{x}_{i(r)}', \underline{y}_{i(r)}')'$ a vector of length $m_i + n_i$ for $i=1,\ldots,k$ and $r=1,2$, and assume that the components are generated by the model

$$x_{ij(r)} = \mu_{x(r)} + a_{xi(r)} + e_{xij(r)} \tag{4a}$$

$$y_{ij(r)} = \mu_{y(r)} + a_{yi(r)} + e_{yij(r)} \tag{4b}$$

where $\mu_{x(r)}$ and $\mu_{y(r)}$ are fixed, $a_{xi(r)}$ and $a_{yi(r)}$ are random cluster effects for the x- and y-measurements respectively for the i^{th} cluster and r^{th} diagnostic test with

- $a_{xi(r)} \sim N(0, \sigma_{ax(r)}^2)$ for $i=1.,\ldots,k$ and $r=1,2$,
- $a_{yi(r)} \sim N(0, \sigma_{ay(r)}^2)$ for $i=1.,\ldots,k$ and $r=1,2$,

The terms $e_{xij(r)}$ and $e_{yij(r)}$ include the within clusters variation (between-subjects including measurement error) for the x- and y-measurements respectively for the i^{th} cluster, j^{th} subject, and r^{th} diagnostic test ($r=1,2$) with distributions

- $e_{xij(r)} \sim N(0, \sigma_{ex(r)}^2)$ for $i=1.,\ldots,k;\ j=1,\ldots,m_i$, and
- $e_{yij(r)} \sim N(0, \sigma_{ey(r)}^2)$ for $i=1.,\ldots,k;\ j=1,\ldots,n_i$.

The impact of measurement error on inference about the ROC curve has been discussed in non-survey settings [11, 12].

We assume that within cluster measurements on normal and diseased subjects are independent [6]. The four sets of random variables $\{a_{xi(r)}\}$, $\{a_{yi(r)}\}$, $\{e_{xij(r)}\}$, and $\{e_{yij(r)}\}$ are assumed to be mutually independent; also, all random variables defined in different clusters are assumed to be independent while the following dependence is allowed within a cluster. The random cluster effects for the x-measurements for the two diagnostic tests have correlation (Cor) coefficient ρ_{ax} (with a similar definition for y-measurements)

- $\rho_{ax} = Cor(a_{xi(1)}, a_{xi(2)})$ & $\rho_{ay} = Cor(a_{yi(1)}, a_{yi(2)})$ for all i

Furthermore, the x-measurements for the two diagnostic tests have correlation coefficient ρ_{ex} for every subject (with a similar definition for y-measurements)

- $\rho_{ex} = Cor(e_{xij(1)}, e_{xij(2)})$ & $\rho_{ey} = Cor(e_{yij(1)}, e_{yij(2)})$ for all i ,j.

It follows that $Var(x_{ij(r)}) = \sigma_{x(r)}^2$ with $\sigma_{x(r)}^2 = \sigma_{ax(r)}^2 + \sigma_{ex(r)}^2$, and $Cov(x_{ij(r)}, x_{ij'(r)}) = \sigma_{ax(r)}^2 \equiv \rho_{x(r)}\sigma_{x(r)}^2$ for $j \neq j'$ with the intra-cluster correlation coefficient defined by $\rho_{x(r)} = \sigma_{ax(r)}^2 / \sigma_{x(r)}^2$ with similar definitions for $\sigma_{y(r)}^2$ and $\rho_{y(r)}$.

Thus, within a cluster the observations $\underline{z}_{i(r)}$ have a (block-diagonal) equi-correlated multivariate normal (MVN) model of the form

$$\underline{z}_{i(r)} \sim N\left(\begin{pmatrix}\mu_{x(r)}\underline{1}_{m_i} \\ \mu_{y(r)}\underline{1}_{n_i}\end{pmatrix}, \begin{pmatrix}R(\sigma_{x(r)}^2, \rho_{x(r)}\sigma_{x(r)}^2) & 0 \\ 0 & R(\sigma_{y(r)}^2, \rho_{y(r)}\sigma_{y(r)}^2)\end{pmatrix}\right) \tag{5}$$

where $N(\underline{\mu}, \Sigma)$ denotes the MVN distribution with mean $\underline{\mu}$ and covariance matrix Σ, $\underline{1}_n$ denotes a vector of n ones, and $R(v,c)$ denotes the equi-correlated matrix with all diagonal elements (variances) v and all off diagonal elements (covariances) equal to c.

The expected value of the ROC AUC estimate (3) can be calculated from the model assumptions. Now, $E\left(\Psi(x_{is(r)}, y_{jt(r)})\right) = \Pr(y_{jt(r)} - x_{is(r)} < 0)$, and from (5) $y_{jt(r)} - x_{is(r)} \sim N(\mu_{z(r)}, \sigma_{z(r)}^2)$ with $\mu_{z(r)} = \mu_{y(r)} - \mu_{x(r)}$ and $\sigma_{z(r)}^2 = \sigma_{x(r)}^2 + \sigma_{y(r)}^2$. Thus, $E\left(\Psi(x_{is(r)}, y_{jt(r)})\right) = \Phi\left(\delta_{z(r)}\right)$ where $\delta_{z(r)} = -\mu_{z(r)} / \sigma_{z(r)}$ and Φ denotes the standard normal cumulative distribution function (c.d.f.). From equation (3) we have that $E\left(\hat{\theta}_{(r)}\right) = \Phi\left(\delta_{z(r)}\right)$ since both sets of weights are normalized. To test the hypothesis $\eta = 0$, we need the variance/covariance matrix of $\underline{\hat{\theta}}' = \left(\hat{\theta}_{(1)}, \hat{\theta}_{(2)}\right)$. In the Appendix, we provide an exact expression for this 2x2 matrix.

This approach can be extended to include linear covariates in equations (4a) and (4b). A useful extension would be to allow correlation between normal and diseased subjects within a cluster.

2.4 Estimation of the model parameters

The measurements from the two diagnostic tests on *N* subjects can be written as $\underline{z}' = (\underline{x}', \underline{y}')$ with $\underline{y}$ a vector of length *2n* defined by $\underline{y}' = (\underline{y}_1', \ldots, \underline{y}_k')$ where $\underline{y}_i$ is of length $2n_i$ with $\underline{y}_i' = (\underline{y}_{i1}', \ldots, \underline{y}_{in_i}')$ with $\underline{y}_{ij}' = (y_{ij(1)}, y_{ij(2)})$; the x-measurements can be written in a similar fashion. Then, $\underline{x}$ and $\underline{y}$ are independent and follow the mixed-linear model

$$\left\langle \begin{aligned} \underline{y} &= \mathrm{X}_y \underline{\mu}_y + \mathrm{Z}_y \underline{a}_y + \underline{\varepsilon}_y \\ \underline{x} &= \mathrm{X}_x \underline{\mu}_x + \mathrm{Z}_x \underline{a}_x + \underline{\varepsilon}_x \end{aligned} \right.$$

where, $\mathrm{X}_y = \underline{1}_n \otimes I_2$, $Z_y = diag(\underline{1}_{n_1} \otimes I_2, \ldots, \underline{1}_{n_k} \otimes I_2)$ with $\otimes$ the "direct product", $\underline{\mu}_y' = (\mu_{y(1)}, \mu_{y(2)})$, $\underline{a}_y$ the *2k* dimensional vector of random components is defined by $\underline{a}_y' = (\underline{a}_{y1}', \ldots, \underline{a}_{yk}')$ with $\underline{a}_{yi}' = (a_{yi(1)}, a_{yi(2)})$ for *i=1,…,k*, and $\underline{e}_y$ the error vector is defined by $\underline{e}_y' = (\underline{e}_{y1}', \ldots, \underline{e}_{yk}')$ where $\underline{e}_{yi}$ is a vector of length $2n_i$ given by $\underline{e}_{yi}' = (\underline{e}_{yi1}', \ldots, \underline{e}_{yin_i}')$ with $\underline{\varepsilon}_{yij}' = (\varepsilon_{yij(1)}, \varepsilon_{yij(2)})$. The joint distribution of the random effects and the errors of $\underline{y}$ are given by

$$\begin{pmatrix} \underline{a}_y \\ \underline{\varepsilon}_y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} G_y & 0 \\ 0 & R_y \end{pmatrix}\right)$$

where, $G_y = \Sigma_{ay} \otimes I_k$ and $R_y = \Sigma_{\varepsilon y} \otimes I_n$ with

$$\Sigma_{ay} = \begin{pmatrix} \sigma^2_{ay(1)} & \rho_{ay}\sigma_{ay(1)}\sigma_{ay(2)} \\ \rho_{ay}\sigma_{ay(1)}\sigma_{ay(2)} & \sigma^2_{ay(2)} \end{pmatrix} \quad \Sigma_{\varepsilon y} = \begin{pmatrix} \sigma^2_{\varepsilon y(1)} & \rho_{\varepsilon y}\sigma_{\varepsilon y(1)}\sigma_{\varepsilon y(2)} \\ \rho_{\varepsilon y}\sigma_{\varepsilon y(1)}\sigma_{\varepsilon y(2)} & \sigma^2_{\varepsilon y(2)} \end{pmatrix}$$

The 2x2 covariance matrices Σ_{ay} and $\Sigma_{\varepsilon y}$ measure the between- and within-cluster (including measurement error) variation for the normal subjects respectively. Similar definitions and distributions are valid for $\underline{x}$.

Either maximum likelihood estimation (MLE) or restricted maximum likelihood (REML) can be used to obtain consistent and asymptotically normal estimates of the model parameters. We used the EM algorithm to obtain the MLE [13]. Due to the separability of the likelihood, the parameter estimates of $\underline{x}$ and $\underline{y}$ can be obtained separately so it is sufficient to describe the estimation process for $\underline{y}$. The variance can be expressed as $Var(\underline{y}) = V_y = Z_y G_y Z_y' + R_y$ and given the estimates of the variance/covariance matrices, the estimates for the fixed and random effects can be obtained by generalized least squares (GLS) and are given by

$$\left\langle \begin{aligned} \underline{\hat{\mu}}_y &= \left(\mathrm{X}_y' \hat{V}_y^{-1} \mathrm{X}_y\right)^{-1} \mathrm{X}_y' \hat{V}_y^{-1} \underline{y} \\ \underline{\hat{a}}_y &= \hat{G}_y Z_y' \hat{V}_y^{-1} \underline{r}_y \end{aligned} \right.$$

where the residual vector is $\underline{r}_y = \underline{y} - \mathrm{X}_y \underline{\hat{\mu}}_y$. The MLE was obtained in an iterative fashion by minimizing

$$-2 * \ln(L) = \ln\left|\hat{V}_y\right| + n\left(\underline{r}_y' \hat{V}_y^{-1} \underline{r}_y\right) \tag{6}$$

with respect to the covariance matrices $\left(\Sigma_{ay}, \Sigma_{\varepsilon y}\right)$. To use the weights in the parameter estimation, we replaced V_y with V_{wy} in equation (6) where $V_{wy} = \tilde{W}_y V_y \tilde{W}_y$, $\tilde{W}_y = W_y \otimes I_2$, and $W_y = Diag\left(\left(w_{jt}^y\right)^{1/2}\right)$ with a similar modification for V_x [14].

3. ROC AUC Estimation using Replication

Here, we sketch how two replication methods, the jackknife and balanced repeated replication (BRR), are used in complex surveys to estimate the variance of nonlinear functions such as the difference in the two ROC AUCs. The two replication procedures are asymptotically valid in the presence of within-cluster correlation and provide a check on the estimation obtained under the parametric model. We assume that the survey design is such that the clusters are formed from *S* strata with *2* primary sampling units (PSUs) per strata. Thus, the total number of clusters is *k=2*S*.

3.1 The Jackknife

Quenouille [15] introduced the jackknife as a bias reduction tool. Later, Tukey [16] conjectured that it could be used to estimate variance for a large class of (possibly non-linear) estimators. Efron [17] and Wolter [18, Chap. 4] provide comprehensive discussion of the jackknife for infinite populations and finite populations respectively.

For variance estimation, the jackknife proceeds by splitting the sample into a set of disjoint groups of the same approximate size. Each of these groups is removed, and the estimate is calculated on the remaining observations. The variability of the resulting estimates can be used to estimate the variance of the full-sample estimator.

When applying the jackknife methodology to multi-stage cluster samples, the groups are formed using the ultimate clusters [18, Sec. 2.4] rather than the elementary units. The estimator $\hat{\eta}_{[s,l]}$ is computed for *s=1,...,S* and *l=1,2* using (2) and (3) but eliminating the observations in stratum *s* and PSU *l* and by doubling the weights in the remaining PSU in stratum *s* [19]. The jackknife estimate of variance is

$$\hat{Var}(\hat{\eta}) = 2^{-k} \sum_{s=1}^{S} \sum_{l=1}^{2} (\hat{\eta}_{[s,l]} - \hat{\eta})^2 \tag{7}$$

with $\hat{\eta}$ the full sample estimate. For linear functions of the strata means, (7) is an unbiased estimate; even for nonlinear functions such as $\hat{\eta}$ the method usually performs well [20].

3.2 Balanced Repeated Replication (BRR)

Whereas, the jackknife procedure retains most of the sample in each replicate, the balanced repeated replication method (BRR) retains about one-half of the sample. Because there are *S* strata with *2* PSUs/strata, the total number of replicate samples that can be formed is 2^S. However, all of the information in the replicates is available in *g* orthogonal or "balanced" replications, where *g* is the smallest integer divisible by 4 that is greater than or equal to *S*.

We used Fay's method, a variant of the BRR method as a second replication procedure. See Judkins [21] for a review of this method, which is based on g- parameter estimates, $\hat{\eta}_{[i]}$. The replication estimates are calculated using the standard full-sample estimator, $\hat{\eta}$, with statistical weights that depend on the replication. For each replication, the sampling weights in the selected half sample are multiplied by *f* while the remaining (half-sample) weights are multiplied by *2-f* where $0<f<1$. Since every observation has positive weight, it is viewed as a compromise between the jackknife and standard BRR. Using Fay's method, the variance estimator is

$$\hat{Var}(\hat{\eta}) = \frac{1}{g(1-f)^2}\sum_{i=1}^{g}(\hat{\eta}_{[i]} - \hat{\eta})^2 \qquad (8)$$

where $\hat{\eta}_{[i]}$ the estimate using the i^{th} replication ($i=1,..,g$).

4. Detecting Undiagnosed Diabetes using NHANES III

In this section, we test the equality of the ROC AUCs for two predictors of undiagnosed diabetes using data from the third National Health and Nutrition Examination Survey (NHANES III). In a related study, Thompson, Smith and Boyle [22] applied a model-based approach for diabetes detection to data from an Egyptian population-based household survey but did not phrase the problem in terms of ROC. Prior to discussing our estimation results, we give a brief description of the NHANES III survey design, data collection, and study population.

4.1 NHANES III: Survey Design

NHANES III was conducted from 1988 to 1994 by the National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention (CDC) [23-24]. A nationally representative sample of the U.S. civilian non-institutionalized population was selected using a complex, stratified, multi-stage probability cluster sampling design. The survey design included planned oversampling of black and Mexican-American persons, as well as of children and the elderly, to provide more precise estimates for these subgroups of the population. Informed consent was obtained from all respondents and the protocol was reviewed and approved by the NCHS NHANES Institutional Review Board. Health and dietary information on sampled persons was obtained through an interview in the home followed by a standardized physical examination in a mobile examination center (MEC).

A four stage sampling design was used: (1) PSUs consisting mostly of single counties, (2) area segments within PSUs, (3) households within area segments, and (4) persons within households. The probability of selection of a person in NHANES III depended on the PSU and the person's age-sex-race/ethnicity domain. For NHANES III, the design had *49* pseudo-strata each with *2* pseudo-PSUs. We use the pseudo-strata as strata and the pseudo-PSUs as PSUs to form the clusters for the jackknife analysis.

For the BRR analysis, *g=52* replicates were used (as the smallest integer divisible by *4* and also greater than *49*). The NHANES III MEC examination replicate weights [24] using Fay's method with *f=0.3* were used to determine the replicate estimates that were used in equation (8).

4.2 Administration of the Oral Glucose Tolerance Test

In NHANES III, an oral glucose tolerance test (OGTT) was administered to person's ages 40-74 years as part of the examination. Each interviewed household was randomly assigned to either the morning or to the afternoon/evening session. Those assigned to the morning session were requested to fast overnight. For the OGTT, a fasting blood sample was drawn and a 75-g glucose-equivalent oral glucose challenge was then administered. A second blood sample was drawn two hours after the glucose challenge.

4.3 Study Population, Sample Design, and Gold Standard

For the purposes of this analysis, the sample consisted of persons who were assigned to the morning session and examined in the morning after an overnight fast of at least 9 hours but less than 24 hours and who had a second blood draw 2 hours ± 15 minutes after the glucose challenge was administered. This group conforms most closely to the World Health Organization (WHO) criteria for OGTT testing [25]. We excluded subjects who had diabetes based on self-report. Persons with a fasting plasma glucose of 140 mg/dL or greater or with a two-hour plasma glucose of 200 mg/dL or greater were considered to have undiagnosed diabetes (i.e., diseased).

4.4 Model-based Estimation Results

We used data for all 3053 NHANES III subjects who met the criteria described in Section 4.3. Of these subjects, 420 (13.8%) had undiagnosed diabetes using the OGTT gold standard. Then, we evaluated the diagnostic capability, as compared to the results from the OGTT test, of each of two alternative procedures for defining diabetes, namely fasting plasma glucose (FPG) and glycohemoglobin (HbA1c).

Scatterplots of the logarithm of two diagnostic measurements on the same scale for diseased and normal subjects (Figure 1a and 1b) yields the following conclusions:

- The normality assumption of the logarithm of the measurements appears reasonable for both normal and diseased subjects.
- The correlation between the two measurements on the same subject is appreciable and must be considered in the analysis. The correlation between the two procedures is much higher for diseased (Pearson correlation, *r*=0.87) than for normal subjects (*r*=0.26).
- In general, diseased patients have larger values for both measurements. However, classification cannot be made perfectly with either measurement (since the distributions of diseased and normal subjects overlap).

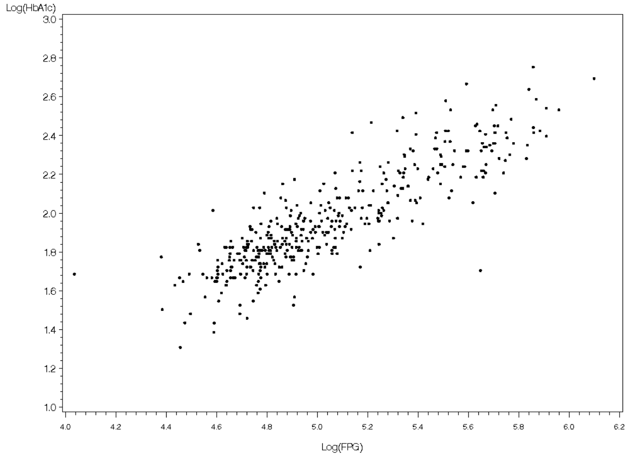

Figure 1a. Log of Diagnostic Measures for Diseased Subjects

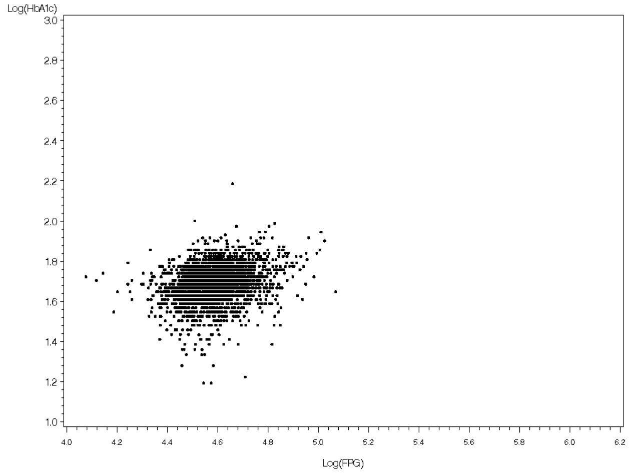

Figure 1b. Log of Diagnostic Measures for Normal Subjects

The mle's of the parameters were obtained using SAS IML [26]. Nested models for normal and diseased subjects were used to test for the statistical significance of the correlation coefficients. For example, for diseased subjects, we used the three nested hypotheses:

- H$_1$: (ρ_{ax},ρ_{ex}) unrestricted
- H$_2$: $\rho_{ex}=0$ and ρ_{ax} unrestricted
- H$_3$: $(\rho_{ax},\rho_{ex})=(0,0)$

Parameters were estimated for each of these hypotheses, and the difference in the log-likelihood obtained from (6) was used to test the significance of the correlations by comparison with a chi-square distribution with 1 degree of freedom. Table 1 shows the values of -2ln(L), the differences, and the p-value of the hypothesis test. For diseased subjects, both tests were statistically significant at the 5% level while for normal subjects neither test was significant (estimated values for H$_1$ were $(\hat{\rho}_{ay}=.18,\hat{\rho}_{ey}=-.02)$). This analysis showed that both correlations were necessary in the analysis; at least, for diseased subjects. For example, the test of the null hypothesis $\rho_{ex}=0$ can be obtained by testing H$_2$ versus H$_1$; from the table the test statistic is 14.1 (=1340.96-1326.82) and the p-value of the test is <0.0001.

Table 1. Nested models for normal and diseased subjects

Diseased subjects	-2ln(L)	Difference (p-value)
unrestricted	-1340.96	
$\rho_{ex}=0$	-1326.82	14.1 (<0.0001)
Both zero	-1275.17	51.6 (<0.0001)
Normal subjects		
unrestricted	-19565.78	
$\rho_{ev}=0$	-19564.60	1.2 (0.30)
Both zero	-19562.36	2.2 (0.11)

Based on the tests summarized in table 1, we included both correlations for diseased subjects but neither for normal subjects. The estimated covariance matrices of the random effects and measurement errors (with correlation coefficients below the diagonal) for both diseased (x) and normal (y) subjects were given by

$$\Sigma_{ax}=\begin{pmatrix}0.02860 & 0.01924\\ 0.9959 & 0.01304\end{pmatrix}\quad \Sigma_{ex}=\begin{pmatrix}0.08114 & -0.0126\\ -0.1826 & 0.05875\end{pmatrix} \tag{9a}$$

$$\Sigma_{ay}=\begin{pmatrix}0.000594 & 0.000\\ 0.000 & 0.000676\end{pmatrix}\quad \Sigma_{ey}=\begin{pmatrix}0.00936 & 0.000\\ 0.000 & 0.00798\end{pmatrix} \tag{9b}$$

The distribution of the (unweighted) cluster mean is

$$\bar{\underline{x}}_{i.}\sim N(\underline{\mu}_x,\Sigma_{ax}+m_i^{-1}\Sigma_{ex}) \tag{10}$$

where m_i is the number of diseased subjects in the i^{th} cluster and 'dot' denotes average. The unweighted Pearson correlation of cluster means is 0.914. However, measurement error (and within-cluster variation) can reduce the correlation, so it is not surprising the mixed-model gives a higher correlation estimate for cluster means, $\rho_{ax}=0.9959$. In summary, the model predicts an almost co-linear relationship between the cluster means of diseased subjects -- if large numbers of subjects are obtained in each cluster. Rohlfing *et al.* [27] gave correlation results for diseased subjects for FPG and Hb1Ac from a large multi-center clinical trial.

4.5 Design of future studies

The variance inflation factor (*IF*) due to clustering and the effective sample size (n$_{\text{eff}}$) can be calculated from the estimated covariance matrices. The *IF* is given by

$$IF=1+(\bar{n}-1)\rho \tag{11}$$

where $\bar{n}$ is the average cluster size and ρ is the intra-cluster correlation coefficient. The *IF* measures the ratio of the variance of a cluster sample mean to an independent sample mean with the same number of subjects [28]. The effective cluster sample size is the equivalent number of independent subjects ($n_{eff}=n/IF$ where n is the total number of subjects). Table 2 shows the inflation factors, the effective sample size, and the actual sample size for both normal and diseased subjects. For example, for FPG the *IF* is obtained from (11) where $\bar{n}=420/96$ and the intra-cluster correlation is given by $\rho=0.02860/(0.02860+0.08114)=0.26$ from (9a). The inflation factors are useful in conducting power analyses for the design of future studies. The table shows a total effective sample size of 1195 for FPG and 1072 for HBA1c respectively; an appreciable reduction in both cases from the actual number of 3053 subjects. In the table, the *IFs* range from 1.80 to 3.14, and the corresponding increase in the length of the associated confidence intervals range from 34% to 77%.

Table 2. Inflation factor and effective sample sizes

	Fasting plasma glucose (FPG)		Glycohemoglobin (HbA1c)		
	IF	n_{eff}	IF	n_{eff}	Actual n
Diseased	2.14	196	1.80	234	420
Normal	2.64	999	3.14	838	2633
Total		1195		1072	3053

4.6 Model-Based ROC AUC estimation and testing

Figure 2 shows the ROC curves for FPG and HbA1c. Since the ROC curve for FPG lies completely above the curve for HbA1c, FPG seems to be a better predictor of undiagnosed diabetes than HbA1c. We use (FPG, HbA1c) as the ordering of the tests; from equation (3), we estimate the ROC AUCs and obtain $\underline{\hat{\theta}}$ =(0.9310, 0.8665) so their difference is estimated by $\hat{\eta} = 0.9310 - 0.8665 = 0.0645$ using (2).

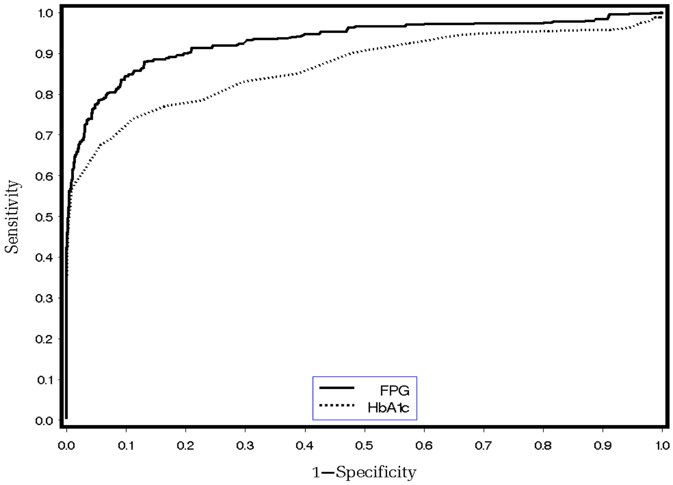

Figure 2. Weighted ROC Curves for Undiagnosed Diabetes

To test the hypothesis of the equality of the two diagnostic procedures, we replace parameters by their estimates in equation (A-2) and (A-10) and obtain the estimated covariance matrix of $\underline{\hat{\theta}}$

$$Var(\underline{\hat{\theta}}) = 10^{-4}\begin{pmatrix} 1.691 & 1.415 \\ 0.617 & 3.109 \end{pmatrix} \qquad (12)$$

where, again, the element below the diagonal in matrix (12) is the correlation coefficient. From (12), the variance of the difference $\hat{\eta} = \hat{\theta}_{(1)} - \hat{\theta}_{(2)}$ is estimated by

$$Var(\hat{\eta}) = 10^{-4}(1.691 + 3.109 - 2.830) = (0.0140)^2 . \qquad (13)$$

The distribution of $\hat{\eta}$ is approximated as a Student t-distribution with *49* degrees of freedom (the difference between the number of sampled PSUs and strata). Thus, the 95% confidence interval for the difference in the ROC AUCs is (0.037, 0.093). Since the confidence interval does not contain zero, we reject the null hypothesis at the 5% level.

4.7 Comparison with alternative procedures

In table 3, we compare the model-based results with the two data replication procedures. The model-based and replication methods do not differ in the estimate of the ROC AUCs; so the difference is estimated by 0.0645 for all three procedures. However, the estimated standard error, t-statistic, and confidence interval for the difference differ. For all three, the table shows that we obtain the same general conclusion; namely, FPG is a better predictor than HbA1c for undiagnosed diabetes since the 95% confidence intervals do not overlap the origin (In fact, all differences are statistically significant at p<0.001). The model-based procedure yields the smallest standard error; hence the largest t-statistic and shortest confidence interval. The model-based procedure is the only exact test and also allows checks of the analytic assumptions.

Table 3. Statistics for the difference of the ROC AUCs

Method	St. Err.	T-Stat	95% CI
Model-Based	0.0140	4.60	(0.037, 0.093)
Jackknife	0.0153	4.21	(0.034, 0.095)
BRR	0.0162	3.98	(0.032, 0.097)

It is interesting to compare the results with those that would be obtained if the sampling design were ignored, and the 3053 subjects were treated as an independent and identically distributed sample so that the ROC AUC estimation procedure proposed in [7] is valid. Applying the method of [7], a 95% confidence interval for the ROC AUC difference is 0.0584 ± 0.0198 . Since the 95% confidence interval does not contain zero, the result of the hypothesis test is the same as the result obtained from the procedures that utilize the design information (table 3).

In general, ignoring the sampling design could lead to erroneous conclusions since the estimated standard error is overly optimistic -- hence confidence limits are too narrow. As a measure of the reduction in variance, we calculate the inflation factor (IF) for the variance of the ROC AUC difference; that is we calculate the ratio of $Var(\hat{\eta})$ from the model-based (i.e., table 3) to the independent analysis from [7]

$$IF = Var(\text{model} - \text{based}) / Var(\text{indep.}) = 2.02 \qquad (14)$$

Equation (14) shows that the variance from the independent analysis is approximately half as much as it would be if the design were utilized.

5. DISCUSSION

When two empirical ROC curves are constructed to evaluate two diagnostic tests, statistical tests on the difference between the curves must take into account the correlated nature of the data. Techniques have been developed to deal with intra-subject (between two measurements on the same subject) correlation.

In most national surveys, the data are obtained from a multi-stage design with cluster sampling so the correlation of the measurements may be higher for subjects within the same PSU than for subjects in different PSUs due to the similarity of subjects who reside in a small area. Here, we provide a model-based approach that can be used to test whether the difference between two ROC AUC curves is statistically significant when the data are obtained from a cluster sample with different selection probabilities. The ROC AUC variance is expressed in terms of parameters of a mixed-model, which include parameters measuring the between- and within-cluster variation in both measurements. Estimation of the parameter yields a test of the equality of two ROC AUCs.

The model-based procedure is applied to NHANES III data to test for the difference between two predictors of undiagnosed diabetes. Previously, the only procedure for testing the equality of ROC AUCs for cluster samples [6] assumed equal selection probabilities, which is not satisfied for NHANES III. The hypothesis test of the ROC AUC yields similar conclusions to that obtained from the replication procedures (the jackknife method and balanced repeated replication), which are often used to approximate the variance of non-linear estimators (such as the ROC AUC) for complex survey designs. However, the model-based analysis yields the exact variance under the model assumptions, allows checks for assumptions, yields insights that are difficult to obtain without a model, and can be useful in designing future studies. In addition, the model-based results are contrasted with the analysis that ignores the sample design [7], which yields overly optimistic confidence limits.

Acknowledgement

The authors wish to thank David DeLong and Amrik Shah for sharing SAS IML software, John McPeek, Yishu Yang, and Tom Evans for programming assistance, and Grant Izmirlian, Rocky Feuer, and Jennifer Madans for useful comments.

6. References

1. Hanley JA, McNeil, BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**, 29-36.
2. Hanley JA. The use of the 'Binormal' model for parametric ROC analysis of quantitative diagnostic tests. *Stat. in Med.* 1986; **15**, 1575-1585.
3. Meta CE, Herman BA, Shen J-H. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. *Stat. in Med.* 1988; **17**, 1033-1053.
4. Smith PJ, Thompson TJ, Engelgau MM, Herman WH. A generalized linear model for analysing receiver operating characteristic curves. *Stat. in Med.* 1996; **15**, 323-333.
5. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. of Math. Psych.* 1975; **12**, 387-415.
6. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. *Biometrics* 1997; **53**, 567-578.
7. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 1988; **44**, 837-845.
8. Kish L. Weighting for unequal P_i. *J. of Off. Stat.* 1992*;* **8**, 183-200.
9. Korn EL, Graubard BI. *Analysis of Health Surveys*. John Wiley, New York, 1999, Sec. 4.6.
10. Shah BV, LaVange LM. Mixed Models for Survey Data. *Proceedings of the Amer. Stat. Assn: Survey Research Section* 1996*;* 111-116.
11. Coffin M, Sukhatme S. Receiver operating characteristic studies and measurement error. *Biometrics* 1997; **53**, 823-837.
12. Faraggi D. The effect of random measurement error on receiver operating characteristic (ROC) curves. *Stat. in Med.* 2000; **19**, 61-70.
13. Shah A, Laird N, Schoenfeld D. A random-effects model for multiple characteristics with possibly missing data. *J. of Amer. Stat. Assn.* 1997; **92**, 775-779.
14. Goldstein H. *Multilevel Statistical Models*. Edward Arnold, London, 1995, Sec. 3.4.
15. Quenouille MH. Notes on bias in estimation. *Biometrika* 1956; **43**, 353-360.
16. Tukey JW. Bias and confidence in not-quite large samples. *Annals of Math. Stat.* 1958; **29**, 614.
17. Efron B. *The Jackknife, the Bootstrap, and other Resampling Plans*. SIAM Monograph, Philadelphia, 1982.
18. Wolter KM. *Introduction to Variance Estimation*. Springer-Verlag, New York, 1985.
19. Kalton G. Practical methods for estimating survey sampling errors. *Bull. of the Int. Stat. Inst.* 1977; **47**, 495-514.
20. Rust KF. Variance estimation for complex estimators in sample surveys. *J. of Off. Stat.* 1985; **4**, 381-397.
21. Judkins DR. Fay's Method for Variance Estimation. *J. of Off. Stat.* 1990*;* **6**, 223-239.
22. Thompson TJ, Smith Pj, Boyle JP. Finite mixture models with concomitant information: assessing diagnostic criteria for diabetes. *App. Stat.* 1996; 47, 393-404.
23. National Center for Health Statistics. Plan and operation of the Third National Health and Nutrition Examination Survey, 1988-94. *Vital and Health Statistics* 1995; Ser 1, No. 32. DHHS Publ No. (PHS) 94-1308. U.S. Public Health Service. Washington, DC: GPO. 3
24. National Center for Health Statistics *NHANES III Reference Manuals and Reports*, CD-ROM, 1996.
25. World Health Organization. Diabetes Mellitus: Report of a WHO Study Group. World Health Organization, 1985. (Technical Report Series No. 646).
26. SAS Institute Inc. *SAS/IML Software: Usage and Reference Manual*. Ver. 6, SAS Institute Inc., Cary, NC, 1995.
27. Rohfling CL, Wiedmeyer HM, Little RR, England JD, Tennill A, Goldstein DE. Defining the relationship between Plasma Glucose and HbA1c. *Diabetes Care* 2002 (25) 275-278.
28. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold, London, 2000, Sec 1.3.

APPENDIX: ROC AUC Covariance Matrix

We calculate the covariance matrix of $\underline{\hat{\theta}}$. The elements of the 2x2 covariance matrix $C = \left(C_{r,r'}\right) = \left(Cov\left(\hat{\theta}_{(r)}, \hat{\theta}_{(r')}\right)\right)$ are

$$C_{r,r'} = \sum_{i,s} \sum_{i',s'} \sum_{j,t} \sum_{j',t'} w_{is}^{x} w_{jt}^{y} w_{i's'}^{x} w_{j't'}^{y} Cov(\Psi(x_{is(r)}, y_{jt(r)}), \Psi(x_{i's'(r')}, y_{j't'(r')}))$$

First, we calculate the variance; that is, $C_{r,r}$ .The covariance term on the right side of the equation involves two x-subjects defined by pairs (i,s) and (i',s') and two y-subjects defined by pairs (j,t) and (j',t') . If we define the subject pairs by the four-dimensional vectors $\underline{i} = (i, s, i', s')$ and $\underline{j} = (j, t, j', t')$ then the covariance term is constant for $(\underline{i}, \underline{j}) \in A_u^x x B_v^y$ where $1 \le u, v \le 3$ and the sets A_u^x and B_v^y are defined as follows:

- $A_1^x = \{\underline{i} : i = i', s = s'\}$ and
- $A_2^x = \{\underline{i} : i = i', s \ne s'\}$ and $B_2^y = \{\underline{j} : j = j', t \ne t'\}$
- $A_3^x = \{\underline{i} : i \ne i'\}$ and $B_1^y = \{\underline{j} : j \ne j'\}$

These sets are characterized as follows. In A_1^x and B_1^y, the two subjects are the same. In A_2^x and B_2^y, the subjects are different but the clusters are the same. In A_3^x and B_3^y, the clusters are different. Now ,

$$Var\left(\hat{\theta}_{(r)}\right) = \sum_{u=1}^{3} \sum_{v=1}^{3} W_{uv} \varsigma_{uv(r)} \qquad \text{(A-1)}$$

with $\varsigma_{uv(r)} = Cov(\Psi(x_{is(r)}, y_{jt(r)}), \Psi(x_{i's'(r)}, y_{j't'(r)}))$ for $(\underline{i}, \underline{j}) \in A_u^x x B_v^y$ and

$$W_{uv} = \left(\sum_{\underline{i} \in A_u^x} w_{is}^x w_{i's'}^x \right) \left(\sum_{\underline{j} \in B_v^y} w_{jt}^y w_{j't'}^y \right) \qquad \text{(A-2)}$$

The 3x3 matrix W=(W$_{uv}$) can be expressed as the product

$$W = \begin{pmatrix} W_x \\ \tilde{W}_x - W_x \\ 1 - \tilde{W}_x \end{pmatrix} \begin{pmatrix} W_y & \tilde{W}_y - W_y & 1 - \tilde{W}_y \end{pmatrix} \qquad \text{(A-3)}$$

with $W_x = \sum_{is} \left(w_{is}^x\right)^2$, $W_y = \sum_{jt} \left(w_{jt}^y\right)^2$, $\tilde{W}_x = \sum_i \left(\sum_s w_{is}^x \right)^2$, and $\tilde{W}_y = \sum_j \left(\sum_t w_{jt}^y \right)^2$.

Now, we express $\varsigma_{uv(r)}$ in terms of the parameter vector $\underline{\alpha}_{(r)} = (\mu_{x(r)}, \mu_{y(r)}, \sigma_{x(r)}^2, \sigma_{y(r)}^2, \rho_{x(r)}, \rho_{y(r)})$. Since

$$Cov(\Psi(x_{is(r)}, y_{jt(r)}), \Psi(x_{i',s'(r)}, y_{j',t'(r)})) =$$
$$\Pr\left((y_{jt(r)} - x_{is(r)}, y_{j't'(r)} - x_{i's'(r)}) < (0,0)\right) - \Phi^2\left(\delta_{z(r)}\right) \qquad \text{(A-4)}$$

and with $\rho_{(r)} = \rho_{(r)}(\underline{i}, \underline{j})$

$$\begin{pmatrix} y_{jt(r)} - x_{is(r)} \\ y_{j't'(r)} - x_{i's'(r)} \end{pmatrix} \sim \mathrm{N}\left(\mu_{z(r)} \underline{1}, \sigma_{z(r)}^2 \begin{pmatrix} 1 & \rho_{(r)} \\ \rho_{(r)} & 1 \end{pmatrix} \right) \qquad \text{(A-5)}$$

it follows that $\Pr\left((y_{jt(r)} - x_{is(r)}, y_{j't'(r)} - x_{i's'(r)}) < (0,0)\right) = F_{\rho(r)}(\delta_{z(r)} \underline{1})$ where $F_\rho(.)$ is the c.d.f. of the normalized bivariate normal distribution with correlation ρ. Since the x- and y- values are uncorrelated, from equation (A-5) we have

$$\rho_{(r)} \sigma_{z(r)}^2 = Cov(\Psi(x_{is(r)}, y_{jt(r)}), \Psi(x_{i's'(r)}, y_{j't'(r)}))$$
$$= Cov(y_{jt(r)}, y_{j't'(r)}) + Cov(x_{is(r)}, x_{i's'(r)}) \qquad \text{(A-6)}$$

The two covariance terms on the right hand side of equation (A-6) can be calculated using equation (5). For $(\underline{i}, \underline{j}) \in A_u^x x B_v^y$, the right hand side of equation (A-6) is constant and we define the value as $\gamma_{uv(r)}$. Then, the matrix $\Gamma_{(r)} = \left(\gamma_{uv(r)}\right)$ is given by

$$\Gamma_{(r)} = \begin{pmatrix} \sigma_{x(r)}^2 + \sigma_{y(r)}^2 & \sigma_{x(r)}^2 + \rho_{y(r)} \sigma_{y(r)}^2 & \sigma_{x(r)}^2 \\ \rho_{x(r)} \sigma_{x(r)}^2 + \sigma_{y(r)}^2 & \rho_{x(r)} \sigma_{x(r)}^2 + \rho_{y(r)} \sigma_{y(r)}^2 & \rho_{x(r)} \sigma_{x(r)}^2 \\ \sigma_{y(r)}^2 & \rho_{y(r)} \sigma_{y(r)}^2 & 0 \end{pmatrix} \qquad \text{(A-7)}$$

The correlation $\rho_{(r)} = \rho_{uv(r)}$ of (A-5) and (A-6) can be obtained from $\rho_{uv(r)} = \gamma_{uv(r)} / \sigma_{z(r)}^2$.

In summary, the variance of the ROC AUC estimate (3) can be calculated exactly from equation (A-1) where W$_{uv}$ for $1 \le u, v \le 3$ is defined by equation (A-2),

$$\varsigma_{uv(r)} = F_{\rho_{uv(r)}}(\delta_{z(r)} \underline{1}) - \Phi^2\left(\delta_{z(r)}\right) \qquad \text{(A-8)}$$

with $\delta_{z(r)} = -\mu_{z(r)} / \sigma_{z(r)}$, $\mu_{z(r)} = \mu_{y(r)} - \mu_{x(r)}$, $\sigma_{z(r)}^2 = \sigma_{x(r)}^2 + \sigma_{y(r)}^2$, $\rho_{uv(r)} = \gamma_{uv(r)} / \sigma_{z(r)}^2$, $\Gamma_{(r)} = \left(\gamma_{uv(r)}\right)$ is defined in equation (A-7), Φ denotes the standard normal c.d.f., and $F_\rho(.)$ is the c.d.f. of the normalized bivariate normal distribution with correlation ρ.

Calculation of $Cov\left(\hat{\theta}_{(1)}, \hat{\theta}_{(2)}\right)$ is similar but involves

$$Cov(\underline{z}_{i(1)}, \underline{z}_{i(2)}) = \begin{pmatrix} Cov(\underline{x}_{i(1)}, \underline{x}_{i(2)}) & Cov(\underline{x}_{i(1)}, \underline{y}_{i(2)}) \\ Cov(\underline{y}_{i(1)}, \underline{x}_{i(2)}) & Cov(\underline{y}_{i(1)}, \underline{y}_{i(2)}) \end{pmatrix}$$
$$= \begin{pmatrix} R(v_x, c_x) & 0 \\ 0 & R(v_y, c_y) \end{pmatrix} \qquad \text{(A-9)}$$

where the diagonal elements are $v_x = \rho_{ax} \sigma_{ax(1)} \sigma_{ax(2)} + \rho_{ex} \sigma_{ex(1)} \sigma_{ex(2)}$ and the off-diagonal elements are $c_x = \rho_{ax} \sigma_{ax(1)} \sigma_{ax(2)}$ for x, with similar expressions for the elements of y ($v_y = \rho_{ay} \sigma_{ay(1)} \sigma_{ay(2)} + \rho_{ey} \sigma_{ey(1)} \sigma_{ey(2)}$ and $c_y = \rho_{ay} \sigma_{ay(1)} \sigma_{ay(2)}$).

Using similar arguments as above, we can show that

$$Cov(\hat{\theta}_{(1)}, \hat{\theta}_{(2)}) = \sum_{u=1}^{3} \sum_{v=1}^{3} W_{uv} \varsigma_{uv(1,2)} \qquad \text{(A-10)}$$

where W=(W$_{uv}$) is defined by equation (A-2),

$$\varsigma_{uv(1,2)} = F_{\rho_{uv(1,2)}}(\delta_{z(1)}, \delta_{z(2)}) - \Phi\left(\delta_{z(1)}\right) \Phi\left(\delta_{z(2)}\right)$$

with $\rho_{uv(1,2)} = \gamma_{uv(1,2)} / \left(\sigma_{z(1)} \sigma_{z(2)}\right)$ and $\Gamma_{(1,2)} = \left(\gamma_{uv(1,2)}\right)$ is the 3x3 matrix defined by

$$\Gamma_{(1,2)} = \begin{pmatrix} v_x + v_y & v_x + c_y & v_x \\ c_x + v_y & c_x + c_y & c_x \\ v_y & c_y & 0 \end{pmatrix}$$

where v$_x$, c$_x$, v$_y$, and c$_y$ are defined below equation (A-9). Equation (A-10) shows that the covariance of the estimates $\left(\hat{\theta}_{(1)}, \hat{\theta}_{(2)}\right)$ involves the four within cluster correlation coefficients $\left(\rho_{ax}, \rho_{ay}, \rho_{ex}, \rho_{ey}\right)$ while the variance of the two components do not.