UW Biostatistics Working Paper Series

9-16-2004

A Marginal Model Approach for Analysis of Multi-reader Multi-test Receiver Operating Characteristic (ROC) Data

Xiao Song
University of Washington, songx@u.washington.edu

Xiao-Hua Zhou
University of Washington, azhou@u.washington.edu

1. Introduction

The Receiver Operating Characteristic (ROC) curve is the most commonly used index for assessing the accuracy of diagnostic imaging tests. In radiology, one common design for ROC studies involves multiple readers and multiple tests (Zhou et al, 2002). In such studies, all readers read all test results from the same patients. An example of such a study was conducted by Jiang et al (1999) to test whether computer-aided diagnosis (CAD) could improve radiologists' performance in breast cancer diagnosis. In the study, the 10 radiologists read mammograms of 104 patients using both computer-aided and unaided methods, and the response of the radiologist on the presence of malignant cancer had a continuous-scale. The true disease status of each patient was verified using a near-consecutive biopsy series. Among the 104 patients, 46 were malignant and 58 were benign. The objective was to compare the accuracy of computer-aided and unaided diagnoses.

The main complication for analysis of data obtained from such studies is that the test results from the same subject or the same reader may be correlated. Approaches that ignore the correlations might lead to erroneous conclusions. To deal with such data, several approaches have been proposed which assume that the estimated area (or other function) of the ROC plot follows a mixed-effects ANOVA model (Dorfman, Berbaum and Metz, 1992; Obuchowski and Rockette, 1995). Beiden, Wagner and Campell (2000) proposed a bootstrap approach under the same model without the normality assumption on the random effects but with the same correlation structure, which was further relaxed to allow different variances for different tests (2001). However, it might be difficult to check the assumptions on the correlation structure of these estimated measures in practice. Among these methods, the most widely used approach is the Dorfman-Berbaum-Metz (DBM) approach (Dorfman, Berbaum and Metz, 1992), which is regularly used in studies that the FDA and industry rely upon to quantify the benefits of new diagnostic and screening technologies. The DBM approach assumes a mixed-effects ANOVA model for the jackknife pseudovalues for the area under the ROC curve (AUC). Roe and Metz (1997) indicated by simulations that the DBM method works well for testing the null hypothesis that the diagnostic tests are equivalent when the response variable follows a standard mixed-effects ANOVA model with normal random effects and errors; however, they did not evaluate the performance of the DBM approach in estimation of the AUCs in situations of inequivalent tests. Moreover, there are other concerns for using the DBM method (Zhou et al., 2002). First, since pseudovalues are not real observations, the ANOVA model for pueudovalues does not have straightforward interpretation. Second, since pseudovalues are generally correlated, it lacks firm theoretical basis to utilize the standard inference procedures for ANOVA model.

In this paper, we first investigate the theoretical basis of the DBM method. Our results indicate that the DBM method does not satisfy the regular assumptions for standard mixed-effects ANOVA models in general because the variance of the response variable may vary across tests and subjects. Hence this approach might lead to erroneous inference. However, our theoretical results do show that correlations among the AUC jackknife pseudovalues from different subjects tend to zero as the number of subjects goes to infinity. We then propose a marginal regression model approach based on the AUCs. The estimators of the regression coefficients are consistent and asymptotically normal. Thus in contrast to the DBM approach, the marginal model approach has solid theoretical basis. We then compare the relative finite sample performance of our method with the DBM method via simulation studies. Our results show that our new method has similar coverage accuracy as the DBM method for the difference of two AUCs and for individual AUCs when the AUCs are the same for all tests but has better coverage accuracy for individual AUCs when the AUCs are different for different tests.

We organize the paper as follows. In Section 2, we describe the data structure for continuous outcomes. We investigate the theoretical basis for the DBM method in Section 3. The marginal model approach is given in Section 4. We compare the two approaches via simulation in Section 5 and by application to the breast cancer data in Section 6. We discuss the extension of the marginal model approach to ordinal outcomes in Section 7.

2. Data Structure

Suppose there are $K = n_0 + n_1$ subjects, of which n_0 are non-diseased and n_1 are diseased, I tests taken on each subject, and J readers each reading all test results from each subject. Let the subscripts $k = 1, \ldots, n_0$ denote non-diseased subjects and $k = n_0 + 1, \ldots, K$ denote diseased subjects. Let Y_{ijk} be the test value for subject k from the test i assigned by reader j, with larger values being more indicative of disease. In this paper, we consider the case that Y_{ijk} is continuous. Extensions to ordinal outcomes are discussed in Section 7. We assume that the test values from different subjects and different readers are independent, but they can be correlated if they are from the same reader or the same subject.

3. Existing DBM jackknife approach

The DBM approach computes the AUCs based on the ROC estimator under the binormal assumption (Dorfman and Alf, 1969), which is the maximum likelihood estimator (MLE) in the case of ordinal outcomes. Continuous outcome data need to be discretized before using this estimator, which is not MLE anymore in this case (Pepe, 2003). For simplicity, we use the nonparametric Wilcoxon estimator for the AUCs. We investigate the possible loss of efficiency of using the Wilcoxon estimator in Section 5.

Let A_i be the AUC for test i. Then $A_i = \Pr(Y_{ijk} > Y_{ijs})$ for $s = 1, \ldots, n_0$ and $k = n_0 + 1, \ldots, K$. The Wilcoxon estimator for A_i based on the observations from reader j is $A_{ij} = (n_0 n_1)^{-1} \sum_{s=1}^{n_0} \sum_{k=n_0+1}^{K} \varphi_{ijks}$, where $\varphi_{ijks} = I(Y_{ijk} > Y_{ijs})$ with $I(\cdot)$ being the indicator function; that is, φ_{ijks} is equal to 1 if $Y_{ijk} > Y_{ijs}$ and 0 otherwise. The corresponding jackknife pseudovalue is $A^*_{ijk} = KA_{ij} - (K-1)A_{ij(k)}$ $(k = 1, \ldots K)$, where $A_{ij(k)}$ is the "leave-1-out" estimator obtained by deleting subject k. The DBM method assumes that the pseudovalues A^*_{ijk} follow the ANOVA model

$$A^*_{ijk} = \mu^* + \alpha^*_i + R^*_j + C^*_k + (\alpha R)^*_{ij} + (\alpha C)^*_{ik} + (RC)^*_{jk} + \varepsilon^*_{ijk}, \tag{1}$$

where μ^* is the population mean, α^*_i is the fixed effect of test i, R^*_j is the random effect of reader j, C^*_k is the random effect of subject k, $(\alpha R)^*_{ij}$, $(\alpha C)^*_{ik}$ and $(RC)^*_{jk}$ are the corresponding two-way interactions, and ε^*_{ijk} is the random error. The random variables R^*_j, C^*_k, $(\alpha R)^*_{ij}$, $(\alpha C)^*_{ik}$, $(RC)^*_{jk}$ and ε^*_{ijk} are normally and independently distributed with mean zero and variances σ^2_{R*}, σ^2_{C*}, $\sigma^2_{\alpha R*}$, $\sigma^2_{\alpha C*}$, σ^2_{RC*} and $\sigma^2_{\varepsilon*}$, respectively. Standard techniques for ANOVA models are used for inference. For example, the F-test is used to test the fixed and random effects. The estimator for $\mu^* + \alpha^*_i$ is used to estimate A_i.

Roe and Metz (1997) conducted simulations using the DBM method. They assume the response variable Y_{ijk} follows the ANOVA model

$$Y_{ijk} = \mu_t + \alpha_{it} + R_j + C_{kt} + (\alpha R)_{ij} + (\alpha C)_{ikt} + (RC)_{jkt} + \varepsilon_{ijkt}, \tag{2}$$

where $t = I(k > n_0)$, μ_t is the population mean, α_{it} is the fixed effect of test i, R_j is the random effect of reader j, C_{kt} is the random effect of subject k, $(\alpha R)_{ij}$, $(\alpha C)_{ikt}$, and $(RC)_{jkt}$ are the corresponding two-way interactions, and ε_{ijkt} is the random error. The random variables R_j,C_{kt}, $(\alpha R)_{ij}$, $(\alpha C)_{ikt}$, $(RC)_{jkt}$, and ε_{ijkt} are normally and independently distributed with mean zero and variances σ^2_R, σ^2_{Ct}, $\sigma^2_{\alpha R}$, σ^2_{aCt}, σ^2_{RCt} and $\sigma^2_{\varepsilon t}$, respectively. Their simulation results indicate that the DBM method works well for testing the null hypothesis that the diagnostic tests are equivalent, i.e., $\alpha^*_i = 0$.

In Roe and Metz's simulation, they implicitly assume that (1) holds under (2). Notice that the following two conditions about the covariance structure hold under the standard mixed-effects ANOVA model (1): (i) $\text{corr}(A^*_{ijk}, A^*_{i'j'k'}) = 0$ for $i \neq i'$, $j \neq j'$, and $k \neq k'$; (ii) $\text{var}(A^*_{ijk}) = \sigma^2_{R*} + \sigma^2_{C*} + \sigma^2_{\alpha R*} + \sigma^2_{\alpha C*} + \sigma^2_{RC*} + \sigma^2_{\varepsilon*}$, which is a constant independent of i, j

and k. Now we check whether these conditions hold under model (2). With some tedious algebra, we can show that

$$\operatorname{cov}(A^*_{ijk}, A^*_{i'j'k'}) = \\ \left\{ \begin{array}{l} \frac{n_1-2K+2}{(n_0-1)^2 n_0} \left\{\theta_{ii'b11} + (n_0-1)\theta_{ii'b10} + (n_1-1)\theta_{ii'b01}\right\} + \frac{(K-1)^2}{(n_0-1)^2 n_1}\theta_{ii'b10}, \\ \qquad k \neq k', k \leq n_0, k' \leq n_0; \\ \frac{n_0-2K+2}{(n_1-1)^2 n_1} \left\{\theta_{ii'b11} + (n_0-1)\theta_{ii'b10} + (n_1-1)\theta_{ii'b01}\right\} + \frac{(K-1)^2}{(n_1-1)^2 n_0}\theta_{ii'b01}, \\ \qquad k \neq k', k > n_0, k' > n_0; \\ \frac{1}{(n_0-1)(n_1-1)} \left(1 - \frac{K-1}{n_0 n_1}\right) \left\{\theta_{ii'b11} + (n_0-1)\theta_{ii'b10} + (n_1-1)\theta_{ii'b01}\right\}, \\ \qquad k \neq k', k \leq n_0, k' > n_0; \text{ or } k > n_0, k' \leq 0; \\ \frac{n_1-2K+2}{(n_0-1)^2 n_0} \left\{\theta_{ii'b11} + (n_0-1)\theta_{ii'b10} + (n_1-1)\theta_{ii'b01}\right\} + \frac{(K-1)^2}{(n_0-1)^2 n_1} \left\{\theta_{ii'b11} + (n_1-1)\theta_{ii'b10}\right\}, \\ \qquad k = k' \leq n_0; \\ \frac{n_0-2K+2}{(n_1-1)^2 n_1} \left\{\theta_{ii'b11} + (n_0-1)\theta_{ii'b10} + (n_1-1)\theta_{ii'b01}\right\} + \frac{(K-1)^2}{(n_1-1)^2 n_0} \left\{\theta_{ii'b11} + (n_0-1)\theta_{ii'b01}\right\}, \\ \qquad k = k' > n_0; \end{array} \right. \tag{3}$$

where $\theta_{ii'bcd} = \operatorname{cov}(\varphi_{ijks}, \varphi_{i'j'k's'})$ with $b = I(j = j')$, $c = I(k = k')$ and $d = I(s = s')$. Expression (3) holds even without the normality assumption on the random effects and error in model (2). Hence both conditions (i) and (ii) assumed under the standard mixed-effects ANOVA model (1) are violated for finite samples in general. To have a better understanding of the covariance structure, we consider its asymptotic form when the number of subjects K goes to infinity. Write $\xi(K) = O\{\eta(K)\}$. If, for some positive constants v and K_0, $|\xi(K)| \leq v|\eta(K)|$ when $K \geq K_0$. By assuming $O(n_0) = O(n_1)$, we can show that

A. $\operatorname{corr}(A^*_{ijk}, A^*_{i'j'k'}) = O(K^{-1})$ for $k \neq k'$;

B. $$\lim_{K\to\infty} \operatorname{var}(A^*_{ijk}) = \lim_{K\to\infty} \left\{ I(k \leq n_0) \left(\frac{K}{n_0}\right)^2 \theta_{ii110} + I(k > n_0) \left(\frac{K}{n_1}\right)^2 \theta_{ii101} \right\}, \tag{4}$$
which depends on i and k;

C. a sufficient condition for (4) to be independent of i and k is that the random effects and error in model (2) are normal and $\mu_0 = \mu_1$, $\alpha_{i0} = \alpha_{i1}$, $\sigma^2_{C0} = \sigma^2_{C1}$, $\sigma^2_{\alpha C0} = \sigma^2_{\alpha C1}$, $\sigma^2_{RC0} = \sigma^2_{RC1}$, $\sigma^2_{\varepsilon 0} = \sigma^2_{\varepsilon 1}$ and $\lim_{K\to 0} n_0/K = 1/2$.

The proof is sketched in the Appendix.

Result (A) indicates that the correlations among the pseudovalues from different subjects are asymptotically equal to 0 as the number of subjects goes to infinity and hence condition (i) holds asymptotically, which might explain the unbiased estimation for the AUCs observed in our simulations in Section 5. However, result (B) implies that the variances of the pseudovalues A^*_{ijk} can differ across tests and subjects when the tests are not equivalent or when the sample sizes for diseased and non-diseased subjects are not equivalent; that is, condition (ii) can be violated even for large samples. This deviation from condition (ii) might lead to biased variance estimators, as we illustrate in Section 5.

4. Marginal Model Approach

We propose a marginal generalized linear model for the AUCs which allows to include types of tests and other covariates in the model; our regression model is an extension of the marginal regression model for the AUCs for independent ROC data, proposed by Dodd and Pepe (2003), to multi-reader multi-test ROC data. Let X^1_k denote covariates for diseased subject k $(k = n_0 + 1, \ldots, K)$, X^0_s denote covariates for non-diseased subject s $(s = 1, \ldots, n_0)$, and Q_j denote covariates for reader j $(j = 1, \ldots, J)$. Let $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{iI})^T$ $(i = 1, \ldots, I)$, where $Z_{ir} = I(i = r)$ is the indicator for the test r. Here Q_j's are independent across j, X^1_k's are independent across k, X^0_s's are independent across s, and Z_i's are independent across i. Following Dodd and Pepe (2003), we define the covariate-specific AUC A_{ijks} as

$$A_{ijks} = E(\varphi_{ijks}|Z_i, Q_j, X^1_k, X^0_s) = P(Y_{ijk} > Y_{ijs}|Z_i, Q_j, X^1_k, X^0_s).$$

Then we propose the following regression model for A_{ijks}:

$$A_{ijks} = g(\beta_1^T Z_i + \beta_2^T Q_j + \beta_3^T X^1_k + \beta_4^T X^0_s), \qquad (5)$$

where $g(\cdot)$ is a monotone link function, and $\beta = (\beta_1^T, \beta_2^T, \beta_3^T, \beta_4^T)^T$ is a vector of regression parameters. Under model (5), we have the regression model for φ_{ijks}:

$$\Pr\left(\varphi_{ijks} = 1 | Z_i, Q_j, X_k^1, X_s^0\right) = g(\beta_1^T Z_i + \beta_2^T Q_j + \beta_3^T X_k^1 + \beta_4^T X_s^0). \tag{6}$$

For this marginal model, the set of "observations" is $\{(\varphi_{ijks}, Z_i, Q_j, X_k^1, X_s^0) : i = 1, \ldots, I; j = 1, \ldots, J; k = n_0+1, \ldots, K; s = 1, \ldots, n_0\}$. Since φ_{ijks} are not independent, standard methods for generalized linear models can not be applied directly. We consider three different assumptions on the correlation structure. First, as conforming to the ANOVA model (2), we assume that φ_{ijks} and $\varphi_{i'j'k's'}$ are correlated only when $k = k'$ or $s = s'$. Then φ_{ijks} are sparsely correlated as defined by Lumley (1998) and Lumley and Hamblett (2003). In their notation, for each "observation" φ_{ijks}, we define the set $S_{ijks} = \{(i', j', k', s') : k' = k \text{ or } s' = s\}$, which contains the indices of all "observations" correlated to φ_{ijks}. It is easy to see that the number of "observations" in S_{ijks} is $M = IJ(K-1) = O(IJn_0 + IJn_1)$. Now consider a subset $\mathcal{T}$ of $\{(i, j, k, s) : i = 1, \ldots, I; j = 1, \ldots, J; k = n_0 + 1, \ldots, K; s = 1, \ldots, n_0\}$ which satisfies that, for any two elements $(i', j', k', s') \in \mathcal{T}$ and $(i'', j'', k'', s'') \in \mathcal{T}$, $(i', j', k', s') \notin S_{i''j''k''s''}$ and $(i'', j'', k'', s'') \notin S_{i'j'k's'}$. Thus any two elements in $\mathcal{T}$ must have different k and s. Hence the maximum number of elements in $\mathcal{T}$ is $m = \min(n_0, n_1)$. Therefore $Mm = O(IJn_0n_1)$. Noticing that IJn_0n_1 is the number of "observations", we can conclude that the condition of sparse correlation is satisfied. This assumption can be relaxed to that φ_{ijks} and $\varphi_{i'j'k's'}$ are correlated only when $j = j'$ or $k = k'$ or $s = s'$; in this case, $S_{ijks} = \{(i', j', k', s') : j' = j \text{ or } k' = k \text{ or } s' = s\}$, $M = O(In_0n_1 + IJn_0 + IJn_1)$ and $m = \min(J, n_0, n_1)$. We can further relax the assumption to that φ_{ijks} and $\varphi_{i'j'k's'}$ are correlated only when $i = i'$ or $j = j'$ or $k = k'$ or $s = s'$; the corresponding $S_{ijks} = \{(i', j', k', s') : i' = i \text{ or } j' = j \text{ or } k' = k \text{ or } s' = s\}$, $M = O(Jn_0n_1 + In_0n_1 + IJn_0 + IJn_1)$ and $m = \min(I, J, n_0, n_1)$. Notice that under each of these conditions we always have $Mm = O(IJn_0n_1)$ and hence the data have sparse correlation structure. To discriminate these assumptions, we call them assumption I, II and

III, respectively.

Now consider the pseudo-likelihood

$$L = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{s=1}^{n_0}\prod_{k=n_0+1}^{K} \left\{g(\beta^T W_{ijks})\right\}^{\varphi_{ijks}} \left\{1 - g(\beta^T W_{ijks})\right\}^{1-\varphi_{ijks}}, \tag{7}$$

where $W_{ijks} = (Z_i^T, Q_j^T, X_k^{1T}, X_s^{0T})^T$. Expression (7) is the likelihood when φ_{ijks} are independent. The log pseudo-likelihood equation is

$$U = \frac{\partial \log L}{\partial \beta^T} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{s=1}^{n_0}\sum_{k=n_0+1}^{K} U_{ijks}(\beta) = 0, \tag{8}$$

where

$$U_{ijks}(\beta) = \left[\frac{\left\{\varphi_{ijks} - g(\beta^T W_{ijks})\right\} g'(\beta^T W_{ijks})}{g(\beta^T W_{ijks})\left\{1 - g(\beta^T W_{ijks})\right\}}\right] W_{ijks},$$

$g'(\cdot)$ is the derivative of $g(\cdot)$. Let $\hat{\beta}$ be the solution to (8) that maximizes (7). By Theorem 7 of Lumley and Hamblett (2003), under some regularity conditions, as $m \to \infty$, $\hat{\beta}$ is consistent and asymptotically normal with variance consistently estimated by $C^{-1}B(C^{-1})^T$, where

$$C = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{s=1}^{n_0}\sum_{k=n_0+1}^{K} \frac{\partial U_{ijks}(\hat{\beta})}{\partial \beta^T},$$

$$B = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{s=1}^{n_0}\sum_{k=n_0+1}^{K}\sum_{(i',j',k',s')\in S_{ijks}} U_{ijks}(\hat{\beta})U^T_{i'j'k's'}(\hat{\beta}).$$

A consistent estimator for the AUCs can then be obtained by substituting $\hat{\beta}$ for β in (5). In practice, to utilize the asymptotic normality, m is required to be large. Under assumption I, this corresponds to that both the number of diseased subjects and the number of non-diseased subjects are large. Assumption II further requires that the number of readers is large and assumption III in addition requires that both the number of tests and the number of readers are large. Hence assumption I is more reasonable in practice.

It is easy to see that the mixed-effects ANOVA model (2) is a special case of the proposed marginal model, that is,

$$\Pr\left(\varphi_{ijks} = 1 | Z_i, Q_j, X_k, X_s\right) = \phi(\beta_1^T Z_i),$$

where $\phi(\cdot)$ is the distribution function for the standard normal, and $\beta_1 = (\mu_1 - \mu_0 + \alpha_{11} - \alpha_{10}, \ldots, \mu_1 - \mu_0 + \alpha_{I1} - \alpha_{I0})^T / \left\{\sum_{t=0}^{1} \left(\sigma_{Ct}^2 + \sigma_{\alpha Ct}^2 + \sigma_{RCt}^2 + \sigma_{\varepsilon t}^2\right)\right\}^{1/2}$. Compared to the jackknife pseudovalue model (1), the marginal model (6) has advantage in interpretation, since the dependent variable is the more meaningful covariate-specific AUC; for example, the regression parameter for Z_{ir} indicates the effect of test r on the AUC, with larger values imply better accuracy. In addition, it is easy to incorporate covariates in (6). The asymptotic properties provide sound theoretical basis for statistical inference. Since the estimating equation (8) has the same form as that for the generalized estimating equation (GEE) with independent working correlation structure, the estimates can be obtained by standard statistics softwares such as SAS and SPlus, although the standard errors need to be recomputed using the sandwich estimator described above. We have written the code for computing the standard errors under the logit and probit links using PROC IML in SAS, which can be easily extended to other links or converted to SPlus. Therefore the implementation of the marginal model approach includes three steps: i) derive the data set $\{\varphi_{ijks}, Z_i, Q_j, X_k, X_s\}$ from the original data; ii) obtain the estimates using PROC GENMOD or PROC LOGISTIC in SAS or function gee() or glm() in SPlus; iii) compute the standard errors using self-coded functions based on the sandwich estimator and the output obtain from step ii).

5. Simulation Studies

To compare the performance of the marginal model approach and the DBM approach, we conducted simulations under scenarios similar to those in Roe and Metz (1997). We generated the data according to model (2), with $I = 2$, $J = 5$, $\mu_0 = 0$, $\mu_1 = 0.75$ or 1.5,

$\alpha_{10} = \alpha_{11} = \alpha_{20} = 0$, σ^2_R =0.0055, σ^2_{Ct} =0.3, $\sigma^2_{\alpha R}$ =0.0055, σ^2_{aCt} =0.3, σ^2_{RCt} =0.2, $\sigma^2_{\varepsilon t} = 0.2$ $(t = 0, 1)$. For $\mu_1 = 0.75$ and 1.5, $A_1 = 0.702$ and 0.856, respectively. We considered two scenarios representing equal and unequal accuracy of tests, respectively; that is, $\alpha_{21} = 0$ and 1. When $\alpha_{21} = 0$, the AUCs are the same for the two tests, that is, $A_1 = A_2$; when $\alpha_{21} = 1$, $A_2 - A_1 = 0.190$ and 0.106 for $\mu_1 = 0.75$ and 1.5, respectively.

We carried out simulations for $n_0 = n_1 = 20$ and 50. For each scenario, 500 Monte Carlo data sets were simulated. Table 1 presents the results of estimation for A_1 and $A_2 - A_1$ using both the DBM method based on the Wilcoxon estimator for the AUCs and the marginal model method. For both methods, the coverage probabilities of the 95% confidence intervals were computed; the confidence intervals were obtained based on the t-statistics for the DBM approach and the Wald statistics for the marginal model approach. In the case when $\alpha_{21} = 0$, both estimators are close to the truth and the standard errors track the empirical standard deviations well. The coverage probabilities for both approaches are close to the nominal level, which tend to be larger for the DBM approach. In the case when $\alpha_{21} = 1$, the performance of the marginal model approach is similar to that when $\alpha_{21} = 0$. In contrast, for the DBM estimator, the inference on $A_2 - A_1$ works well while the standard error for A_1 is much less than the corresponding empirical standard deviation and the coverage probability is well below the nominal level, which tends to decrease with increased number of patients. A possible explanation is that the variances of the pseudovalues vary across the tests in this case.

In summary, for inference on the difference of the AUCs, the performance of the marginal model approach is comparable to that of the DBM approach. However, the marginal model method has better performance for inference on individual AUCs when the accuracy of tests are not equivalent although their performances are comparable in the case of equal accuracy. To assess possible loss of efficiency by using the Wilcoxon estimator for AUCs in the DBM approach, we also computed AUCs parametrically assuming both the diseased and non-diseased distributions of Y are normal; that is, the AUC estimates are based on the MLEs of the normal parameters. The results are similar with only slight loss of efficiency for those using the Wilcoxon estimator (not shown).

6. Application

We apply the marginal model approach and the DBM approach to the breast cancer data (Jiang et al., 1999) described in Section 1. In this case, there were $I = 2$ tests (computer-aided and unaided diagnoses), $J = 10$ readers (radiologists), $n_0 = 58$ non-diseased (benign) and $n_1 = 46$ diseased (malignant) subjects. For each diagnosis, the radiologists were asked to give their degree of suspicion that a lesion was malignant by reading the mammagrams and then placing a mark on a 5-cm line labeled "benign" at the left end and "malignant" on the right end. These marks were then converted to numerical scores with a ruler. Here the scores were the observed test results. For the computer-aided diagnosis, the radiologists were given an additional computer-estimated likelihood of malignancy based on eight computer-extracted image features from the mammagrams.

To compare the computer-aided and unaided diagnoses, we estimated their AUCs and the difference between them using both the DBM approach and the marginal model approach. Since the AUC for each diagnostic method is the expectation of the reader-specific AUCs, the estimated AUC can be viewed as the estimate of mean of the AUCs for the 10 readers in this case. The results are shown in table 2. For both approaches, the estimates for the AUCs are very close and the lower bounds of the 95% confidence intervals for the AUC difference were greater than 0. It seems that CAD can improve radiologists' performance in breast cancer diagnosis.

7. Discussion

For multi-reader multi-test ROC data, the correlated data structure makes the analysis more complicate than that for the independent case. One popular analytic method for such data is the DBM jackknife method that assumes that the pseudovalues for areas under ROC curves follow a standard mixed-effects ANOVA model, and this method has been found widely used in practice. In this paper, we have conducted a theoretical study on the validity of the DBM method and have explored situations when the DBM method is appropriate and situations when the DBM method may lead to erroneous inference. We have also proposed a new marginal model approach for the analysis of multi-reader multi-test ROC data. Our new approach has the advantage in interpretation, sound theoretical foundation, and good finite sample performance. Our method is also computationally simpler than the DBM jackknife method. When we implemented both approaches in SAS for the breast cancer data described in Section 6, the marginal model approach was about 4 times faster than the DBM approach.

This paper has focused on continuous test outcome Y_{ijk}. Like DBM approach, the marginal model approach can be extended to ordinal outcomes as well. If Y_{ijk} is ordinal, then $A_i = \Pr(Y_{ijk} \geq Y_{ijs})$. We define $\varphi_{ijks1} = I(Y_{ijk} > Y_{ijs})$ and $\varphi_{ijks2} = I(Y_{ijk} = Y_{ijs})$, $s = 1, \ldots, n_0$, $k = n_0 + 1, \ldots, K$; alternatively, we can define $\varphi_{ijks2} = I(Y_{ijk} \geq Y_{ijs})$. Assume the following marginal model

$$\Pr\left(\varphi_{ijksr} = 1 | Z_i, Q_j, X_k^1, X_s^0\right) = g(\beta_{1r}^T Z_i + \beta_{2r}^T Q_j + \beta_{3r}^T X_k^1 + \beta_{4r}^T X_s^0).$$

Since the set of "observations" $\{(\varphi_{ijksr}, Z_i, Q_j, X_k^1, X_s^0) : r = 1, 2; i = 1, \ldots, I; j = 1, \ldots, J; k = n_0 + 1, \ldots, K; s = 1, \ldots, n_0\}$ is sparsely correlated, the estimation and inference procedure will be similar to that for the continuous test outcomes.

Acknowledgements

We would like to Thank Dr. Yulei Jiang for providing us with the breast cancer data set used in this paper. We would also like to thank an anonymous referee for many helpful comments that resulted in an improved version of this manuscript. This work is supported in part by NIH grant AHRQ R01HS013105-01 (Zhou), U01-AI46702 (Song), and career development funds from the University of Washington Department of Biostatistics (Song).

References

Beiden, S.V., Wagner, R.F. and Campbell, G. (2000). Components-of-variance models and multiple bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis. *Academic Radiology* **7**, 341–349.

Beiden, S.V., Wagner, R.F., Campbell, G., Metz, C.E. and Jiang, Y. (2001). Components-of-variance models for random-effects ROC analysis. *Academic Radiology* **8**, 605–615.

Dorfman, D.D. and Alf E. Jr. (1969). Maximum likelihood estimation of parameters of signal detection theory and dertermination of confidence interevals — rating method data. *Journal of Mathematical Psychology* **6**, 487–496.

Dorfman, D.D., Berbaum, K.S. and Metz, C.E. (1992). Receiver operating characteristic rating analysis: generaliztion to the population of readers and patients with the jackknife method. *Investigative Radiology* **27**, 723–731.

Dodd, L.E. and Pepe, M.S. (2003). Partial AUC estimation and regression. *Biometrics* **59**, 614-623.

Jiang, Y., Nishikawa, R.M., Schmidt, R.A., Metz, C.E., Giger M.L. and Doi, K. (1999). Improving breast cancer diagnosis with computer-aided diagnosis. *Academic Raiology* **6**, 22–33.

Lumley, T. (1998). Marginal regression modelling of weakly dependent data. *Ph.D Dissertation*, Department of Biostatistics, University of Washington, Seattle, WA.

Lumley, T. and Hamblett, N. M. (2003). Asymptotics for marginal generalized linear models with sparse correlations. *Annals of Statistics*, re-submitted, a preprint copy available at http://www.bepress.com/uwbiostat/paper207/.

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

Obuchowski, N.A. and Rockette, H.E. (1995). Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations. *Communications in Statistics - Simulations* **24**, 285–308.

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

Pepe, M.S. (1998). Three approaches to regression analysis of receiver operating characteristic curves in medical diagnostic testing. *Biometrics* **54**, 124–135.

Pepe, M.S. (2000). Receiver operating characteristic methodology. *Journal of the American Statistical Association* **95**, 308–311.

Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press.

Roe, C.A. and Metz, C.E. (1997). Variance-component modeling of reciver operating characteristic index estimates. *Academic Radiology* **4**, 587–600.

Zhou, X.H., McClish, D.K. and Obuchowski, N.A. (2002). *Statistical Methods in Diagnostic Medicine.* Wiley and Sons, New York.

Appendix

By straightforward algebra, the pseudovalues A^*_{ijk} can be represented through φ_{ijks}, that is,

$$A^*_{ijk} = \begin{cases} \frac{n_0-K}{(n_0-1)n_0n_1} \sum_{s=1}^{n_0} \sum_{s'=n_0+1}^{K} \varphi_{ijss'} + \frac{K-1}{(n_0-1)n_1} \sum_{s=n_0+1}^{K} \varphi_{ijsk}, & k = 1, \ldots, n_0; \\ \frac{n_1-K}{(n_1-1)n_0n_1} \sum_{s=1}^{n_0} \sum_{s'=n_0+1}^{K} \varphi_{ijss'} + \frac{K-1}{(n_1-1)n_0} \sum_{s=1}^{n_0} \varphi_{ijks}, & k = n_0+1, \ldots, K. \end{cases}$$

Note that $\varphi_{ijks} = I(\delta_{ijks} > 0)$, where

$$\begin{aligned}\delta_{ijks} &= \mu_1 - \mu_0 + \alpha_{i1} - \alpha_{i0} + C_{k1} - C_{s0} \\ &\quad + (\alpha C)_{ik1} - (\alpha C)_{is0} + (RC)_{jk1} - (RC)_{js0} + \varepsilon_{ijk1} - \varepsilon_{ijs0}.\end{aligned}$$

We can write $\text{cov}(\varphi_{ijks}, \varphi_{i'j'k's'}) = \theta_{ii'bcd}$, where $b = I(\ j = j')$, $c = I(k = k')$ and $d = I(s = s')$. It is easy to see that $\theta_{ii'b00} = 0$. Then (3) follows with some simple but tedious algebra. Note that the normality assumption is not required here. Results (A) and (B) follow immediately from (3).

Now we consider the normal case. It is easy to see that $(\delta_{ijks}, \delta_{i'j'k's'})^T$ is normally distributed with mean $(\mu_1 - \mu_0 + \alpha_{i1} - \alpha_{i0}, \mu_1 - \mu_0 + \alpha_{i'1} - \alpha_{i'0})^T$ and variance

$$\Sigma_{ijks,i'j'k's'} = V \begin{pmatrix} 1 & \rho_{abcd} \\ \rho_{abcd} & 1 \end{pmatrix},$$

where $V = \sum_{t=0}^{1} (\sigma^2_{Ct} + \sigma^2_{\alpha Ct} + \sigma^2_{RCt} + \sigma^2_{\varepsilon t})$, $a = I(i = i')$,

$$\begin{aligned}\rho_{abcd} &= V^{-1} \left[\sigma^2_{C1} I(k = k') + \sigma^2_{C0} I(s = s')\right. \\ &\quad + \sigma^2_{\alpha C1} I(i = i', k = k') + \sigma^2_{\alpha C0} I(i = i', s = s') \\ &\quad + \sigma^2_{RC1} I(j = j', k = k') + \sigma^2_{RC0} I(j = j', s = s') \\ &\quad \left. + \sigma^2_{\varepsilon 1} I(i = i', j = j', k = k') + \sigma^2_{\varepsilon 0} I(i = i', j = j', s = s')\right].\end{aligned}$$

Hence

$$\begin{aligned}\theta_{abcd} &= \Pr(\delta_{ijks} > 0, \delta_{i'j'k's'} > 0) - \Pr(\delta_{ijks} > 0) \Pr(\delta_{i'j'k's'} > 0) \\ &= \begin{cases} \phi(h_{ijks}) - \phi^2(h_{ijks}), & i = i',\ b = c = d = 1 \\ \phi_2(h_{ijks}, h_{i'j'k's'}, \rho_{abcd}) - \phi(h_{ijks})\phi(h_{ijk's'}), & \text{otherwise,} \end{cases}\end{aligned}$$

where $\phi(\cdot)$ and $\phi_2(\cdot)$ are the standard univariate and bivariate normal distribution functions, respectively, and $h_{ijks} = V^{-1/2}(\mu_1 - \mu_0 + \alpha_{i1} - \alpha_{i0})$. Note when $\mu_0 = \mu_1$, $\alpha_{i0} = \alpha_{i1}$, $\sigma^2_{C0} = \sigma^2_{C1}$, $\sigma^2_{\alpha C0} = \sigma^2_{\alpha C1}$, $\sigma^2_{RC0} = \sigma^2_{RC1}$, and $\sigma^2_{\varepsilon 0} = \sigma^2_{\varepsilon 1}$, we have $h_{ijks} = 0$ and $\rho_{ab10} = \rho_{ab01}$. Hence $\theta_{ii110} = \theta_{ii101}$. Then (C) follows.

Table 1

Simulation results. MM, marginal model approach; B, bias; SD, empirical standard deviation across simulated data sets; SE, average of estimated standard errors; CP, coverage probability of the 95% confidence interval; CL, length of the 95% confidence interval.

			$\alpha_{21} = 0$				$\alpha_{21} = 1$			
			$n_0 = 20$		$n_0 = 50$		$n_0 = 20$		$n_0 = 50$	
μ_1			DBM	MM	DBM	MM	DBM	MM	DBM	MM
0.75	A_1	B	0.002	0.002	0.000	0.000	0.002	0.002	0.000	0.000
		SD	0.065	0.064	0.045	0.045	0.064	0.064	0.045	0.045
		SE	0.068	0.063	0.043	0.041	0.056	0.063	0.035	0.041
		CP	0.957	0.942	0.944	0.926	0.904	0.942	0.878	0.926
	$A_2 - A_1$	B	-0.002	-0.003	0.002	0.002	-0.002	-0.002	0.000	0.000
		SD	0.068	0.067	0.043	0.043	0.058	0.058	0.039	0.038
		SE	0.069	0.064	0.043	0.042	0.061	0.056	0.039	0.037
		CP	0.959	0.926	0.962	0.944	0.944	0.936	0.954	0.934
1.5	A_1	B	0.002	0.002	0.001	0.000	0.002	0.002	0.000	0.000
		SD	0.047	0.047	0.032	0.032	0.047	0.047	0.032	0.032
		SE	0.048	0.043	0.030	0.028	0.036	0.043	0.023	0.028
		CP	0.940	0.928	0.942	0.932	0.862	0.928	0.844	0.932
	$A_2 - A_1$	B	-0.002	-0.002	0.001	0.001	-0.002	-0.001	0.000	0.000
		SD	0.049	0.049	0.030	0.030	0.043	0.043	0.028	0.028
		SE	0.050	0.045	0.031	0.030	0.043	0.039	0.027	0.026
		CP	0.964	0.938	0.972	0.954	0.938	0.912	0.942	0.918

Table 2

Results for the breast cancer data. MM, marginal model approach; Est, Estimate; SE, standard error; CI, 95% confidence interval.

	DBM			MM		
AUC	Est	SE	CI	Est	SE	CI
Unaided	0.597	0.038	(0.522, 0.671)	0.597	0.036	(0.524, 0.666)
With Aid	0.742	0.038	(0.667, 0.816)	0.741	0.036	(0.666, 0.807)
With Aid – Unaided	0.145	0.035	(0.076, 0.214)	0.145	0.034	(0.079, 0.211)