Regression Model for the AUC of Clustered Ordinal Test Results and Working Independent Optimal Weights

Johan Lim, Woojoo Lee, Sin-Ho Jung, Kyeong Eun Lee,
and Sung-Cheol Yun [*]

Abstract

We study a regression model on the area under the receiver operating characteristic curves (AUC) for clustered (or repeatedly measured) test results. To account for cluster information, we consider a weighted estimating equation for Dodd and Pepe (2003)'s regression model with working independence weights. We find the optimal weight in the given class of working independence weights to minimize the variance (or MSE) of regression estimators. We apply the proposed procedure to analyzing our recent experiment on diagnosing a liver disorder. In the experiment, we investigated MRI images of patients having symptoms of potential liver disorder to compare the performance of different MRI picturing methods in testing for liver disorders.

Key words: Area under curve, clustered ordinal results, generalize estimating equation, optimal weight.

1 Introduction

The receiver operating characteristic (ROC) curve plots two accuracy measures of diagnostic tests: the false positive rate (FPR) and the true positive rate (TPR). FPR is the probability of falsely identifying a non-disease (negative) group, say ND, as a disease (positive) group, say D. TPR is the probability of correctly classifying a subject from the disease group. For

[*] Johan Lim and Woojoo Lee are at Department of Statistics, Seoul National University, Seoul, 151-747, Korea (Email: johanlim,lwj221@snu.ac.kr). Sin-Ho Jung is at Department of Biostatistics and Bioinformatice and CALGB Statistical Center, Duke University, NC, USA (Email:sinho.jung@duke.edu). Kyeong Eun Lee is at Department of Statistics, Kyungpook National University, Dae-Gu, Korea (Email: artlee@knu.ac.kr). Sung-Cheol Yun is at Department of Preventive Medicine, College of Medicine, University of Ulsan and at Aasan Medical Center, Seoul, Korea (Email: ysch97,leems@amc.seoul.kr). Johan Lim is the corresponding author.

example, suppose Y is the continuous result of a diagnostic test, where, for a cutoff value c, $Y \geq c$ implies classifying a subject to the disease (positive) group and $Y < c$ implies classifying him to the disease (negative) group. Then, the FPR is $\mathrm{P}\big(Y \geq c \big| \mathrm{ND}\big)$ and the TPR is $\mathrm{P}\big(Y \geq c \big| \mathrm{D}\big)$. The ROC curve plots these FPRs and TPRs for various choices of c.

The performance of diagnostic test often depends on covariates such as gender and age as in the hearing impairment example in Dodd and Pepe (2003). The covariate dependent ROC models are studied extensively in the previous literature to take into account the covariate effect. A common approach is to introduce covariate dependent error distributions. For example, Tosteson and Begg (1988) model the survival functions of ND and D as a function of covariates directly and calculate the induced covariate-specific ROC curve. Similar approaches are taken by many other authors including Le (1997), Pepe (1998), Pepe (2000), Alonzo and Pepe (2002), Cai and Pepe (2002), Faraggi (2003), and Schisterman et al. (2004).

Another approach is the regression model for the area under the ROC curve (AUC). The AUC is a single index performance measure of medical diagnostics. When the diagnostic result is continuous, it is equivalent to the probability that the result from the disease group Y^{D} is larger than that of the non-disease group Y^{ND}. Using this result, Dodd and Pepe (2003) propose a regression model for the AUC, and estimate it by solving the generalized estimating equations (GEE) constructed by pairing test results of the non-disease group with that of the disease group. To be specific, suppose $\big(Y_i^{\mathrm{D}}, X_i^{\mathrm{D}}\big)$ and $\big(Y_j^{\mathrm{ND}}, X_j^{\mathrm{ND}}\big)$ are observations from D and ND for $i = 1, 2, \ldots, n_{\mathrm{D}}$, and $j = 1, 2, \ldots, n_{\mathrm{ND}}$, respectively. They propose a regression model $g\big(\theta_{ij}\big) = \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}$, where $\mathbf{X}_{ij} = \big(X_i^{\mathrm{D}}, X_j^{\mathrm{ND}}\big)^{\mathrm{T}}$, $\theta_{ij} = P\big(Y_i^{\mathrm{D}} > Y_j^{\mathrm{ND}} \big| X_i^{\mathrm{D}}, X_j^{\mathrm{ND}}\big)$, and g is the probit or the logit function. It is worth noting that Dodd and Pepe (2003) do not assume repeated observations from a single subject or a patient.

In this paper, we are interested in the regression model for clustered (or repeatedly measured) test results which often arises in many medical practices (Zeltzer et al., 1991; Tang and Balakrishnan, 2011). The analysis of clustered results (or repeatedly measured data) are well understood under the conventional regression model. For example, Liang and Zeger (1986) propose a generalized estimating equation (GEE) method to estimate regression parameters and point out that disregarding the cluster information results in inefficiency of the estimators and biased variance estimate. The AUC of clustered ordinal results is previously studied by Obuchowski (1997). She estimates the AUC of each method using the Mann-Whitney type statistics and compares two diagnostic methods. In doing so, she uses working independence weights which are equal to all lesions (or observations).

We modify the regression model for the AUC by Dodd and Pepe (2003) to account for the cluster information. Dodd and Pepe (2003) construct generalized estimating equations (GEE) by pairing test results of the non-disease group with that of the disease group to estimate the model. In this paper, we propose to use a weighted GEE to take the cluster information into account. The optimal weight for the weighted GEE is known to be the inverse covariance matrix of estimating equations in a cluster; however, it is unfeasible to get the covariance matrix of estimating equations of randomly paired test results.

We consider a class of working independence weights in constructing GEE to incorporate the cluster information. It is a convex set of two special weights, equal weights to observations (lesions in our example) $\mathbf{w}^{\mathrm{L}}$ and equal weights to clusters (subjects or patients in our example) $\mathbf{w}^{\mathrm{S}}$. These two weights are shown to be optimal in opposite situations in conventional repeatedly measured data analysis (Ju, 2004). To be specific, when there exists a strong dependency among observations within a cluster, the optimal independent weight is known to be the equal weights to clusters. On the other hand, if observations within a cluster are independent or nearly independent, the optimal independent weight assigns equal weight to observations. We suggest to use the weight in the given class to minimize the variance (or the mean squared error, MSE) of regression coefficient estimators. We call the weight to achieve the minimum as "optimal independent weight" and use a bootstrap procedure to compute it.

The paper is organized as follows. In Section 2, we introduce an estimating equation approach for clustered ordinal test results. We further introduce the class of working independence weights we consider and propose a bootstrap procedure to find the optimal independent weights. In Section 3, we numerically investigate the aforementioned relationship between the optimal independent weights and the dependency among observations within a cluster. It shows that, when the observations within a cluster are strongly dependent (or independent) to each other, the independent optimal weight is close to $\mathbf{w}^{\mathrm{S}}$ (or $\mathbf{w}^{\mathrm{L}}$). In the section, we also investigate the performance of the bootstrap procedure we use in this paper. In Section 4, we apply the model to analyzing the liver disorder data set which motivates the paper. We briefly summarize the paper in Section 5.

2 Model and estimation

In this section, we propose the regression model and its estimation for the clustered test results. We assume that test results are ordinal which is more common than binary results in many medical diagnostics. We also assume that all covariates are discrete and its extension to continuous covariates can be done without much additional efforts as in Dodd and Pepe (2003).

Suppose we observe $\{(Y_{ik}^{\mathrm{D}}, X_{ik}^{\mathrm{D}}\}$ for $i = 1, 2, \ldots, n_{\mathrm{D}}$, $k = 1, 2, \ldots, n_i$, from the disease population, and $\{(Y_{jl}^{\mathrm{ND}}, X_{jl}^{\mathrm{ND}})\}$ for $j = 1, 2, \ldots, n_{\mathrm{ND}}$ and $l = 1, 2, \ldots, n_j$ from the non-disease population. Given covariates x with $X_{ik}^{\mathrm{D}} = X_{jl}^{\mathrm{ND}} = x$, the covariate specific AUC is

$$\mathrm{AUC}(x) = \mathrm{P}\big(Y_{ik}^{\mathrm{D}} > Y_{jl}^{\mathrm{ND}} | x\big) + \big(1/2\big)\mathrm{P}\big(Y_{ik}^{\mathrm{D}} = Y_{jl}^{\mathrm{ND}} | x\big). \tag{1}$$

We propose the model that

$$g\big\{\mathrm{AUC}(x)\big\} = \beta_0 + \beta_1^T u(x), \tag{2}$$

where g is a link function such as the logistic or the probit function, and $u(x)$ is the covariate vector.

To estimate the model, we introduce new variables and a set of estimating equations. We define new variables U_{ijkl} for $\{(i,k),(j,l)\}$ with $i \in \mathrm{D}$, $j \in \mathrm{ND}$, and $X_{ik}^{\mathrm{D}} = X_{jl}^{\mathrm{ND}} = x$ as:

$$U_{ikjl} = \begin{cases} 1 & \text{if } Y_{ik}^{\mathrm{D}} > Y_{jl}^{\mathrm{ND}} \\ 1/2 & \text{if } Y_{ik}^{\mathrm{D}} = Y_{jl}^{\mathrm{ND}} \\ 0 & \text{if } Y_{ik}^{\mathrm{D}} < Y_{jl}^{\mathrm{ND}} \end{cases},$$

whose mean is $\theta_{ikjl}(x) \equiv \mathrm{E}\big(U_{ikjl}\big|x\big) = \mathrm{P}\big(Y_{ik}^{\mathrm{D}} > Y_{jl}^{\mathrm{ND}}\big|x\big) + \big(1/2\big)\mathrm{P}\big(Y_{ik}^{\mathrm{D}} = Y_{jl}^{\mathrm{ND}}\big|x\big) = \mathrm{AUC}(x)$. Thus, $g\big(\theta_{ikjl}(x)\big) = \beta_0 + \beta_1^T u(x)$.

We estimate β_0 and β_1 by solving the equations

$$\sum_{i}^{n_{\mathrm{D}}}\sum_{j}^{n_{\mathrm{ND}}} w_{ij} \sum_{k=1}^{n_i}\sum_{l=1}^{n_j} \left\{ \begin{array}{c} \begin{pmatrix} 1 \\ u(x) \end{pmatrix} h\big(\beta_0 + \beta_1^{\mathrm{T}} u(x)\big) \\ \big(\partial\theta_{ikjl}/\partial\beta\big)\nu(\theta_{ikjl})^{-1} I\big(X_{ik}^{\mathrm{D}} = X_{jl}^{\mathrm{ND}}\big)\big(U_{ikjl} - \theta_{ikjl}\big) \end{array} \right\} = 0, \quad (3)$$

where $\beta = \big(\beta_0, \beta_1^T\big)^T$, w_{ij} is non-negative weight, $h = \big(g^{-1}\big)'$, $\nu(\theta_{ikjl}) = \theta_{ikjl}(1-\theta_{ikjl})$, and $I(A)$ is the indicator function of the event A.

When the link function is logit (or probit), we could solve the estimating equation simply by using a logistic (or probit) regression model between U_{ijkl} and X_i for pairs $X_i^{\mathrm{D}} = X_j^{\mathrm{ND}}$. Solving (3) is equivalent to solve generalized estimating equations with independent diagonal weights to estimate the model. The most efficient weights $\mathbf{w}^{\mathrm{eff}} = \{w_{ij}, i = 1,2,\ldots,n_{\mathrm{D}}, j = 1,2,\ldots,n_{\mathrm{ND}}\}$ rely on the dependent structure of estimating equations $e_{ikjl} = I\big(X_{ik}^{\mathrm{D}} = X_{jl}^{\mathrm{ND}}\big)\big(U_{ikjl} - \theta_{ikjl}\big)$, for $i = 1,2,\ldots,n_{\mathrm{D}}$, $j = 1,2,\ldots,n_{\mathrm{ND}}$, and $k = 1,2,\ldots,n_i, l = 1,2,\ldots,n_j$. The above estimating equations are constructed from the pairs of observations from the disease and the non-disease groups which are dependent on each other in a complex manner. Thus, $\mathbf{w}^{\mathrm{eff}}$ is almost unfeasible.

We consider two *extreme working independence weights*. One is assigning an equal weight to all lesions (observations) and the other is assigning an equal weight to all subjects (clusters). Suppose we assign an equal weight to all lesions. Then, this is equivalent to assuming

$$w_{ij}^{\mathrm{L}} \propto 1. \quad (4)$$

The weight w_{ij}^{L} is the optimal diagonal weight providing the minimum (determinant of) variance of the regression estimators, when lesions in a cluster are independent or nearly independent. On the other hand, when we assign an equal weight to all subjects, it is equivalent to assigning the weight proportional to $1/n_i^{\mathrm{D}}$ to each lesion of subject $i \in \mathrm{D}$ and the weight proportional to $1/n_j^{\mathrm{ND}}$ to lesions of subject $j \in \mathrm{ND}$. Thus, the weight w_{ij} is

$$w_{ij}^{\mathrm{S}} \propto \big\{1/\big(n_i^{\mathrm{D}} n_j^{\mathrm{ND}}\big)\big\}. \tag{5}$$

It is optimal when lesions within a cluster are dependent. We let $\mathbf{w}^{\mathrm{L}} = \big\{w_{ij}^{\mathrm{L}}, i = 1, 2, \ldots, n_{\mathrm{D}}, j = 1, 2, \ldots, n_{\mathrm{ND}}\big\}$ and $\mathbf{w}^{\mathrm{S}} = \big\{w_{ij}^{\mathrm{S}}, i = 1, 2, \ldots, n_{\mathrm{D}}, j = 1, 2, \ldots, n_{\mathrm{ND}}\big\}$.

In this paper, we consider the class of weights $\mathbf{w} = \big\{w_{ij}, i = 1, 2, \ldots, n_{\mathrm{D}}, j = 1, 2, \ldots, n_{\mathrm{ND}}\big\}$, where

$$\Big\{\mathbf{w} \mid w_{ij} = \lambda w_{ij}^{\mathrm{L}} + \big(1 - \lambda\big) w_{ij}^{\mathrm{S}}, \quad \lambda \in [0, 1]\Big\}. \tag{6}$$

and find λ^* minimizing the variance (or the MSE) of regression estimators.

The variance (or the MSE) of the regression estimators are unknown. We note that we obtain the regression estimates $\widehat{\beta}$, the solution to (3) using the logistic regression. However, the standard errors reported by the logistic regression are no longer accurate due to the dependency among observations U_{ijkl}s. Here, we propose to use a bootstrap procedure to compute the variance (or the MSE) of the regression coefficients. The computed standard errors further provide a method for selecting optimal λ. The procedures are summarized as follows:

- Using resampling techniques, get a data set from the original data set which has the same structure. In other words, the generated data set has the same number of diseased and non-diseased patients for each AUC group as the original data set. Thus, we resample patients separately for each AUC group.

- Let this resampled data set be $\big\{(Y_{ik}^{\mathrm{D}*}, X_{ik}^{\mathrm{D}*}), (Y_{jl}^{\mathrm{ND}*}, X_{jl}^{\mathrm{ND}*})\big\}$. Using this resampled data set, obtain the regression coefficient estimate $\widehat{\beta}_1^*$.

- Repeat the above procedure B times so that we have B regression coefficient estimates $\{\widehat{\beta}_1^{(k)}, k = 1, 2, \ldots, B\}$.
- Compute the variances (or MSEs) of the bootstrapped regression estimates and find λ^* to minimize their sum.
- We here remark that the standard errors of the regression estimators corresponding to λ^* are straightforwardly from the bootstrapped estimates.

3 Numerical study

3.1 Independent Optimal Weights

In this section, we study the relationship between the independent optimal weight and the dependency among observations within a cluster. To be specific, we numerically show that (i) equal weight to all lesions is the optimal independent weight providing the minimum variance (or MSE) of the regression estimators, when lesions in a cluster are independent or nearly independent; (ii) equal weight to all subjects (clusters) is optimal independent weight, when lesions within a cluster are strongly dependent to each other. In the study, we generate data sets with the differential strength of within cluster dependence, and apply different weights in (6) for $\lambda = 0.0, 0.1, \ldots, 0.9, 1.0$ including $w_{ij}^{\mathrm{L}}(\lambda = 0.0)$, $\bigl(1/2\bigr)\bigl(w_{ij}^{\mathrm{L}} + w_{ij}^{\mathrm{S}}\bigr)$, and $w_{ij}^{\mathrm{S}}(\lambda = 1.0)$.

We first explain the model used for the simulation study. The test results Y^{ND} and Y^{D} have values of $1, \ldots, K$ with probabilities $\mathbf{p}_0 = \bigl(\alpha(x), \alpha(x), \ldots, \alpha(x), 1 - (K-1)\alpha(x)\bigr)$ and $\mathbf{p}_1 = \bigl(1 - (K-1)\alpha(x), \alpha(x), \ldots, \alpha(x)\bigr)$, respectively. We set $1\big/K < \alpha(x) < 1\big/(K-1)$. The AUC for covariate x is

$$\mathrm{AUC}(x) = \alpha(x)\bigl\{\bigl(K^2 - 4K + 2\bigr)\alpha(x) + 2\bigr\}\big/2$$

and it is modelled using the logit link as

$$\mathrm{AUC}(x) = \exp(\beta_0 + \beta_1 x)\big/\big\{1 + \exp(\beta_0 + \beta_1 x)\big\}. \tag{7}$$

To generate clustered ordinal results, we use a multivariate normal distribution with mean 0 and covariance Σ whose $(i, j)-$th element is $\sigma_{ij} = \rho^{|i-j|}$. To be more specific, suppose $\big(v_{i1}, v_{i2}, \ldots, v_{in_i}\big)$ is a random vector from a multivariate normal distribution. Then, we define the clustered ordinal results $\big(y_{i1}, y_{i2}, \ldots, y_{in_i}\big)$ of the $i-$th subject as

$$y_{ij} = \begin{cases} 1 & \text{if} \quad v_{ij} \in \big(-\infty, \psi_1\big) \\ 2 & \text{if} \quad v_{ij} \in \big[\psi_1, \psi_2\big) \\ \vdots & \quad\vdots \\ K-1 & \text{if} \quad v_{ij} \in \big[\psi_{k-1}, \psi_K\big) \\ K & \text{if} \quad v_{ij} \in \big[\psi_K, \infty\big), \end{cases} \tag{8}$$

where $\psi_k = \Phi^{-1}\big(k\alpha(x)\big)$, for $k = 1, 2, \ldots, K-1$, and $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. The dependence in ordinal results is induced by that in the multivariate normal distribution.

The data set we consider is composed of two types of clusters, clusters with size c_1 and c_2. In both disease (D) and non-disease group (ND), the number of clusters with size c_1 is (approximately) equal to that with size c_2. We set (c_1, c_2) as $(2, 5)$ or $(3, 4)$. We further assume equal number of clusters for non-disease and disease group with size n, i.e., $n_{\mathrm{D}} = n_{\mathrm{ND}} = n$. We consider a two sample problem where there exists a single covariate $x = 0$ or $x = 1$, where $x = 0$ indicates the control diagnostic method and $x = 1$ indicates the new method. We also assume $\beta_0 = 0$,$\beta_1 = 0.5$ or 0.8, and $\rho = 0.1$, 0.5 or 0.9. The sample size n is chosen as $50, 100, 150$ and 200.

We generate 1000 data sets and compute the estimates of β_1. For each data set, we apply the aforementioned three weights. We compute the bias and the variance for each case (of strength of dependency and different choices of weights). We report the results for $w_{ij}^{\mathrm{L}}(\lambda = 0.0)$, $\big(1/2\big)\big(w_{ij}^{\mathrm{L}} + w_{ij}^{\mathrm{S}}\big)$, and $w_{ij}^{\mathrm{S}}(\lambda = 1.0)$. In all cases, the square of the bias significantly smaller than the variance; thus, the variance is approximately equal to the MSE.

We report the weight (λ_{opt}) which provides the minimum MSE (approximately equals to the minimum variance) among $\lambda = 0.0, 0.1, \ldots, 0.9$, and 1.0. The results are reported in Tables 1-4.

Tables 1-4 also report the coverage probability of confidence interval of β. Suppose $\widehat{\beta}_k$ is the regression estimate of the $k-$th data set. The interval estimator for the $k-$th data set is defined by $\widehat{\beta}_k \pm z_{(1-\alpha)/2}\text{s.e.}\big(\widehat{\beta}_k\big)$, where $\text{s.e.}\big(\widehat{\beta}_k\big)$ is estimated by the standard error of $\widehat{\beta}_1, \ldots, \widehat{\beta}_{1000}$. The empirical coverage probability is the proportion of data sets, whose confidence interval contains the true value of regression coefficient. The empirical coverage probabilities are close to their aimed levels in all cases we consider.

The results confirm that $\mathbf{w}^{\text{L}}$ (or $\mathbf{w}^{\text{S}}$) is optimal when observations within clusters are weakly (or strongly) dependent. This is more pronounced for the case $(c_1, c_2) = (2, 5)$ than the case $(c_1, c_2) = (3, 4)$. Here, the results considering $\mathbf{w}^{\text{L}}$ are consistent with those in Obuchowski (1997).

3.2 Bootstrap for clustered data

Bootstrapping clustered data is an important problem and is much studied in previous literature. The literature includes randomized cluster bootstrap, two-stage bootstrap by Davison and Hinkley (1997), reverse two-stage bootstrap by McCullagh (2000), and residual bootstrap by Andersson and Karlsson (2001). There are also many other bootstrap procedures not listed above and a good review is given in Davison and Hinkley (1997) and Field and Welsh (2007).

The bootstrap procedure in Section 2 is a special case of randomized cluster bootstrap and is denoted by cluster bootstrap. However, it should be remarked that the overall procedure in the paper is flexible to the choice of bootstrap procedure. In other words, in the procedure, we can replace the cluster bootstrap with the others such as the two-stage bootstrap, the reverse two-stage bootstraps, or others.

ρ	n	$\mathbf{w}^{\mathrm{L}}$	$\mathbf{w}^{\lambda=0.5}$	$\mathbf{w}^{\mathrm{S}}$	λ_{opt}	$\mathbf{w}^{\lambda_{\mathrm{opt}}}$
0.1	50	0.0103(0.2655)	0.0128(0.2710)	0.0157(0.2891)	1.0	0.0103(0.2655)
		[0.896, 0.942]	[0.895, 0.945]	[0.901, 0.944]		[0.896, 0.942]
	100	0.0008(0.1894)	-0.0007(0.1922)	-0.0019(0.2046)	0.9	0.0005(0.1892)
		[0.899, 0.950]	[0.901, 0.941]	[0.902, 0.945]		[0.902, 0.950]
	150	0.0036(0.1510)	0.0019(0.1543)	0.0004(0.1649)	1.0	0.0036(0.1510)
		[0.908, 0.968]	[0.907, 0.959]	[0.906, 0.951]		[0.908, 0.968]
	200	0.0051(0.1358)	0.0033(0.1380)	0.0016(0.1465)	0.9	0.0047(0.1357)
		[0.890, 0.955]	[0.898, 0.955]	[0.897, 0.953]		[0.891, 0.956]
0.5	50	0.0119(0.3371)	0.0126(0.3328)	0.0139(0.3441)	0.6	0.0124(0.3324)
		[0.901, 0.950]	[0.904, 0.950]	[0.905, 0.948]		[0.906, 0.949]
	100	0.0139(0.2480)	0.0114(0.2430)	0.0093(0.2493)	0.5	0.0114(0.2430)
		[0.899, 0.960]	[0.902, 0.954]	[0.906, 0.954]		[0.902, 0.954]
	150	0.0079(0.1986)	0.0064(0.1955)	0.0051(0.2015)	0.6	0.0067(0.1954)
		[0.886, 0.946]	[0.886, 0.946]	[0.898, 0.948]		[0.885, 0.945]
	200	0.0091(0.1750)	0.0071(0.1718)	0.0052(0.1764)	0.5	0.0071(0.1718)
		[0.900, 0.949]	[0.894, 0.948]	[0.898, 0.948]		[0.894, 0.948]
0.9	50	0.0022(0.4734)	0.0025(0.4479)	0.0039(0.4432)	0.1	0.0036(0.4424)
		[0.906, 0.953]	[0.911, 0.941]	[0.901, 0.944]		[0.903, 0.944]
	100	0.0249(0.3368)	0.0212(0.3194)	0.0181(0.3156)	0.1	0.0187(0.3152)
		[0.898, 0.953]	[0.903, 0.949]	[0.896, 0.946]		[0.898, 0.946]
	150	0.0155(0.2799)	0.0144(0.2656)	0.0136(0.2618)	0.1	0.0137(0.2617)
		[0.902, 0.936]	[0.897, 0.949]	[0.896, 0.948]		[0.893, 0.945]
	200	0.0169(0.2330)	0.0153(0.2200)	0.0139(0.2166)	0.1	0.0142(0.2164)
		[0.909, 0.945]	[0.904, 0.949]	[0.902, 0.950]		[0.900, 0.952]

Table 1: $\beta_1 = 0.5$ and clusters with size 2 and 5. The values in each cell are bias and $\sqrt{MSE}$. $\sqrt{MSE}$ are in the parentheses.

ρ	n	$\mathbf{w}^{\mathrm{L}}$	$\mathbf{w}^{\lambda=0.5}$	$\mathbf{w}^{\mathrm{S}}$	λ_{opt}	$\mathbf{w}^{\lambda_{\mathrm{opt}}}$
0.1	50	0.0015(0.2667)	0.0006(0.2675)	-0.0002(0.2697)	1.0	0.0015(0.2667)
		[0.911, 0.950]	[0.909, 0.945]	[0.906, 0.943]		[0.911, 0.950]
	100	0.0076(0.1880)	0.0071(0.1883)	0.0067(0.1896)	0.9	0.0075(0.1880)
		[0.905, 0.953]	[0.906, 0.952]	[0.911, 0.957]		[0.906, 0.951]
	150	0.0047(0.1529)	0.0044(0.1527)	0.0042(0.1537)	0.7	0.0045(0.1527)
		[0.896, 0.947]	[0.900, 0.947]	[0.902, 0.946]		[0.896, 0.947]
	200	0.0066(0.1338)	0.0062(0.1338)	0.0058(0.1345)	0.8	0.0064(0.1337)
		[0.892, 0.947]	[0.896, 0.949]	[0.896, 0.945]		[0.895, 0.944]
0.5	50	0.0031(0.3409)	0.0027(0.3408)	0.0024(0.3425)	0.7	0.0029(0.3406)
		[0.897, 0.946]	[0.898, 0.945]	[0.900, 0.949]		[0.897, 0.944]
	100	0.0025(0.2402)	0.0018(0.2396)	0.0012(0.2402)	0.5	0.0018(0.2396)
		[0.898, 0.941]	[0.903, 0.940]	[0.903, 0.937]		[0.903, 0.940]
	150	0.0141(0.1942)	0.0138(0.1935)	0.0135(0.1940)	0.4	0.0137(0.1935)
		[0.893, 0.944]	[0.898, 0.948]	[0.900, 0.947]		[0.898, 0.947]
	200	0.0083(0.1689)	0.0078(0.1684)	0.0073(0.1687)	0.4	0.0077(0.1684)
		[0.887, 0.951]	[0.892, 0.950]	[0.890, 0.948]		[0.893, 0.951]
0.9	50	0.0436(0.4358)	0.0423(0.4329)	0.0413(0.4326)	0.2	0.0417(0.4324)
		[0.897, 0.947]	[0.898, 0.948]	[0.903, 0.951]		[0.897, 0.950]
	100	0.0174(0.3159)	0.0158(0.3138)	0.0143(0.3131)	0.0	0.0143(0.3131)
		[0.899, 0.951]	[0.898, 0.952]	[0.899, 0.955]		[0.899, 0.955]
	150	0.0169(0.2577)	0.0162(0.2557)	0.0156(0.2550)	0.0	0.0156(0.2550)
		[0.896, 0.944]	[0.893, 0.944]	[0.892, 0.946]		[0.892, 0.946]
	200	0.0157(0.2176)	0.0152(0.2157)	0.0148(0.2149)	0.0	0.0148(0.2149)
		[0.898, 0.950]	[0.898, 0.951]	[0.901, 0.951]		[0.901, 0.951]

Table 2: $\beta_1 = 0.5$ and clusters with size 3 and 4. The values in each cell are bias and $\sqrt{MSE}$. $\sqrt{MSE}$ are in the parentheses.

ρ	n	$\mathbf{w}^{\mathrm{L}}$	$\mathbf{w}^{\lambda=0.5}$	$\mathbf{w}^{\mathrm{S}}$	λ_{opt}	$\mathbf{w}^{\lambda_{\mathrm{opt}}}$
0.1	50	0.0013(0.2583) [0.915, 0.954]	-0.0002(0.2625) [0.921, 0.952]	-0.0011(0.2803) [0.905, 0.956]	0.9	0.0009(0.2580) [0.916, 0.953]
	100	0.0048(0.1894) [0.900, 0.958]	0.0025(0.1927) [0.904, 0.955]	0.0006(0.2058) [0.904, 0.955]	0.9	0.0043(0.1892) [0.902, 0.955]
	150	0.0011(0.1557) [0.903, 0.957]	-0.0015(0.1585) [0.901, 0.957]	-0.0038(0.1688) [0.897, 0.956]	0.9	0.0006(0.1556) [0.902, 0.955]
	200	0.0044(0.1374) [0.888, 0.955]	0.0026(0.1396) [0.894, 0.950]	0.0010(0.1483) [0.901, 0.943]	0.9	0.0040(0.1373) [0.885, 0.955]
0.5	50	0.0143(0.3377) [0.904, 0.949]	0.0132(0.3339) [0.892, 0.953]	0.0135(0.3482) [0.894, 0.949]	0.6	0.0133(0.3332) [0.892, 0.953]
	100	0.0174(0.2459) [0.896, 0.954]	0.0151(0.2435) [0.898, 0.952]	0.0134(0.2523) [0.892, 0.950]	0.6	0.0155(0.2430) [0.895, 0.951]
	150	0.0083(0.2024) [0.896, 0.952]	0.0064(0.1989) [0.893, 0.948]	0.0048(0.2044) [0.894, 0.946]	0.6	0.0068(0.1989) [0.893, 0.948]
	200	0.0094(0.1757) [0.907, 0.949]	0.0073(0.1731) [0.903, 0.949]	0.0055(0.1783) [0.904, 0.951]	0.6	0.0077(0.1730) [0.903, 0.947]
0.9	50	0.0416(0.4715) [0.905, 0.946]	0.0397(0.4443) [0.904, 0.947]	0.0395(0.4377) [0.908, 0.953]	0.1	0.0394(0.4373) [0.905, 0.951]
	100	0.0242(0.3291) [0.899, 0.955]	0.0220(0.3128) [0.898, 0.951]	0.0208(0.3110) [0.904, 0.945]	0.2	0.0212(0.3099) [0.902, 0.947]
	150	0.0136(0.2681) [0.893, 0.943]	0.0119(0.2537) [0.891, 0.945]	0.0108(0.2504) [0.890, 0.939]	0.1	0.0110(0.2501) [0.890, 0.939]
	200	0.0176(0.2347) [0.909, 0.951]	0.0157(0.2220) [0.899, 0.956]	0.0142(0.2187) [0.905, 0.949]	0.1	0.0145(0.2186) [0.903, 0.953]

Table 3: $\beta_1 = 0.8$ and clusters with size 2 and 5. The values in each cell are bias and $\sqrt{MSE}$. $\sqrt{MSE}$ are in the parentheses.

ρ	n	$\mathbf{w}^{\mathrm{L}}$	$\mathbf{w}^{\lambda=0.5}$	$\mathbf{w}^{\mathrm{S}}$	λ_{opt}	$\mathbf{w}^{\lambda_{\mathrm{opt}}}$
0.1	50	0.0202(0.2622) [0.904, 0.951]	0.0194(0.2622) [0.905, 0.948]	0.0188(0.2637) [0.901, 0.947]	0.8	0.0199(0.2621) [0.905, 0.950]
	100	0.0074(0.1869) [0.891, 0.958]	0.0065(0.1872) [0.893, 0.959]	0.0057(0.1885) [0.898, 0.959]	0.9	0.0072(0.1869) [0.890, 0.958]
	150	0.0028(0.1512) [0.906, 0.960]	0.0025(0.1517) [0.904, 0.957]	0.0022(0.1530) [0.908, 0.952]	1.0	0.0028(0.1512) [0.906, 0.960]
	200	0.0058(0.1338) [0.893, 0.944]	0.0054(0.1338) [0.892, 0.939]	0.0050(0.1345) [0.895, 0.943]	0.8	0.0056(0.1337) [0.896, 0.942]
0.5	50	0.0020(0.3350) [0.906, 0.946]	0.0030(0.3349) [0.906, 0.944]	0.0041(0.3366) [0.905, 0.946]	0.7	0.0026(0.3348) [0.907, 0.944]
	100	0.0118(0.2461) [0.895, 0.954]	0.0112(0.2452) [0.899, 0.953]	0.0106(0.2454) [0.894, 0.953]	0.3	0.0110(0.2451) [0.894, 0.954]
	150	0.0038(0.1927) [0.898, 0.949]	0.0036(0.1926) [0.898, 0.948]	0.0035(0.1935) [0.898, 0.948]	0.7	0.0037(0.1925) [0.896, 0.947]
	200	0.0085(0.1697) [0.895, 0.954]	0.0079(0.1692) [0.900, 0.953]	0.0073(0.1695) [0.904, 0.948]	0.5	0.0079(0.1692) [0.900, 0.953]
0.9	50	0.0083(0.4303) [0.899, 0.952]	0.0069(0.4271) [0.901, 0.954]	0.0057(0.4261) [0.898, 0.952]	0.1	0.0059(0.4261) [0.898, 0.952]
	100	0.0388(0.3212) [0.892, 0.952]	0.0379(0.3186) [0.896, 0.955]	0.0371(0.3176) [0.899, 0.952]	0.0	0.0371(0.3176) [0.899, 0.952]
	150	0.0147(0.2562) [0.898, 0.951]	0.0145(0.2544) [0.900, 0.950]	0.0143(0.2538) [0.903, 0.945]	0.0	0.0143(0.2538) [0.903, 0.945]
	200	0.0162(0.2192) [0.899, 0.950]	0.0157(0.2173) [0.896, 0.950]	0.0152(0.2166) [0.893, 0.949]	0.0	0.0152(0.2166) [0.893, 0.949]

Table 4: $\beta_1 = 0.8$ and clusters with size 3 and 4. The values in each cell are bias and $\sqrt{MSE}$. $\sqrt{MSE}$ are in the parentheses.

In this section, we numerically investigate the performance of the cluster bootstrap procedure. In the study, we set the number and the size of clusters as those of our real example in Section 4. The example has 36 clusters (patients) and 106 bservations in total as

cluster size	1	2	3	4	5	7	8
number of observations	6	13	6	4	5	1	1

.

We generate samples from the same model used in Section 3.1. We generate 500 data sets for each case of within-cluster correlation $\rho = 0.1, 0.5$, and 0.9. In each data set:

- we resample 400 bootstrapped data sets.
- we compute the regression estimates using different weights $w = 0.0, 0.1, \ldots, 0.9, 1.0$.
- for each w, we construct 95% confidence interval of β using 400 bootstrap estimates.
- we then count the number of data sets among 500 data sets, whose confidence interval contains the true regression parameter β, and compute the empirical coverage probability.

The empirical coverage probabilities are plotted in Figure 1. It shows the cluster bootstrap slightly underestimates the coverage probabilities, particularly when ρ is small. However, the difference between the empirical coverage probabilities and the aimed level of coverage is small and also decreases as ρ increases. The better performance of cluster bootstrap for large ρ would be in accord with the optimality of large λ in Equation (6).

4 Example

The liver is the most frequent site of metastases from various extrahepatic malignancies. Determining the presence of hepatic metastases is important in order to provide the optimal plan for patients who are candidates for surgery and in order to assess prognosis after

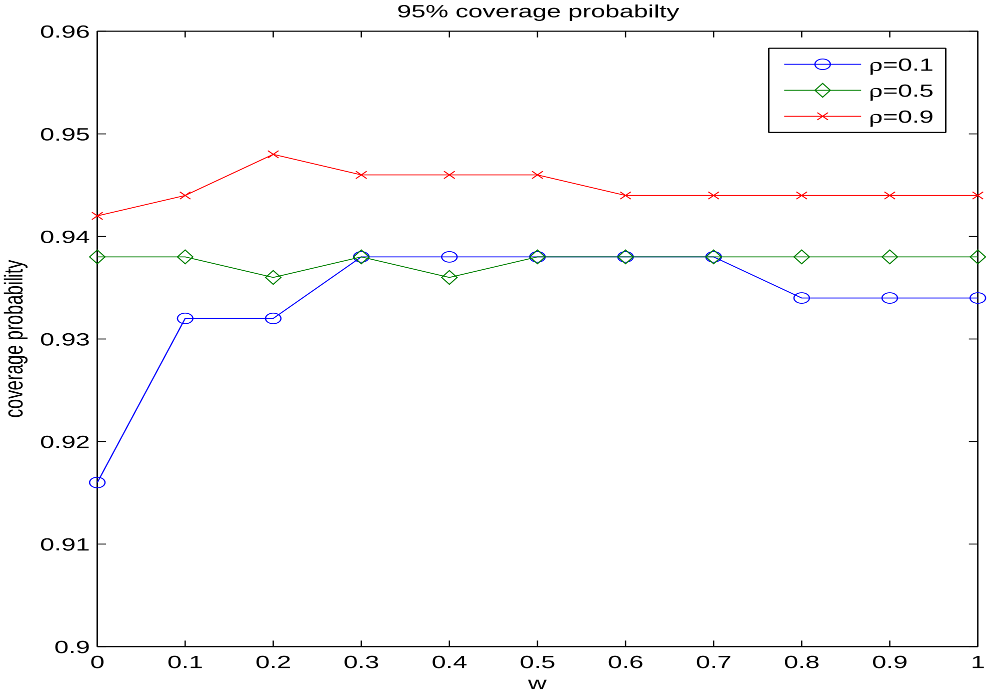

Figure 1: Empirical coverage probabilities for the bootstrap procedure based on sub-sampling clusters.)

initial treatment. Thirty-six (17 men, 46-76 years old; 19 women, 31-74 years old) patients are enrolled as our study population. The extrahepatic malignancies among these patients included 31 colorectal cancers, 3 stomach cancers, 1 duodenal cancer, and 1 renal cell carcinoma. A total of 106 focal lesions (51 metastases and 55 benign lesions) are found among the 36 patients included in our study. The maximum number of focal hepatic lesions seen in a patient is 8 and the minimum number is 1. Eighteen patients with 53 lesions undergo surgery and are confirmed as having either metastases (n = 29) or benign lesions (n = 24). Among the remaining 18 patients with 53 lesions, the histopathologic results are obtained by percutaneous liver biopsies of 14 lesions (13 metastases and 1 eosinophilic abscess). The remaining 39 lesions of these 18 patients are diagnosed as metastases (n = 14) or benign lesions (n = 25) by means of all available data as well as follow-up imaging (ultrasonography, CT, MRI, and other imaging if available) and levels of serum carcinoembryonic antigen. To compare the gadobenate dimeglumine (Gd-BOPTA)-enhanced and the ferucarbotran-enhanced MR imaging methods, an MRI is performed on the following settings:

(Set A) : Gd-BOPTA-enhanced dynamic imaging.
(Set B) : Combination of Gd-BOPTA-enhanced dynamic and delayed imaging.
(Set C) : Ferucarbotran-enhanced delayed imaging.
(Set D) : Combination of ferucarbotran-enhanced dynamic and delayed imaging
(Set E) : Combination of Gd-BOPTA-enhanced dynamic imaging and ferucarbotran-enhanced delayed imaging.

We are interested in comparing **(G1)**(Set A) and (Set B), **(G2)** (Set C) and (Set D) and **(G3)** (Set C) and (Set E) to investigate the effect of additional delayed or dynamic imaging.

All MR images are evaluated on a PACS (picture archiving and communicating system) (Radpia, Hyundai Information Technology) with the annotations masked. Three radiologists (7, 7, and 6 years' experience of abdominal MRI, respectively) perform the image analysis; they are unaware of any other information regarding the patients' clinical history, laboratory results, and the findings of other imaging modalities except that the patients had focal hepatic lesions suspected of being metastases. The MR images are independently reviewed and were interpreted by each observer in five reading sessions separated by at least 4-week intervals in order to avoid learning effects. The readers determine the possibility of malignancy of the detected lesions using a 5-point confidence rating scale (definitely not=0, probably not=1, possibly=2, probably=3, definitely=4).

We first disregard the cluster information by subjects and plot the empirical ROC curves in Figure 2. The figure shows that the additional delayed or dynamic imaging improves the accuracy in diagnosing abnormality. To be specific, we find that:

- The ROC curve of (Set B) has higher value than that of (Set A);equivalently, (Set B) has higher TPR than (Set A) at all FPR.

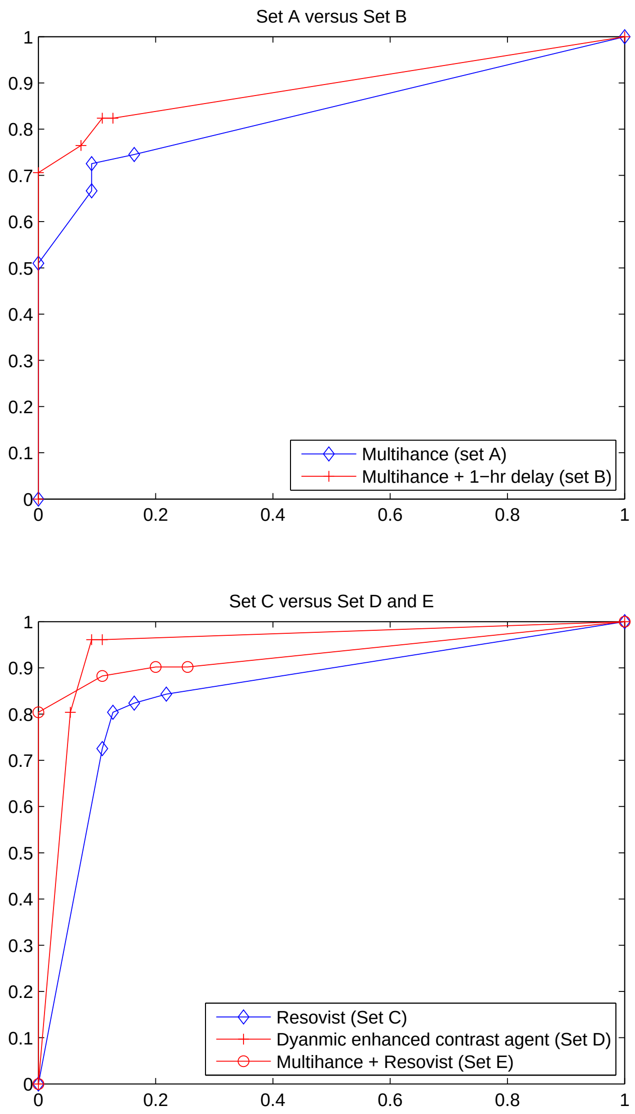

Figure 2: The empirical ROC curves of (Set A)-(Set E)

- The ROC curve of (Set D) has higher value than that of (Set C);equivalently, (Set D) has higher TPR than (Set C) at all FPR.

- The ROC curve of (Set E) has higher value than that of (Set C);equivalently, (Set E) has higher TPR than (Set C) at all FPR.

Now, we test whether these improvements are statistically significant. To do it, we apply the proposed AUC regression model with two discrete covariates, picturing methods and readers. We consider the class of working independence weights in (6). We consider λ to minimize the sum of variances of all estimates without the intercept or the estimates of (Set B)-(Set E. The grid search shows that, in both cases, the λ to minimize the sum of variances is 0.9 (see Figure 3.). Here, the variances at a given λ are estimated using 1000 bootstrap samples.

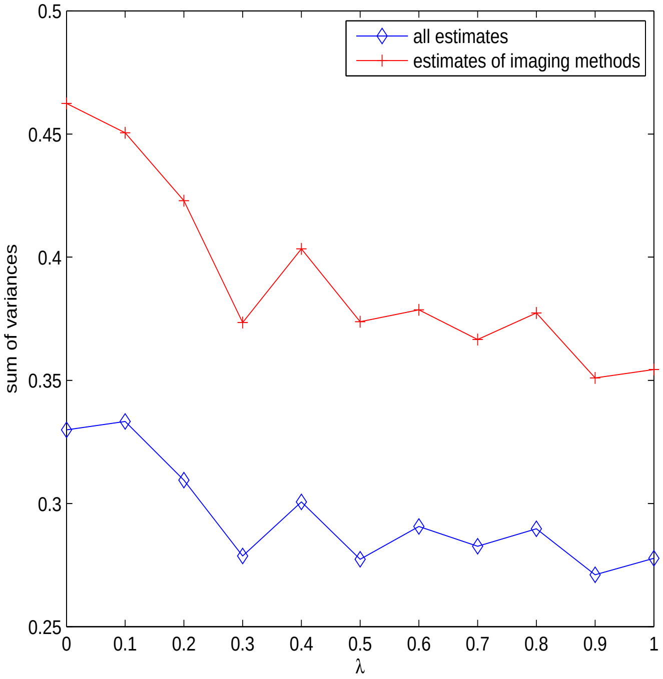

Figure 3: λ versus the sum of variances of estimates.

We now estimate the model with working independence weight $w_{ij} = 0.9w_{ij}^{\text{L}} + 0.1w_{ij}^{\text{S}}$ and the results are reported in Table 5. For Gd-BOPTA-enhanced MRI, there is no significant difference in AUC with adjustment for clustering effect between the dynamic set and the combination set (set A versus set B) in detecting hepatic metastases (p-value = 0.1778). For ferucarbotran-enhanced MRI, however, there is a significant difference between the delay set and the combination set (set C versus set D) (p-value = 0.0462). In comparison of AUCs between double- and single-contrast sets, set E does not have significant improvement from set C for detecting hepatic metastases (p-value = 0.1252).

In summary, there is no significant difference in detecting hepatic metastases between the two contrast-enhanced MR imaging method, but the combination of Gd-BOPTA-enhanced dynamic and delayed imaging methods has an additive value in detecting hepatic metastases.

r	Est.	SE	p-value	Est.	SE	p-value
Intercept	1.8999	0.2881	<0.0001	1.8844	0.2927	<0.0001
Reader2	0.2111	0.2181	0.3331	0.2111	0.2181	0.3331
Reader3	-0.0267	0.1815	0.8830	-0.0267	0.1815	0.8830
Set A	−−	−−	−−	0.0155	0.2892	0.9573
Set B	0.1865	0.1384	0.1778	0.2019	0.3120	0.5176
Set C	-0.0155	0.2892	0.9572	−−	−−	−−
Set D	0.4437	0.3430	0.1958	0.4592	0.2303	0.0462
Set E	0.4471	0.2262	0.4081	0.4625	0.3016	0.1252

Table 5: Results of the analysis of liver abnormality tests. The left set of estimates assumes the estimate of Set A is 0. The right set assumes the estimate of Set C is 0.

5 Conclusion

In this paper, extending the results of Dodd and Pepe (2003), a weighted regression model is proposed, along with a procedure to estimate it when the test has clustered ordinal results. Here, the weights reflect the cluster information. We consider a class of working independence weights which are convex combinations of two extreme working independence weights. One is assigning an equal weight to all lesions (observations) and the other is assigning an equal weight to all subjects (clusters). We propose to use a cluster bootstrap procedure to estimate the mean squared error of regression estimators. We then propose to use the weight in the class to minimize the variance (or the mean squared error) of regression estimates. We numerically show that (i) equal weight to all lesions is the optimal independent weight providing the minimum variance (or MSE) of the regression estimators, when lesions in a cluster are independent or nearly independent; (ii) equal weight to all subjects (clusters) is optimal, when lesions in a cluster are strongly dependent to each other. We also numerically investigate the performance of the cluster bootstrap for the data sets having same sizes with the motivating example. We apply the proposed procedure to testing several picturing methods to detect liver abnormality and find that the combination of Gd-BOPTA-enhanced dynamic and delayed imaging methods has an additive value in detecting hepatic metastases.

Acknowledgements

We thank the associate editor and the reviewer for their helpful comments and suggestions. Johan Lim's research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by Ministry of Education, Science and Technology(No. 23 2010-0011448).

References

Alonzo, T. and Pepe, M. (2002). Distribution-free ROC analysis using binary regression techniques. *Biostatistics*, 3:421–432.

Andersson, M. and Karlsson, S. (2001). Bootstrapping error component models. *Computational Statistics*, 16:221–231.

Cai, T. and Pepe, M. (2002). Semi-parametric ROC analysis to evaluate biomakers for disease. *Journal of the American Statistical Association*, 97:1099–1107.

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and Their Application.* Cambridge: Cambridge University Press.

Dodd, L. and Pepe, M. (2003). Semiparametric regression for the area under the receiver operating charateristic curve. *Journal of the American Statistical Association*, 98:409–417.

Faraggi, D. (2003). Adjusting receiver operating characteristic curves and related indices for covariates. *The Statistician*, 52:179–192.

Field, C. and Welsh (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society- Series B*, 69(3):289–301.

Ju, H. (2004). Topics in analyzing longitudinal data. *Unpublished Ph.D. Thesis, Texas A&M University.*

Le, C. (1997). Evaluation of confounding effects in ROC studies. *Biometrics*, 53:998–1007.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–21.

McCullagh, P. (2000). Re-sampling and exchangable arrays. *Bernoulli*, 6:285–301.

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

Pepe, M. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*, 54:124–135.

Pepe, M. (2000). An interpretation for the ROC curve and inference using GLM procedures. *Biometrics*, 56:352–359.

Schisterman, E., Faraggi, D., and Reiser, B. (2004). Adjusting the generalized ROC curve for covariates. *Statistics in Medicine*, 23:3319–3331.

Tang, L. and Balakrishnan, N. (2011). A random-sum Wilcoxon statistic and its application to analysis of ROC and LROC data. *Journal of Statistical Planning and Inference*, 141:335–344.

Tosteson, A. and Begg, C. (1988). A general regression methodology for ROC curve estimation. *Medican Decision Making*, 8:204–215.

Zeltzer, S., Swensson, R., Nawfel, R., Lentini, J., Kazda, L., and Judy, P. (1991). Visualization and detection-localization on computed tomographic images. *Investigative Radiology*, 26(4):285–294.