View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by D-Scholarship@Pitt

STATISTICAL METHODS FOR EVALUATING BIOMARKERS SUBJECT TO DETECTION LIMIT

by

Yeonhee Kim

B.S. and B.B.A., Ewha Womans University, Korea, 2001

M.S., North Carolina State University, 2005

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

DEPARTMENT OF BIOSTATISTICS

This dissertation was presented

by

Yeonhee Kim

It was defended on

May 26, 2011

and approved by

Lan Kong, Ph.D., Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Andriy Bandos, Ph.D., Research Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Jong-Hyeon Jeong, Ph.D., Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

(Joyce) Chung-Chou Ho Chang, Ph.D., Associate Professor, School of Medicine, University of Pittsburgh

Dissertation Director: Lan Kong, Ph.D., Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

STATISTICAL METHODS FOR EVALUATING BIOMARKERS SUBJECT TO DETECTION LIMIT

Yeonhee Kim, PhD

University of Pittsburgh, 2011

As a cost effective diagnostic tool, numerous candidate biomarkers have been emerged for different diseases. The increasing effort of discovering informative biomarkers highlights the need for valid statistical modeling and evaluation. Our focus is on the biomarker data which are both measured repeatedly over time and censored by the sensitivity of given assay. Inappropriate handling of these types of data can cause biased results, resulting in erroneous medical decision.

In the first topic, we extend the discriminant analysis to censored longitudinal biomarker data based on linear mixed models and modified likelihood function. The performance of biomarker is evaluated by area under the receiver operation characteristic (ROC) curve (AUC). The simulation study shows that the proposed method improves both parameter and AUC estimation over substitution methods when normality assumption is satisfied for biomarker data. Our method is applied to the biomarker study for acute kidney injury patients. In the second topic, we introduce a simple and practical evaluation method for censored longitudinal biomarker data. A modification of the linear combination approach by Su and Liu [1] enables us to calculate the optimum AUC as well as relative importance of measurements from each time point. The simulation study demonstrates that the proposed method performs well in a practical situation. The application to real-world data is provided. In the third topic, we consider censored time-invariant biomarker data to discriminate time to event or cumulative events by a particular time point. C-index and time dependent ROC curve are often used to measure the discriminant potential of survival model. We extend

these methods to censored biomarker data based on joint likelihood approach. Simulation study shows that the proposed methods result in accurate discrimination measures. The application to a biomarker study is provided.

Both early detection and accurate prediction of disease are important to manage serious public health problems. Because many of diagnostic tests are based on biomarkers, discovery of informative biomarker is one of the active research areas in public health. Our methodology is important for public health researchers to identify promising biomarkers when the measurements are censored by detection limits.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to thank my family - mom, dad, JongHoon, MinYoung, Charlotte and Banibani for their love.

My advisor, Professor Lan Kong, deserves much gratitude for her constant guidance and mentoring. Also I want to thank my committee members, Dr.Bandos, Dr.Chang and Dr.Jeong, for their time and insight.

I am also thankful to the Clinical and Translational Science Institute for supporting my study.

1.0 INTRODUCTION

Biomarkers are measurable factors that can be used as an indicator of disease or a progression of disease. For example, cholesterol level works as a risk predictor of vascular disease, and serum creatinine is a surrogate for renal disease progression. Due to a biomarkers' cost-effective benefit for the diagnosis and prognosis of acute and chronic diseases, discovery of a new biomarker is one of the active areas in medical research. Researchers have developed several evaluation tools for biomarker discovery. Diagnostic measures quantify biomarker's ability of discrimination. It focuses on whether the biomarker can separate patients into event/non-event group. On the other hand, prognostic measures indicate biomarker's predictive capacity of disease occurrence. The risk can be expressed as a function of biomarker through a statistical model such as logistic regression or Cox proportional hazard model.

Biomarker data are collected from many different procedures, designs and sampling schemes. For instance, biomarkers can be collected only at one time point, or collected repeatedly over several time points. Regardless of the data structure, it is tempting to use only the most recent data in the analysis because of a complexity in handling longitudinal data. Besides the high dimensionality of longitudinal data, analysis of biomarker data becomes more complicated if some measurements are censored. The censoring occurs due to a limit of detection (LOD). In this case, only measurements which lie between lower and upper detection limits are observable.

We encountered longitudinal censored data from two biomarker studies: The Genetic and Inflammatory Markers of Sepsis (GenIMS) study and the Biological Markers of Recovery for the Kidney (BioMaRK) study. The GenIMS study is a multicenter, cohort study of 2320 patients with community acquired pneumonia (CAP) followed over time. The CAP is the most common cause of sepsis that can lead to death. A set of biomarkers were measured

daily for a week or longer during the hospitalization. One of the goals of this study was to find the relationship between pathways of biomarkers and the risk of sepsis and death. Because of the sensitivity of the assays used to measure the biomarkers, concentration of some biomarkers was below the detectable limit, resulting in a portion of unquantifiable data. The BioMaRK study was conducted as a part of a large randomized clinical trial [2], and enrolled patients who have a renal-replacement therapy for acute kidney injury. The acute kidney injury is a clinically challenging problem for both physicians and patients. Although it is directly related to the health care cost and well-being of patients, effective treatment of acute kidney injury is still not available. Hence, many clinical studies were initiated to explore informative biomarkers for the outcome of renal function. In BioMaRK study, multiple plasma and urinary biomarkers are measured repeatedly, and the measurements of some biomarkers are censored due to detection limits. In the previous analysis of longitudinal data, it was common to analyze the biomarker at each time point separately. Censored data were usually deleted or substituted by LOD or LOD/2 with the justification that it is easy to implement and widely understood [3]. However, investigators are frequently interested in longitudinal performance of biomarkers. Furthermore, disregarding the censored data often causes significant biases in the estimates of the fixed effects and variance components, inaccurate estimates of summary statistics, and inaccuracies in risk assessments [4] [5]. The objectives of our research are (1) to develop a classification method for the longitudinal biomarkers subject to left or right censoring due to lower or upper detection limit, and (2) to evaluate the censored biomarker performance for both binary and survival outcomes. The organization of the dissertation is as follows. In chapter 2, we review the models for longitudinal data and existing methods for handling censored data. Underlying theory on the classification method is introduced, followed by statistical evaluation tools for binary and survival outcomes. Chapter 3 contains the classification methods for longitudinal censored data. In chapter 4, we present how to incorporate the longitudinal biomarkers in the ROC analysis for both censored and non-censored cases. In chapter 5, we change the outcome of interest from binary data to survival data. With the baseline censored biomarker measurements, we calculate the discrimination accuracy for survival outcome by modifying

the original estimation methods for time dependent ROC and C-index. In chapter 6, we close the dissertation with summary on the proposed methods and discussion about future extensions.

2.0 LITERATURE REVIEW

2.1 CLASSIFICATION METHODS

Linear discriminant analysis (LDA) and logistic regression are two standard statistical methods for classification. They are similar in terms of comparing the posterior probabilities that a subject is from group k ($Group_k$) when deciding a group membership. Suppose biomarker Y from $Group_k$ is an $n \times 1$ vector of observations with mean μ^k and covariance matrix Σ_k. The π_k is the prior probability of a subject belonging to $Group_k$ ($k = 0, 1$). LDA assumes that biomarker data follow a normal distribution with common covariance matrix, $\Sigma_0 = \Sigma_1 = \Sigma$. The probability density function of Y from $Group_k$ is

$$f_k(Y) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp\left[-\frac{(Y-\mu^k)^t\Sigma^{-1}(Y-\mu^k)}{2}\right].$$

Using the Bayes's rule, the posterior probability of $Group_k$ is calculated as

$$Pr(Group_k|Y) = \frac{f_k(Y)\pi_k}{f_0(Y)\pi_0 + f_1(Y)\pi_1}.$$

Two posterior probabilities are compared in a log scale so that the log ratio of posterior probabilities leads to an equation linear in Y [6]. We call it as a discriminant function of LDA. Because when $\Sigma_0 = \Sigma_1 = \Sigma$,

$$log\frac{Pr(Group_1|Y)}{Pr(Group_0|Y)} = log\frac{f_1(Y)}{f_0(Y)} + log\frac{\pi_1}{\pi_0} = log\frac{\pi_1}{\pi_0} - \frac{1}{2}(\mu^1+\mu^0)^t\Sigma^{-1}(\mu^1-\mu^0) + Y^t\Sigma^{-1}(\mu^1-\mu^0). \tag{2.1}$$

The assumption of LDA is generalized in quadratic discriminant analysis (QDA) by allowing different covariance matrices between groups. If Y from $Group_k$ is distributed according to N(μ^k, Σ_k), the quadratic discriminant function is

$$\frac{Y^t\left({\Sigma_0}^{-1}-{\Sigma_1}^{-1}\right)Y}{2}+Y^t\left({\Sigma_1}^{-1}\mu^1-{\Sigma_0}^{-1}\mu^0\right)+log\frac{\pi_1}{\pi_0}-\frac{log|\Sigma_1|/|\Sigma_0|}{2}-\frac{1}{2}\left(\mu^{1t}{\Sigma_1}^{-1}\mu^1-\mu^{0t}{\Sigma_0}^{-1}\mu^0\right).$$

While LDA has a linear discriminant boundary, the discriminant function of QDA has a quadratic term of Y, leading to a quadratic boundary. Non-linear boundary for classification works better especially in case of non-normal data and heterogeneous covariance matrix for two groups.

More generally, likelihood ratio method has long been recognized as an optimal classification rule and it does not require assumptions such as normality or homogeneous covariance matrix. Using the Bayes rule, it can be shown that the likelihood ratio rule is equivalent to rules based on the posterior probability $Pr(Group_k|Y)$. In this sense, the discriminant analysis provides classification which achieves optimality [7].

The discriminant function is compared with a cutoff point to determine a group membership. A cutoff point c is set by the decision theory. The most common goal in the decision theory is to minimize the expected loss. Let $L(Group_k, Group_j)$ be a loss function that indicates the loss by misclassifying a subject in $Group_k$ as in $Group_j$ ($j = 1 \cdots d$). The minimum expected loss can be written in a functional as

$$min_c(\text{Expected loss}) = min_c\left[\sum_{j=1}^{d} L(Group_k, Group_j)Pr(Group_j|Y)\right].$$

The loss function is chosen depending on the cost of a false positive and false negative.

In a logistic regression, log odds of a posterior probability is assumed to be linear in Y :

$$log\frac{Pr(Group_1|Y)}{Pr(Group_0|Y)} = \beta_0 + \beta_1 Y. \tag{2.2}$$

It follows from the equations (2.1) and (2.2) that

$$log\frac{\pi_1}{\pi_0}-\frac{1}{2}(\mu^1+\mu^0)^t\Sigma^{-1}(\mu^1-\mu^0)+Y^t\Sigma^{-1}(\mu^1-\mu^0)=\beta_0+\beta_1 Y.$$

The only difference between LDA and logistic regression is a distributional assumption. LDA assumes that the biomarker in each group follows a normal distribution with common covariance matrix. In contrast, logistic regression does not impose any restrictions on the distribution. It is known that logistic regression is more flexible and performs better when the normal assumption is violated. However, LDA is shown to perform better and yield more efficient estimates of parameters with smaller variance when the assumption is satisfied. In addition, results from LDA are more stable when subjects are classified into more than two groups [6].

Fisher's linear discriminant analysis

Fisher's discriminant analysis is closely linked to LDA. Fisher's discriminant analysis finds a coefficient λ that can best discriminate the data in different classes. The principle of the best discrimination is to maximize the ratio of between class variance to within class variance. The objective function to maximize is expressed as

$$J(\lambda) = \frac{\lambda^t S_B \lambda}{\lambda^t S_W \lambda},$$

where

$$S_B = \sum_{k=1}^{d} \pi_k (\mu^k - \mu)(\mu^k - \mu)^t, \quad S_W = \sum_{k=1}^{d} \pi_k \left[\frac{1}{n_k} \sum_{y \in Group_k} (Y - \mu^k)(Y - \mu^k)^t \right],$$

μ is the grand mean, n_k is the number of subjects in $Group_k$ and λ indicates a linear subspace within which the projection of observations from different classes are best separated. When there are two classes, the solution for λ is $S_W^{-1}(\mu_1 - \mu_0)$ [8]. We can recognize that the linear coefficient of Y in the discriminant function (2.1) is exactly the same as Fisher's linear discriminant coefficient, given the fact that

$$Y^t \Sigma^{-1}(\mu^1 - \mu^0) = (\mu^1 - \mu^0)^t \Sigma^{-1} Y = \{\Sigma^{-1}(\mu^1 - \mu^0)\}^t Y = \{S_W^{-1}(\mu^1 - \mu^0)\}^t Y.$$

LDA projects biomarker measurements into the linear subspace generated by $\Sigma^{-1}(\mu^1 - \mu^0)$, which is the Fisher's linear discriminant coefficient, and clusters them into different groups that are separated by a linear boundary based on the minimum expected loss.

2.2 STATISTICAL METHODS FOR CENSORED DATA

When an instrument is not sensitive enough to measure very high or low values, only observable values are reported for the analysis. Several parametric and non-parametric methods such as deletion, substitution, imputation, and maximum likelihood method have been proposed to resolve the problems.

Deletion means the elimination of all censored data. It reduces the sample size and could produce a large bias. The missing pattern due to elimination is 'nonignorable missing' because the absence of data depends on detection limits. Alternative method is a substitution of censored values by LOD, LOD/2 or $LOD/\sqrt{2}$ [9]. The substitution method is widely used in practice due to its simplicity. However, the substitution still leads to a biased estimation if the distribution of a biomarker beyond LOD is still informative. If the distributional assumption is possible for measurement data, conditional expected value $E(Y|Y < LOD)$ can be assigned to censored data, which is calculated based on the parameters of the distribution and detection limit value [10]. Another, but similar method is single imputation method [11]. From the estimated distribution, it replaces censored data with randomly sampled values. The single imputation method can make estimates minimally biased, but still produces too narrow confidence interval particularly when more than 30% observations are censored. The major problem of single imputation is that the method ignores complexity of the model as well as variability of the imputation process [12]. For left-censored data, one might think that they are not important because the actual values must be extremely small. However, censored data still have a large effect on the estimates of mean and variance, descriptive statistics, regression coefficient, its standard errors, and power of hypothesis tests, especially when the proportion of censoring is not small [13].

To protect against above problems, multiple imputation (MI) method is suggested for censored data [12] [14]. In MI method, maximum likelihood estimates are first obtained for parametric distribution using all available data. With the estimated parameters, censored data are imputed by a sampling procedure. Because the imputed values are not real data, the imputation process is usually repeated several times to create multiple complete data sets. The analysis result from each dataset is combined later to account for variability. The MI

method provides accurate estimates and robust results even though the censoring proportion is high [15]. Another promising statistical approach from methodological perspective is a maximum likelihood estimation (MLE) method. It uses a modified likelihood function that can incorporate the mechanism of censoring in parametric models [16]. The tobit model, which uses a truncated normal distribution for censored data, is one of the widely applied parametric models [17]. MLE method provides less biased estimates and increased standard errors compared to the substitution method when data follow approximately normal distribution [18] [15]. Although some drawbacks exist, for example, MLE works poor when a sample size is small and outliers exist, this method is still preferred to others because MLE itself has several desirable properties such as consistency, asymptotic unbiasedness, and efficiency. It is often considered the gold standard provided that the data are well described by a (log)normal distribution [4] [19] [3] [20].

2.3 EVALUATION OF BIOMARKER

Biomarker's usefulness is often evaluated from either discrimination or risk prediction point of view. Discrimination describes how well a model separates subjects into event and non-event group. Risk prediction concerns a predictive capacity of a biomarker. The predictive capacity is quantified by a risk distribution in the population [21]. According to this definition, a biomarker is said to be useful if predicted risks have a wide distribution in the population so that clinicians can easily divide patients into low and high risk group with fewer subjects being left in the intermediate equivocal risk range. Discrimination and risk prediction are originated from different perspectives. If the objective is a correct classification, discrimination approach is appropriate. If the clinical utility of a biomarker is of interest, risk prediction approach is preferred. There is no gold standard for the evaluation method. It is recommended to choose a proper method depending on the objective of the study. [22]

Table 1: Concept of sensitivity and specificity

	Test negative (T=0)	Test positive (T=1)
Non-disease (D=0)	$Pr(T = 0\|D = 0)$ = TNF	$Pr(T = 1\|D = 0)$ = FPF
Disease (D=1)	$Pr(T = 0\|D = 1)$ = FNF	$Pr(T = 1\|D = 1)$ = TPF

2.3.1 ROC curve

Discrimination performance is usually expressed through sensitivity and specificity. When a test result is dichotomized (i.e. disease/non-disease, positive/negative), sensitivity and specificity directly show a frequency of correct classification. Assuming that a positive test result indicates the presence of a disease, sensitivity is defined as a probability of a positive test result given a patient has a disease. Specificity is defined as a probability of a negative test result given a patient doesn't have a disease. Alternatively, sensitivity can be expressed as true positive fraction (TPF) or 1 - false negative fraction (FNF). Another expression of specificity is true negative fraction (TNF) or 1 - false positive fraction (FPF) (Table 1).

Sometimes researchers want to know the averaged sensitivity over all specificity region to compare overall performances. Especially when the outcome has ordinal or continuous scale, the ROC curve is a useful tool for summarization. The ROC curve is a graph of sensitivity on y-axis as a function of (1-specificity) on x-axis under series of cutoff points. In Figure 1, the larger the AUC, the better the biomarker discriminates between diseased and non-diseased subjects. The perfect accuracy corresponds to AUC of 1, and the practical lower limit for the AUC is 0.5, which can be achieved by a random chance.

Nonparametric ROC curve

Empirical ROC curve is estimated without any assumptions on the distribution of biomarker data. Let $Y_{0i}(i = 1, \cdots, n_0)$ and $Y_{1j}(j = 1, \cdots, n_1)$ be continuous test results from patients without and with a disease, respectively. Nonparametric ROC curve is a non-smooth step

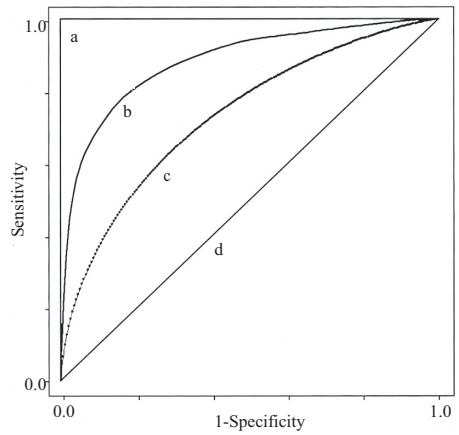

Figure 1: Biomarker a shows the perfect accuracy whereas d shows the worst accuracy

function that changes values at most $n_0 + n_1 + 1$ points. Two coordinates of each point is defined by

$$1 - Specificity = \frac{1}{n_0} \sum_{i=1}^{n_0} I(Y_{0i} > c)$$
$$Sensitivity = \frac{1}{n_1} \sum_{j=1}^{n_1} I(Y_{1j} > c).$$

The AUC is a summation of the areas under the trapezoids and it is also equivalent to Mann-Whitney U-statistics. The nonparametric estimator of AUC is expressed by

$$\widehat{AUC} = \frac{1}{n_0 n_1} \sum_{j=1}^{n_1} \sum_{i=1}^{n_0} \psi(Y_{0i}, Y_{1j}),$$

where

$$\psi(Y_{0i}, Y_{1i}) = \begin{cases} 1 & \text{if } Y_{1i} > Y_{0i} \\ \frac{1}{2} & \text{if } Y_{1i} = Y_{0i} \\ 0 & \text{if } Y_{1i} < Y_{0i} \end{cases}$$

The trapezoidal method is easy to implement, but underestimates the area when the number of distinct test values is small. There are different methods to derive the variance for AUC, such as methods by Bamber [23], Hanley and McNeil [24] and DeLong et al. [25]. Define Y_0-components V_{10} for i^{th} subject and Y_1-components V_{01} for j^{th} subject as

$$V_{10}(Y_{0i}) = \frac{1}{n_1}\sum_{j=1}^{n_1}\psi(Y_{0i}, Y_{1j}), \quad (i = 1, \cdots, n_0)$$
$$V_{01}(Y_{1j}) = \frac{1}{n_0}\sum_{i=1}^{n_0}\psi(Y_{0i}, Y_{1j}), \quad (j = 1, \cdots, n_1).$$

DeLong et al. [25] proposed the variance estimator for nonparametric $\widehat{AUC}$ as

$$\widehat{Var}(\widehat{AUC}) = \frac{1}{n_0}S_{10} + \frac{1}{n_1}S_{01},$$

where

$$S_{10} = \frac{1}{n_0 - 1}\sum_{i=1}^{n_0}(V_{10}(Y_{0i}) - \widehat{AUC})^2$$
$$S_{01} = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(V_{01}(Y_{1j}) - \widehat{AUC})^2.$$

Parametric ROC curve

The binormal ROC model is often employed as a parametric method to obtain a smooth ROC curve. The binormal ROC model postulates a pair of overlapping normal distributions to represent the distribution of two populations [26]. Suppose continuous test results from non-diseased population $Y_0 \sim N(\mu^0, \sigma_0^2)$, and from diseased population $Y_1 \sim N(\mu^1, \sigma_1^2)$. Under each cutoff point c,

$$Sensitivity = Pr(Y_1 > c) = \Phi\left(\frac{\mu^1 - c}{\sigma_1}\right)$$
$$1 - Specificity = Pr(Y_0 > c) = \Phi\left(\frac{\mu^0 - c}{\sigma_0}\right), \tag{2.3}$$

where Φ is a standard normal cumulative distribution function. It follows from the equation (2.3) that

$$Sensitivity = \Phi\left[\frac{\mu^1 - \mu^0}{\sigma_1} + \frac{\sigma_0}{\sigma_1} \times \Phi^{-1}(1 - Specificity)\right].$$

The ROC curve is entirely determined by two parameters u and v, where $u = (\mu^1 - \mu^0)/\sigma_1$ is the standardized difference in the means of diseased and non-diseased population, and $v = \sigma_0/\sigma_1$ is the ratio of the standard deviations of two populations. The AUC is calculated as

$$AUC = Pr(Y_0 < Y_1) = \Phi\left(\frac{\mu^1 - \mu^0}{\sqrt{\sigma_0^2 + \sigma_1^2}}\right) = \Phi\left(\frac{u}{\sqrt{1 + v^2}}\right).$$

By Taylor's expansion, the variance formula for the parametric estimator of AUC is

$$Var(\widehat{AUC}) = \left(\frac{\partial AUC}{\partial u}\right)^2 Var(\hat{u}) + \left(\frac{\partial AUC}{\partial v}\right)^2 Var(\hat{v}) + \left(\frac{\partial AUC}{\partial u}\right)\left(\frac{\partial AUC}{\partial v}\right) Cov(\hat{u}, \hat{v}).$$

Under the asymptotic normality, 100(1 - α)% confidence interval for AUC is given by $\widehat{AUC} \pm Z_{\alpha/2}\sqrt{\widehat{Var}(\widehat{AUC})}$.

ROC curve for censored data

The parametric ROC curve has been extended to incorporate the censored measurements due to detection limit. Perkins et al. [27] [28] developed the method to estimate AUC by obtaining consistent estimates for μ^1, μ^0, σ_0^2 and σ_1^2. Their AUC, $\Phi\left(\frac{\mu^1 - \mu^0}{\sqrt{\sigma_0^2 + \sigma_1^2}}\right)$, yields the similar value to the AUC from completely observed data. Vexler et al. [29] developed the maximum likelihood ratio test to compare AUCs from two biomarkers subject to LOD. Because two biomarker measurements (for example, cholesterol and HDL-cholesterol) from one subject can be correlated, they took both censoring and correlation into account. They employed bivariate normal distribution and used a cumulative distribution function for censored data conditioning on non-censored data.

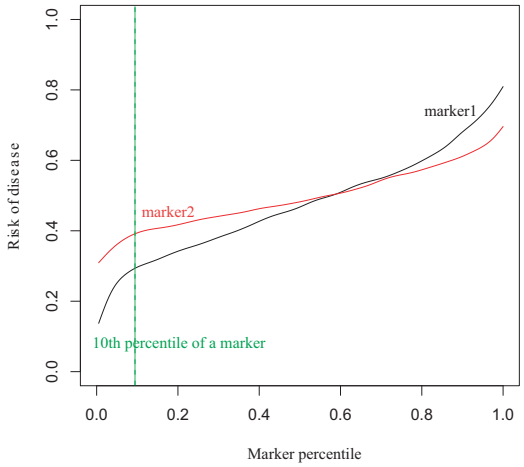

Figure 2: Biomarker 1 is more predictive than biomarker 2.

2.3.2 Predictiveness curve

Although ROC curve has been the most popular method for a biomarker evaluation, it does not take a risk distribution into account. Suppose all diseased subjects have same risk values of 0.52 and all non-diseased subjects have values of 0.51. In the ROC analysis, this would result in a perfect discrimination. Such a weakness triggers researchers to evaluate biomarkers from a different perspective. Huang et al. [21] published a predictiveness curve as a graphical way to present a predicted risk. It is a plot of predicted risk $R(v)$ against the v^{th} percentile of the biomarker, where $R(v) = Pr[D = 1|Y = F^{-1}(v)]$ and F is the cumulative distribution function of biomarker Y. Even though the original scales of the biomarkers are not comparable because the measurement can be different, they are transformed to a common scale in the predictiveness curve by using a percentile of the biomarker. It has been pointed out in many papers that the biomarker with a strong predictive capacity has steeper curves that corresponds to a wide variation in risk [21] [30] [31]. In Figure 2, the biomarker 1 is more predictive than the biomarker 2 because more subjects are in high or low end of

the risk. For example, the subjects in the bottom 10% of the marker distribution have risks in the range of (0.15, 0.30) according to biomarker 1, but in a much higher range (0.30, 0.40) according to biomarker 2. If the predictiveness curve is close to the horizontal line, the biomarker is no more helpful for making a medical decision.

2.3.3 C-index

The C-index was proposed by Harrell et al. [32] as an overall measure of discrimination accuracy for survival outcome. The concept of C-index was motivated by Kendall's τ, a nonparametric version of correlation. In their original paper, the C-index was applied not only to survival data but also to binary data. They pointed out that the C-index for binary outcome is equivalent to the AUC. The definition of AUC for binary outcome is the probability that the diseased subject has worse biomarker value than the non-diseased subject. Changing the outcome from a binary to survival time, the C-index is defined as the probability that the patient with better biomarker value will have a longer survival time than the patient with worse biomarker record, assuming that these two patients are selected at random.

Suppose that (Z_i, U_i, W_i, Y_i) are the actual survival time, predicted survival time, predicted probability of survival at time t, and time-invariant biomarker measurements for i^{th} subject ($i = 1,\cdots, N$), respectively. Harrell et al. [33] expressed C-index as $Pr(U_i < U_j | Z_i < Z_j)$. In practice, it is hard to predict individual's survival time. It is noted that the predicted probability of survival until any fixed time point (W_i) can take place of the predicted survival time(U_i), if two estimates have one-to-one correspondence. One advantage in the application is that this relationship holds when the proportional hazard assumption is satisfied. Under the proportional hazard model, $S(t|Y_i) = (S_0(t))^{\beta^t Y_i}$, where $S(t|Y_i)$ is the survival function given the biomarker value Y_i, $S_0(t)$ is the baseline survival function and β is a regression parameter, the U_i and W_i are exchangeable, because [34]

$$W_i < W_j \Longleftrightarrow S(t|Y_i) < S(t|Y_j) \Longleftrightarrow \int_0^\infty S(t|Y_i)dt < \int_0^\infty S(t|Y_j)dt \Longleftrightarrow$$
$$\int_0^\infty t \cdot f(t|Y_i)dt < \int_0^\infty t \cdot f(t|Y_j)dt \Longleftrightarrow U_i < U_j.$$

Therefore, $Pr(U_i < U_j | Z_i < Z_j) = Pr(W_i < W_j | Z_i < Z_j) = Pr(\beta Y_i > \beta Y_j | Z_i < Z_j)$. The C-index of 1 indicates that the model has a perfect discrimination power, whereas a value of 0.5 corresponds to an uninformative model.

Nonparametric version of estimation is possible for the C-index. A pair of subjects is said to be concordant if $(\beta Y_i > \beta Y_j \ , \ Z_i < Z_j)$ or $(\beta Y_i < \beta Y_j \ , \ Z_i > Z_j)$. In contrast, a pair $(\beta Y_i > \beta Y_j \ , \ Z_i > Z_j)$ or $(\beta Y_i < \beta Y_j \ , \ Z_i < Z_j)$ is said to be discordant. It is noted that not all pairs are usable to determine concordance and discordance. A pair is usable only when one subject has an event before the other experiences an event or censored. For example, we discard pairs if neither of subjects have events or two individuals have the same survival time. Let R be a set of all usable pairs and Q be a total number of usable pairs in R. The C-index is estimated by

$$\hat{C} = \frac{1}{Q} \sum_{(i,j) \in R} c_{ij},$$

where

$$c_{ij} = \begin{cases} 1 & \text{if} \quad (Z_i < Z_j \quad \text{and} \quad \beta Y_i > \beta Y_j) \quad \text{or} \quad (Z_i > Z_j \quad \text{and} \quad \beta Y_i < \beta Y_j) \\ 0 & \text{if} \quad (Z_i < Z_j \quad \text{and} \quad \beta Y_i < \beta Y_j) \quad \text{or} \quad (Z_i > Z_j \quad \text{and} \quad \beta Y_i > \beta Y_j). \end{cases}$$

The original C-index is investigated further to overcome shortcomings. Yan and Greene [35] found that the C-index depends on the number of tied pairs. Therefore, in the presence of large proportion of tied pairs, they recommended to report both C-indices with and without ties. Another modification is done by Uno et al. [36] for censored survival data. To overcome the C-index's dependence on the underlying censoring distribution, they presented the consistent estimates which is free of censoring by using an inverse probability weighting technique. The confidence interval for $\hat{C}$ is developed by Pencina and D'Agostino [34]. The

100(1 - α)% confidence interval is $\hat{C} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{C})}$,

$$\widehat{Var}(\hat{C}) = \frac{4}{N(p_c + p_d)^4}(p_d^2 p_{cc} - 2p_c p_d p_{cd} + p_c^2 p_{dd}),$$

$$p_c = \frac{1}{N(N-1)}\sum_i c_i \quad , \quad p_d = \frac{1}{N(N-1)}\sum_i d_i$$
$$p_{cc} = \frac{1}{N(N-1)(N-2)}\sum_i c_i(c_i - 1) \quad , \quad p_{dd} = \frac{1}{N(N-1)(N-2)}\sum_i d_i(d_i - 1)$$
$$p_{cd} = \frac{1}{N(N-1)(N-2)}\sum_i c_i d_i,$$

and c_i is the number of concordant pairs, d_i is the number of disconcordant pairs with the i^{th} subject in the sample, and $z_{\alpha/2}$ is (1 - $\alpha/2$) percentile of the standard normal distribution.

2.3.4 Time dependent ROC analysis

When a biomarker is used for a diagnosis of disease that changes over time, the original ROC analysis is no longer applicable. In the interval monitoring framework, DeLong et al. [37] and Parker and DeLong [38] developed the new ROC methodology using parameter estimates from discrete logistic regression. When continuous time to event data are available, however, time dependent ROC analysis could be performed. The time dependent ROC was introduced as an extension of the existing concept for sensitivity and specificity to survival outcome. We assume that the higher biomarker values are more indicative of shorter survival time. There are three different definitions of time dependent ROC.

A cumulative/dynamic time dependent ROC is used when the main question is whether the biomarker can distinguish the patients who have experienced the event by time t and who have not [39]. The cumulative case refers to the subject who has experienced the event during the time interval (0, t], whereas the dynamic control refers to the subject with no event by time t. With the cutoff point of c, sensitivity and specificity are defined as

$$Sensitivity(t) = Pr(Y > c | Z \leq t)$$
$$Specificity(t) = Pr(Y \leq c | Z > t).$$

The time dependent sensitivity and specificity are estimated using Kaplan-Meier estimator based on the subset of $Y \geq c$ or weighted Kaplan-Meier estimator based on nearest neighbor kernel. The confidence interval of time dependent ROC curve is calculated by bootstrap method. This ROC method can be used clinically when the sensitivity of standard and new diagnostic measures are compared at certain time points to check whether the new measure provides improved discrimination during the follow-up time. Later, the cumulative/dynamic time dependent ROC is generalized to longitudinal biomarker [40] and competing risk outcomes [41]. For the longitudinal biomarker, the question of interest is how well a biomarker measured at a certain time point after the baseline can discriminate diseased and non-diseased subjects in a subsequent time interval.

Alternative approach is an incident case and dynamic control time dependent ROC [42]. Under this definition, only subject who has an event at time t plays a role of case. The dynamic control corresponds to the subject who is event free by time t. The incident/dynamic time dependent sensitivity and specificity are defined as

$$Sensitivity(t) = Pr(Y > c|Z = t)$$
$$Specificity(t) = Pr(Y \leq c|Z > t).$$

The sensitivity can be estimated under proportional hazard model by computing the expected fraction of failures with a biomarker level greater than c. The specificity is estimated by the empirical distribution function for biomarker among those who survive beyond t. Bootstrap confidence interval can be constructed for nonparametric time dependent ROC. It is particularly useful when investigators want to display the incident discrimination ability over time. It is interesting to know that the C-index is a weighted average of the area under the incident/dynamic time dependent ROC [42].

The last version of time dependent ROC is defined with respect to incident case and static control [43] [44]. Defining the case as the subject who experiences an event at time t, the control is the subject who has not developed an event until a fixed time point at $t^\sharp$. Unlike the other two definitions, the incident/static ROC curve changes over time depending

only on the case group. The sensitivity and specificity is given as

$$Sensitivity(t) = Pr(Y > c|Z = t)$$
$$Specificity(t) = Pr(Y \leq c|Z > t^{\sharp}).$$

Zheng and Heagerty [45] estimated the incident/static time-dependent ROC curve by modeling the biomarker distribution conditional on the event status in a semiparametric way. The ROC curve was expressed as a function of location and scale parameters from the biomarker distribution. Incident/static ROC method is useful in a retrospective study especially when the time to event is certain. As an alternative estimation method, the direct regression approach of ROC curve was comprehensively reviewed and extended by Cai et al. [46] and Pepe et al. [47]. Confidence interval of the ROC curve can be based on bootstrap samples or asymptotic property under certain regularity condition.

In this dissertation, we focus on the cumulative/dynamic time dependent ROC. One of the questions addressed from our study is how well a biomarker can discriminate subjects who had an event until time point t and those who remained event free up to t. To measure a biomarker's discrimination potential for cumulative events by time t, which are time dependent measures, cumulative/dynamic time dependent ROC analysis may be more appropriate than others.

3.0 DISCRIMINANT ANALYSIS FOR CENSORED LONGITUDINAL BIOMARKER DATA

Discriminant analysis is commonly used to evaluate the ability of candidate biomarkers to separate patients into pre-defined groups. Extension of discriminant analysis to longitudinal data enables us to improve the classification accuracy based on biomarker profiles rather than on a single biomarker measurement. However, the biomarker measurement is often limited by the sensitivity of the given assay, resulting in data that are censored either at the lower or upper limit of detection. We develop a discriminant analysis method for censored longitudinal biomarker data based on mixed models. The biomarker performance is assessed by AUC. Through the simulation study, we show that our method is better than the simple substitution methods in terms of parameter estimation and evaluating biomarker performance. We apply our method to a biomarker study aiming to identify biomarkers that are predictive of the recovery from acute kidney injury for patients on renal replacement therapy.

3.1 INTRODUCTION

As a noninvasive and cost-effective tool for diagnosis and prognosis of acute and chronic diseases, biomarkers have received increasing attention for many decades. Two questions raised commonly in the biomarker studies are (1) how to classify subjects into disease and non-disease groups based on their measurements and (2) how to evaluate the clinical utility of the biomarker. For classification and evaluation, several methods such as logistic regression, discriminant analysis, and ROC curve have been widely applied for cross-sectional

data. However, more and more studies highlight the importance of the temporal change of biomarkers which can provide better understanding of the development of a disease [48]. Longitudinal biomarkers have been shown to lead to more accurate diagnosis than single measurement. For example, de Leon et al. [49] stated that the diagnostic accuracy of mild cognitive impairment is improved when longitudinal cerebrospinal fluid marker is used. If biomarkers are measured repeatedly over several time points, we may need specialized techniques for capturing important time-related patterns in the repeated measurements. The other concern in the biomarker study is the LOD. If an instrument is not sensitive enough to detect very high and low concentrations, only measurements which lie between lower and upper detection limits are observable. The results from inappropriate handling of these types of data may mislead physicians in medical decision making.

Our work is motivated from the Biological Markers of Recovery for the Kidney (BioMaRK) study. The recovery of a kidney function following the acute kidney injury (AKI) is an important determinant of morbidity and may have long-term implications for the health and well-being of patients [50]. Hence, identifying informative biomarkers for predicting a 60-day recovery is one of the primary goals of this study. There has been much effort in the biomarker discovery related to AKI due to its unacceptably high mortality rates. Most studies focus on evaluating biomarker performance based on a single measurement. Even when the biomarkers are measured over time, it is common to analyze the biomarker at each time point separately or choose arbitrarily a summary measure such as change score or slope to incorporate the longitudinal information. However, investigators are often more interested in the overall performance of the longitudinal biomarker because biomarker evolution can reveal better the biological process of a disease. In the BioMaRK study, multiple urinary biomarkers are longitudinally measured, and the measurements of some biomarkers are censored due to detection limits. The objective of our research is to develop a classification method for the longitudinal biomarkers subject to left or right censoring due to lower or upper detection limit.

We develop the new classification and evaluation methods to take both censoring and repeated measures into account. Discriminant analysis has been extended to the longitudinal setting with a discriminant function estimated from mixed models [51] [52] [53] [54]. Further

generalization to multivariate longitudinal data has been discussed by Marshall et al. [55] using multivariate nonlinear mixed models. Kohlmann et al. [56] introduced the longitudinal quadratic discriminant analysis and evaluated the classification performance using the ROC curve and Brier score. If longitudinal data are censored due to a detection limit, however, earlier proposed methods cannot produce an expected result. The problem of left-censoring has been studied by many researchers [4] [19] [18]. They considered maximum likelihood approaches to incorporate the censoring issue. As a related work to our objective from a discriminant analysis perspective, Langdon et al. [57] discussed how to classify subjects based on two censored variables. They estimated parameters by maximizing the marginal likelihood function of bivariate normal distribution. These estimates were plugged into the classifier formed by Bayes optimum decision rule. We extend the idea of Langdon et al. [57] to develop classification methods for longitudinal censored data, and show how AUC can be constructed from discriminant analysis.

The organization of this chapter is as follows. In section 3.2, we introduce the underlying theory of our discriminant analysis method. We describe how to classify subjects and how to evaluate biomarker performance in the presence of censoring. In section 3.3, we compare our method with simple substitution methods using simulated data. Finally, our method is applied to the BioMaRK study to predict a patient's recovery status from AKI within 60 days after the enrollment.

3.2 METHOD

3.2.1 Linear mixed model for biomarker data

High dimensionality, serial correlations, unbalanced or unequally spaced repeated measures, and missing data are typical issues that people encounter in the longitudinal analysis. Mixed model is one of the popular approach to handle these problems [58]. The linear mixed model captures the correlations between repeated measurements within a subject via random effects (also called subject specific effects). Also, it can accommodate missing data when the missing

and measurement processes are independent (missing completely at random; MCAR), or the missing process depends only on the observed measurements (missing at random; MAR). Let Y_{ij} be the biomarker measurement on the i^{th} individual at the j^{th} time point, ($i = 1,\cdots, N$; $j = 1,\cdots, n_i$). Thus, $Y_i = (Y_{i1}, \cdots, Y_{in_i})^t$ is an $n_i \times 1$ vector of measurements corresponding to the i^{th} subject. The linear mixed model relating Y_i to a set of covariates can be expressed in the matrix notation as

$$Y_i = X_i\beta + Z_i\gamma_i + e_i,$$

where X_i is an $n_i \times p$ design matrix of fixed effect, β is a $p \times 1$ population parameter vector, and Z_i is an $n_i \times q$ design matrix of random effect. Random error e_i and random effect γ_i are independent and normally distributed with $e_i \sim N(0, R_i)$ and $\gamma_i \sim N(0, G_i)$. Marginally, Y_i is normally distributed with mean $X_i\beta$ and covariance matrix $\Sigma_i = Z_iG_iZ_i^t + R_i$.

Parameters in the linear mixed model can be estimated from the likelihood function formulated given the random effects. A likelihood function is simplified based on the mixed model assumption that longitudinal observations are independent given the random effects. To handle the censoring of biomarker measurements, we use the method similar to Lyles et al. [19]. Suppose lower detection limit and upper detection limit are τ_{lo} and τ_{up}, respectively. The likelihood function is constructed using the normal density function $f(Y_{ij}|\gamma_i)$ for observed measurements and the cumulative distribution function $F(\tau_{lo}|\gamma_i)$ or $1 - F(\tau_{up}|\gamma_i)$ for censored parts. Let θ denote the vector of parameters in the covariance matrices. The final likelihood function for the covariance parameter vector θ and coefficient vector β is given by

$$L(\beta, \theta; Y) = \prod_{i=1}^{N} \left[\int_{R^q} \prod_{j=1}^{n_i} \{f(Y_{ij}|\gamma_i)^{I(d_{ij}=0)} F(\tau_{lo}|\gamma_i)^{I(d_{ij}=1)} \left(1 - F(\tau_{up}|\gamma_i)\right)^{I(d_{ij}=2)}\} f(\gamma_i) d\gamma_i \right],$$

$$d_{ij} = \begin{cases} 0 & \text{if } Y_{ij} \text{ is completely observed} \\ 1 & \text{if } Y_{ij} \text{ is left censored at } \tau_{lo} \\ 2 & \text{if } Y_{ij} \text{ is right censored at } \tau_{up}, \end{cases}$$

$I(\cdot)$ is an indicator function.

Once the likelihood function is defined depending on the censoring types, set of parameters θ and β are estimated by maximizing the likelihood function. Because the likelihood function includes cumulative distribution function to account for censored biomarker, we apply SAS procedure `Proc nlmixed` to obtain the estimates. The `Proc nlmixed` procedure allows us to specify the general form of distribution given the random effects. Integral approximation is done by an adaptive Gaussian quadrature method and the likelihood is maximized by dual quasi-Newton algorithm [59].

3.2.2 Discriminant analysis

We adopt the concept of discriminant analysis to construct the classifier using longitudinal censored biomarker measurements. Discriminant analysis arises from the desire to use an optimal classification rule, and it is often based on the assumption of normal distribution for two separate groups. Let $f_k(y)$ denote the normal density function (with mean μ^k and variance matrix Σ_k) of the longitudinal biomarker measurements for the subjects in group k ($k = 0, 1$). For a subject with biomarker data Y, the posterior probability of assigning the subject into group k is given by

$$Pr(Group_k|Y) = \frac{f_k(Y)\pi_k}{f_0(Y)\pi_0 + f_1(Y)\pi_1},$$

where π_k is the prior probability that a subject belongs to group k without the knowledge of Y. The ratio $Pr(Group_1|Y)/Pr(Group_0|Y)$ is then used as a discriminant function and compared with a pre-defined cutoff point to determine the group membership. Noting that the corresponding log ratio

$$log\frac{Pr(Group_1|Y)}{Pr(Group_0|Y)} = log\frac{f_1(Y)}{f_0(Y)} + log\frac{\pi_1}{\pi_0},$$

we refer to the first term as a risk score $S = log(f_1(Y)/f_0(Y))$. When two groups have same variance matrix (i.e. $\Sigma_0 = \Sigma_1 = \Sigma$), the risk score is simplified as

$$S = \left\{Y - \frac{1}{2}(\mu^1 + \mu^0)\right\}^t \Sigma^{-1}(\mu^1 - \mu^0). \tag{3.1}$$

When variance matrices for two groups have different forms (i.e. $\Sigma_0 \neq \Sigma_1$), the risk score is

$$S = \frac{Y^t \left({\Sigma_0}^{-1} - {\Sigma_1}^{-1}\right) Y}{2} + Y^t \left({\Sigma_1}^{-1}\mu^1 - {\Sigma_0}^{-1}\mu^0\right) - \frac{log|\Sigma_1|/|\Sigma_0|}{2}$$
$$- \frac{1}{2}\left(\mu^{1t}{\Sigma_1}^{-1}\mu^1 - \mu^{0t}{\Sigma_0}^{-1}\mu^0\right).$$

The distributional parameters can be estimated from the mixed model. The estimation of S for each individual depends on whether the subject has censored measurement or not. When Y is completely observed, the risk score S can be directly calculated from equation (3.2.2) using Y. If some components of Y are censored, we will substitute $f_k(Y)$ in equation (3.2.2) by $f_k^*(Y)$, defined as

$$f_k^*(Y) = \int_{R^q} \prod_{j=1}^{n_i} \{f_k(Y_{ij}|\gamma_i)^{I(d_{ij}=0)} F_k(\tau_{lo}|\gamma_i)^{I(d_{ij}=1)} \left(1 - F_k(\tau_{up}|\gamma_i)\right)^{I(d_{ij}=2)}\} f_k(\gamma_i) d\gamma_i$$

A new patient is classified by comparing his/her risk score to a pre-selected threshold. For example, if we use the cutoff point driven by the decision theory with Bayes 0-1 loss function, the subject is classified into group 1 if $\hat{S} > 0$, and classified into group 0, otherwise.

3.2.3 Evaluation of classification performance

AUC has long been defined for cross-sectional test results. Suppose variables T^0 and T^1 are the test results from normal and disease groups. The AUC is defined as $Pr(T^1 > T^0)$, that is, a probability that the test result for a randomly chosen individual with a disease is more indicative of that disease than the test result from a normal subject. The test result can be a continuous biomarker measurement. When the biomarker is measured over time, the test result is a multivariate measure. To summarize the classification performance of a multivariate test result to a univariate measure, we use the risk score S to serve as a test result in the AUC calculation. In the following, we show how non-parametric and parametric estimates of AUC are calculated based on the risk score S.

Nonparametric estimation of AUC

With risk score S used as a test result, we may define S^1 and S^0 as the risk scores for a randomly chosen subject from group 1 and group 0 respectively. Suppose there are n_0 controls and n_1 cases in the data set. The sensitivity and specificity based on the estimated risk score $\hat{S}$ can be formulated as

$$1 - Specificity = \frac{1}{n_0} \sum_{i=1}^{n_0} I\left(\hat{S^0}_i > c\right)$$
$$Sensitivity = \frac{1}{n_1} \sum_{j=1}^{n_1} I\left(\hat{S^1}_j > c\right),$$

for a threshold c. The empirical ROC curve is obtained by connecting points [sensitivity(c), 1-specificity(c)]. The posterior probabilities $Pr(Group_k|Y)$ have also been used to construct the empirical ROC curve [52] [55] [56] in the longitudinal discriminant analysis. Note that the posterior probability of belonging to group 1 is $Pr(Group_1|Y) = e^S \ / \ (e^S + 1)$, which is a monotone transformation of S. Thus these two approaches lead to the same AUC given the invariant property of ROC curve under monotone transformations. We can use trapezoids method to estimate AUC and the method by DeLong et al. [25] for variance estimation. In the presence of censoring, empirical AUC tends to produce lower values because it reflects the actual discrimination ability of incomplete information rather than a potential discrimination ability which could be achieved if LOD is eliminated.

Parametric estimation of AUC

In a special case when two groups have a common covariance matrix, the risk score has a linear form in terms of Y (3.1). Then the AUC can be estimated based on the distributional assumption of the longitudinal biomarker measurements. A smooth ROC curve can be obtained by

$$Sensitivity = \Phi\left(\frac{\lambda^t(\mu^1 - \mu^0)}{\sqrt{\lambda^t \Sigma \lambda}} + \Phi^{-1}(1 - Specificity)\right), \quad \text{with} \quad \lambda = \Sigma^{-1}(\mu^1 - \mu^0),$$

and the AUC is defined as

$$
\begin{aligned}
AUC &= Pr\left[S^1 > S^0\right] \\
&= Pr\left[\left\{Y^1 - \frac{1}{2}(\mu^1 + \mu^0)\right\}^t \Sigma^{-1}(\mu^1 - \mu^0) > \left\{Y^0 - \frac{1}{2}(\mu^1 + \mu^0)\right\}^t \Sigma^{-1}(\mu^1 - \mu^0)\right] \\
&= Pr\left[\left\{\Sigma^{-1}(\mu^1 - \mu^0)\right\}^t Y^1 > \left\{\Sigma^{-1}(\mu^1 - \mu^0)\right\}^t Y^0\right] \\
&= \Phi\left(\frac{\lambda^t(\mu^1 - \mu^0)}{\sqrt{\lambda^t(2\Sigma)\lambda}}\right), \qquad (3.2)
\end{aligned}
$$

where Φ is a standard normal cumulative distribution function. The ROC curve and AUC are estimated using $\hat{\mu}^0$, $\hat{\mu}^1$ and $\hat{\Sigma}$ from the linear mixed model. Because the AUC is entirely determined by parameters μ^1, μ^0, and Σ, it remains intact unless the estimates $\hat{\mu}^1$, $\hat{\mu}^0$, $\hat{\Sigma}$ are biased by the censored data. The standard error of AUC estimate can be calculated following McClish's method [60].

3.3 SIMULATION STUDY

We conduct simulation study to evaluate the performance of the proposed discrimination method. In practice, substitution methods are often used to handle the censored data due to detection limits. Usually the censored observations are replaced by LOD or LOD/2. We investigate how the discrimination measure AUC is affected by the censoring problem and under what scenarios the naive substitution methods tend to introduce significant bias. We also examine the impact of misspecification of covariance structure on the discrimination evaluation of longitudinal biomarker profile. In the simulation, we use the same set of subjects in the estimation of parameters and AUC, which tends to make classification accuracy overoptimistic. However, the bias of AUC due to this will be negligible for considered sample size and number of parameters.

We generate longitudinal biomarker measurement Y_{ij} for the subject i at time point T_{ij} from the mixed model:

$$Y_{ij} = \varphi_1 + \varphi_2 X_i + \varphi_3 T_{ij} + \varphi_4 (X_i \times T_{ij}) + a_i + b_i T_{ij} + e_{ij}, \tag{3.3}$$

where

$$e_{ij} \sim N(0, \sigma^2) \quad \text{and} \quad \begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \right].$$

X_i is a dichotomous variable, indicating the group membership (0 or 1) and T_{ij} is a time factor, indicating the follow up times of measurements (T_{ij} = 1,2,3,4). Random intercept a_i and random slope b_i are included in the model to reflect the deviation of the subject specific trajectory from the population trajectory. We assume that random effects are independent of the random error. Note that the classification performance of a biomarker is determined by the underlying separability in biomarker measurements between groups. The separability not only depends on the regression coefficient parameters which specify the difference in mean, but also the parameters in the covariance matrix. Larger variability in Y makes it more difficult to divide the two groups. We fix the regression parameters at $\varphi_1 = 1.0$, $\varphi_2 = \varphi_3 = 0.5$, $\varphi_4 =$ -1.0, and the covariance parameters at $\sigma^2 = 1.0$, $\sigma_{ab} = 0.0$. This corresponds to the scenario where the trajectories of two groups start at different baseline levels and increase over time for one group and decrease for another group. Moreover, the variabilities of biomarker measurements increase over time. To simulate biomarker data with different discrimination ability, we change the variance in Y through the variance parameters of random effects, i.e., $\sigma_a^2 = \sigma_b^2 = 0.5$, 1.0, and 2.0 with higher values representing poor separation between two groups. We choose lower detection limit τ_{lo} empirically so that the censoring rate of 20% and 40% can be achieved. We simulate 100 datasets, each including 200 subjects from individual group.

The parameter estimates from the linear mixed model as well as parametric estimates of AUCs, associated standard errors (se) and empirical 95% coverage probabilities (CP) are summarized in Table 2. Comparing to the omniscient estimates based on the uncensored complete data, the parameter estimates of the group (φ_2) and interaction (φ_4) effects are

Table 2: Parametric estimation of AUC: Comparison of the discriminant ability between proposed method (PM) and substitution methods (φ_2 = 0.5, φ_4 = -1.0, σ^2 = 1.0)

Pr(censor)	σ_a^2, σ_b^2	Method	$\hat{\varphi}_2$	$\hat{\varphi}_4$	$\hat{\sigma}^2$	$\hat{\sigma}_a^2$	$\hat{\sigma}_b^2$	$\hat{AUC}(se)$	CP
0.2	2.0	Omni	0.488	-0.991	1.001	1.950	2.014	0.687(0.026)	0.960
		LOD	0.140	-0.665	0.808	2.483	1.025	0.669(0.026)	0.890
		LOD/2	0.129	-0.611	0.765	2.218	0.914	0.664(0.026)	0.860
		PM	0.487	-0.991	1.003	1.927	2.012	0.688(0.026)	0.940
	1.0	Omni	0.489	-0.994	1.001	0.963	1.008	0.740(0.024)	0.960
		LOD	0.150	-0.673	0.782	1.266	0.525	0.709(0.025)	0.880
		LOD/2	0.139	-0.628	0.736	1.147	0.481	0.717(0.025)	0.840
		PM	0.487	-0.994	1.001	0.957	1.005	0.741(0.024)	0.950
	0.5	Omni	0.489	-0.995	1.001	0.472	0.505	0.802(0.021)	0.970
		LOD	0.168	-0.688	0.765	0.664	0.273	0.787(0.022)	0.900
		LOD/2	0.158	-0.654	0.725	0.616	0.257	0.783(0.022)	0.850
		PM	0.489	-0.996	1.002	0.469	0.504	0.803(0.021)	0.960
0.4	2.0	Omni	0.488	-0.991	1.001	1.950	2.014	0.687(0.026)	0.960
		LOD	0.198	-0.548	0.567	1.242	0.791	0.660(0.026)	0.810
		LOD/2	0.187	-0.556	0.589	1.351	0.803	0.661(0.026)	0.820
		PM	0.470	-0.987	1.001	1.911	2.056	0.685(0.026)	0.910
	1.0	Omni	0.489	-0.994	1.001	0.963	1.008	0.740(0.024)	0.960
		LOD	0.201	-0.551	0.551	0.616	0.409	0.709(0.025)	0.690
		LOD/2	0.181	-0.569	0.598	0.747	0.424	0.711(0.025)	0.740
		PM	0.488	-0.996	1.001	0.966	1.000	0.742(0.024)	0.920
	0.5	Omni	0.489	-0.995	1.001	0.472	0.505	0.802(0.021)	0.970
		LOD	0.207	-0.555	0.537	0.304	0.216	0.767(0.023)	0.500
		LOD/2	0.181	-0.587	0.611	0.431	0.228	0.773(0.023)	0.620
		PM	0.492	-0.999	0.999	0.479	0.501	0.804(0.021)	0.920

Table 3: Nonparametric estimation of AUC: Comparison of the discriminant ability between proposed method (PM) and substitution methods ($\varphi_2 = 0.5$, $\sigma^2 = 1.0$)

Pr(censor)	σ_0^2, σ_1^2	Method	$\hat{\varphi}_2$	$\hat{\sigma}^2$	$\hat{\sigma}_0^2$	$\hat{\sigma}_1^2$	$\hat{AUC}(se)$
0.2	0.5,1.0	Omni	0.492	1.002	0.498	1.000	0.638(0.023)
		LOD	0.438	0.763	0.273	0.711	0.636(0.023)
		PM	0.492	1.001	0.496	1.000	0.639(0.023)
	0.5,2.0	Omni	0.489	1.002	0.498	1.994	0.681(0.022)
		LOD	0.514	0.779	0.292	1.354	0.673(0.022)
		PM	0.485	0.994	0.483	1.999	0.677(0.022)
0.4	0.5,1.0	Omni	0.492	1.002	0.498	1.000	0.638(0.023)
		LOD	0.357	0.569	0.112	0.456	0.630(0.023)
		PM	0.488	1.001	0.490	1.000	0.634(0.023)
	0.5,2.0	Omni	0.489	1.002	0.498	1.994	0.681(0.022)
		LOD	0.464	0.589	0.117	0.896	0.662(0.022)
		PM	0.501	0.993	0.469	1.998	0.670(0.022)

heavily biased when the censored observations are replaced by LOD or LOD/2. The proposed method (PM) provides approximately unbiased estimates. As expected, the bias of estimates continuously acts on the discriminant analysis and attenuates the AUC. Substitution methods provide increasingly smaller AUCs with poor coverage probabilities as the censoring proportion is increased. Our method, however, presents comparable AUCs to the omniscient values, and the coverage probabilities are close to the nominal level of 0.95.

Generalizing the assumption on the variance matrix, we also estimate the AUC empirically as described in section 3.2.3. We make the variance matrices for two groups different by allowing two random effects in the linear mixed model:

$$Y_{ij} = \varphi_1 + \varphi_2 X_i + \varphi_3 T_{ij} + a_{0i} + a_{1i} + e_{ij},$$

where $e_{ij} \sim N(0, \sigma^2)$, $a_{0i} \sim N(0, \sigma_0^2)$ and $a_{1i} \sim N(0, \sigma_1^2)$. Two random intercepts a_{0i}, a_{1i} and error e_{ij} are assumed to be independent each other. The parameters are fixed at $\varphi_1 = \varphi_3 = 1.0$, $\varphi_2 = 0.5$, $\sigma^2 = 1.0$, and $(\sigma_0^2, \sigma_1^2) = (0.5, 1.0), (0.5, 2.0)$. Individual group includes 300 subjects, each subject having 3 longitudinal time points. Table 3 summarizes

Table 4: Performance measures and fit statistics from different models

Pr(censor)	σ_a^2, σ_b^2	Model	AIC	BIC	AUC(se)
0.2	2.0	True			0.687
		RI	6604.35	6628.30	0.782(0.022)
		RS	5686.15	5710.10	0.683(0.026)
		RI + RS	5599.84	5631.77	0.688(0.026)
	1.0	True			0.741
		RI	5991.66	6015.61	0.844(0.019)
		RS	5347.83	5371.77	0.732(0.024)
		RI + RS	5312.57	5344.50	0.741(0.024)
	0.5	True			0.803
		RI	5492.32	5516.27	0.898(0.015)
		RS	5083.71	5107.66	0.794(0.022)
		RI + RS	5070.54	5102.47	0.803(0.021)
0.4	2.0	True			0.687
		RI	5236.54	5260.49	0.772(0.023)
		RS	4630.07	4654.02	0.689(0.026)
		RI + RS	4568.18	4600.10	0.685(0.026)
	1.0	True			0.741
		RI	4836.06	4860.01	0.829(0.020)
		RS	4401.94	4425.89	0.736(0.024)
		RI + RS	4376.95	4408.88	0.742(0.024)
	0.5	True			0.803
		RI	4513.40	4537.35	0.884(0.016)
		RS	4228.21	4252.16	0.797(0.022)
		RI + RS	4219.54	4251.47	0.804(0.021)

the result averaged over 100 datasets. Unlike the parametric estimate of AUC, the empirical AUC estimate tends to be smaller than a potential AUC because the nonparametric ROC curve is constructed based on the individual's score. If the individual has censored data for at least one time point, his/her score is affected by incomplete measurements and its original discriminative potential is reduced. In contrast, the parametric AUC is close to the omniscient estimate which reflects the potential discrimination ability, because it does not depend on the individual's score, but depends on the mean and variance for each group.

To examine the impact of model misspecification on the biomarker evaluation, we generate the data from model (3.3), which is referred to as random intercept and slope (RI + RS) model, but fit the data with three different models: RI + RS model, random intercept (RI) only model, and random slope (RS) only model. Table 4 shows the AUC estimates, associated standard errors, and goodness-of-fit statistics, Akaike information criterion (AIC) and Bayesian information criterion (BIC), from three mixed models. Both RI only and RS only models yield biased AUC estimates. The RI model overestimates the AUC, while the RS model underestimates the AUC. Only the correct RI + RS model produces AUC estimate close to the true one. In practice when we rarely know the true model, how can we believe that we have a correct performance measure? It appears that the goodness of fit statistics such as AIC and BIC can be used as a general guideline. Overall, the model with better goodness of fit (smaller AIC and BIC) produces AUC estimate closer to the true one. Our results are consistent with what was pointed out by Kohlmann et al. [56], incorrect model specification may lead to spuriously better or worse performance measures.

3.4 APPLICATION TO BIOMARK STUDY

The Biological Markers of Recovery for the Kidney (BioMaRK) is an observational study conducted as an ancillary study of the NIDDK-funded Acute Renal Failure Trial Network (ATN) study. ATN study is a multicenter, prospective trial of two strategies for renal replacement therapy in critically ill patients with acute kidney injury (AKI) [61]. One of the primary goals of the BioMaRK study is to find biomarkers predictive for the recovery

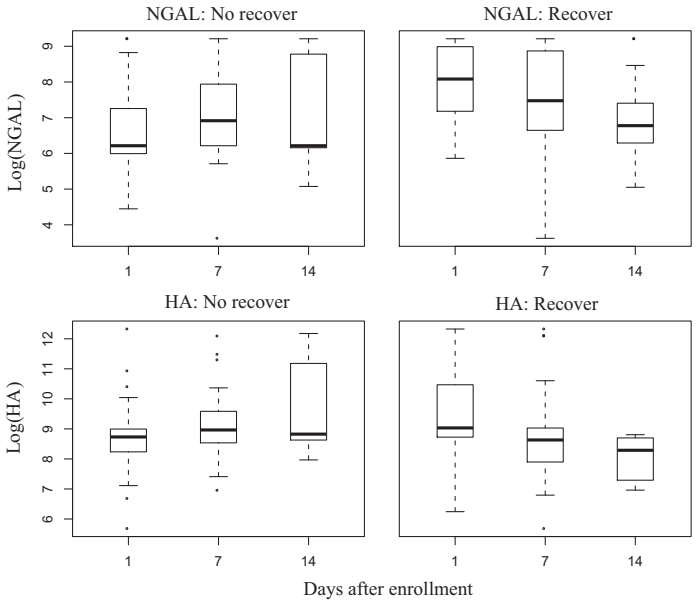

Figure 3: Boxplots of the log transformed NGAL and HA by recovery status

of renal function by 60 days after enrollment. The 'Recover' is defined as a survival with dialysis-independent renal function, and 'No-recover' indicates a death or dependence on dialysis. Serial measurements of plasma and urinary biomarkers are collected from the ATN study participants who signed the consent form for biomarker determination. For illustration purpose, we focus on the analysis of two urinary markers, Neutrophil Gelatinase-Associated Lipocalin (NGAL) and Hyaluronic Acid (HA), that are obtained for 76 patients at day 1, 7, and 14 after enrollment. Among 76 subjects, 38 (50%) recovered from AKI. NGAL is one of the widely used urinary biomarkers for prediction of AKI. HA is a new biomarker that is recently reported to correlate with both proteinuria and renal function in progressive renal disease. NGAL measurements are censored by two upper detection limits at 500 and 10000 ng/mg.Cr. The proportions of censoring at day 1, 7 and 14 are 25.4%, 34.0%, and 27.3%, respectively. HA level over 2029931.3 ng/mg.Cr is not measurable due to the detection limit. Most of HA levels are within the detection limit and only 2.7% are censored at day 1. We take a log transformation for NGAL and HA measurements to normalize the distribution.

Figure 3 presents the group-level boxplots of NGAL and HA on a log scale over three time points, where the censored observations are replaced with the detection limit. It appears that both HA and NGAL levels go up a little at day 7 for the non-recovery group, but go down over time for the recovery group. We consider several candidate models of different covariance matrices (RI only, RS only, and RI + RS model) and different form of group-specific trajectories. We choose the final model with the smallest AIC and BIC as follows.

$$Y_{ij} = \varphi_1 + \varphi_2 Recover_i + \varphi_3 Time_{ij} + \varphi_4 Recover_i \times Time_{ij} + a_i + e_{ij}$$

where

$$e_{ij} \sim N(0, \sigma^2) \quad , \quad a_i \sim N(0, \sigma_a^2)$$

$$Recover_i = \begin{cases} 0 & \text{if subject } i \text{ didn't recover within 60 days} \\ 1 & \text{if subject } i \text{ recovered within 60 days} \end{cases}$$

The parameter estimates are shown in Table 5. The HA level correlates with the group membership a little stronger than the NGAL does, as indicated by the magnitude and significance of the group effect and group by time interaction effect. It gives an evidence that HA may have better discriminant ability than NGAL. The parametric AUC estimate (standard error) for NGAL is 0.822 (0.047), and for HA is 0.853 (0.043) (Figure 4 left: the black solid line and blue dotted line is for NGAL and HA, respectively). Substitution method using LOD/2 produces AUC estimates of 0.612 (0.063) for NGAL and 0.841 (0.040) for HA. We also perform cross sectional analysis to examine the discrimination ability of single biomarker measurement. The AUCs for NGAL day 1, day 7, and day 14 measurements are 0.662, 0.519, and 0.729 respectively. The corresponding AUCs for HA are 0.659, 0.563, and 0.849. Clearly, the discrimination performance is significantly improved by using the biomarker profile rather than the measurement on a single day.

To assess the prediction capacity of the biomarkers, we take the risk distribution into account using a predictiveness curve presented by Huang et al. [21]. Predictiveness curve is

Table 5: Parameter estimates and standard errors from the linear mixed models for NGAL and HA

	NGAL		HA	
parameter	estimate (se)	P-value	estimate (se)	P-value
φ_1	6.960 (0.310)	<0.0001	8.746 (0.321)	<0.0001
φ_2	1.107 (0.440)	0.014	1.363 (0.460)	0.004
φ_3	0.064 (0.039)	0.109	0.071 (0.040)	0.083
φ_4	-0.189 (0.051)	0.001	-0.237 (0.054)	<0.0001
σ_a^2	1.202		1.106	
σ^2	1.816		2.254	

a plot of predicted risk against the percentile of the biomarker. It has been pointed out in many papers that the biomarker with a strong predictive capacity has a steeper curve that corresponds to a wide variation in risk [21] [30] [31]. If the predictiveness curve is close to the horizontal line, the biomarker is no more helpful for making a medical decision. In Figure 4 (right: the black solid line and blue dotted line is for NGAL and HA, respectively), marker HA is more predictive than marker NGAL because more subjects are classified into high or low end of the risk. For example, the subjects in the bottom 10% of the risk score (S) distribution have recovery probabilities in the range of (0.02, 0.19) according to marker HA, but in a higher range (0.11, 0.23) according to marker NGAL. In the same way, the subjects above 90^{th} percentile of the distribution show higher recovery probabilities as predicted by HA (0.91, 0.97) than NGAL (0.80, 0.92).

3.5 DISCUSSION

We propose a new discriminant analysis method to incorporate censored longitudinal biomarker data. In the simulation study, we show that the substitution methods yield biased parameter estimates and different discrimination results. The bias of our method is almost ignorable, and the performance is satisfactory. The empirical AUC computed from the risk score reflects the actual discrimination power of longitudinal censored biomarkers. This AUC nat-

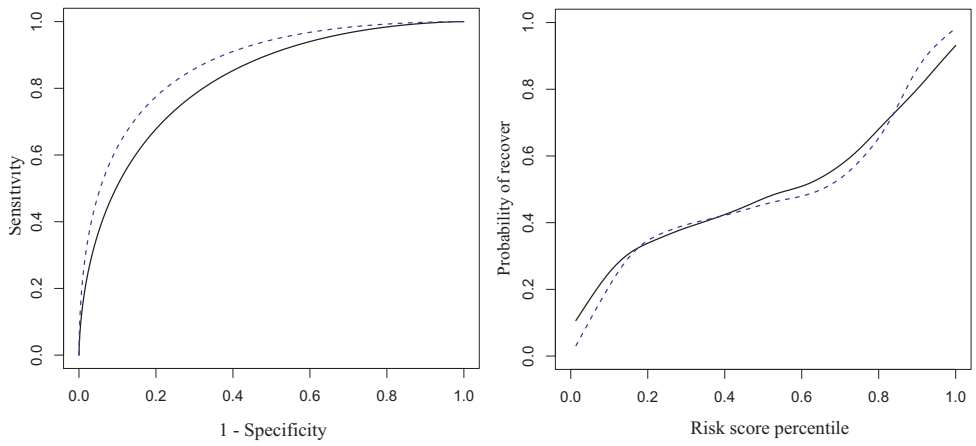

Figure 4: ROC curve (left) and predictiveness curve (right) for NGAL and HA.

urally tends to be lower than the discrimination ability achievable with completely observed biomarker measurements. It is also noted that the model selection is important to correctly evaluate the biomarkers. Thus we recommend selecting the linear mixed model by comparing fit statistics. Our methodology can be widely applied to clinical decision-making when it is necessary to handle below or above the threshold values, such as an investigation of health effects from chronic low-level exposures [62].

The proposed classification method is based on the discriminant analysis that assumes normal distribution for the longitudinal measurements. However, the biomarker data are often highly skewed. Box-Cox transformation is an effective tool to make the distribution of data close to normal. Several papers have discussed the Box-Cox transformation for censored data [63] [64] and for correlated data [65] [66]. Application of Box-Cox transformation to censored multivariate data merits further research. We show discriminant analysis under equality assumption of variance matrix for two groups and later generalize the assumption. The generalized method is more applicable and flexible, but increasing number of parameters may cause a computational issue if the sample size is not large enough. Although we introduce the discriminant analysis based on a single censored longitudinal biomarker, the

extension to the multiple censored longitudinal biomarkers is possible because a complex covariance structure can be used in the linear mixed model to account for the correlations within subjects and between biomarkers.

4.0 BEST LINEAR COMBINATION FOR LONGITUDINAL BIOMARKER DATA

The diagnostic performance of a biomarker is commonly assessed by AUC. Estimation of AUC is often complicated if biomarkers are collected over time, and even censored by the sensitivity limitation of a given assay. For a practical biomarker evaluation, we extend the linear combination method by Su and Liu [1] to censored longitudinal biomarker data. The combination coefficient derived from this method enables us to calculate maximum AUC that can be obtained from fully observed data. Moreover, it can inform investigators which time point is more important in making a medical decision. Simulation studies demonstrate that the proposed method performs better than LOD/2 substitution method. Application is presented for the GenIMS study to evaluate inflammatory and coagulation biomarkers.

4.1 INTRODUCTION

Biomarkers are biochemical, genetic and molecular factors which can monitor biological and pathological processes in human body. Because biomarkers are measurable before a disease is clinically detected, they can help to make an important decision and reduce medical costs. Scientists have made a big effort to develop new biomarkers for several indications such as cancer and Alzheimer's disease. One of the important processes in biomarker research is a performance evaluation. In many cases, ROC analysis is used to measure diagnostic accuracy of a biomarker. As a summary measure of the ROC curve, AUC for a single time point measurement has well been developed in both parametric and non-parametric ways. For the evaluation of repeated measures, researchers have tried to use conditional

probability and posterior probability calculated from Bayesian model or latent class model [67] [68] [44] [69]. However, some methods are only applicable for specific conditions and any standard approaches for longitudinal data have not emerged yet. One way to incorporate high dimensional data is to use a linear combination so that we can use the original definition of ROC curve. Researchers frequently would like to discover a biomarker which provides higher AUC. With the investigators' goal of biomarker discovery in mind, an interesting question at this point is how to condense longitudinal measurements and maximize AUC at the same time. Besides the high dimension problem, researchers face another obstacle when biomarkers are subject to detection limit. It is clear that results are biased if censoring is not appropriately handled. Our goal is to find a linear combination coefficient for each time point to produce the best AUC.

The idea of our method is based on the linear combination of multiple markers to find an ideal biomarker which has high sensitivity and specificity. The linear combination methods were explored by researchers to maximize the sensitivity over the entire specificity range [1] [70], over a range of high specificity [71] and at a fixed specificity [72]. Whereas many approaches are based on the distributional assumption, Pepe and Thompson [73] proposed the distribution free approach to optimize AUC. However, all of these methods are only applicable for fully observed data. For censored measurements due to LOD, a parametric ROC method was investigated by Perkins et al. [27] [28]. They calculated AUC formulated by parameters estimated from the modified likelihood function for the censored observations. In the following sections, we present the formula for the best linear combination coefficient in terms of parameters of binormal distribution. We show how to estimate the parameters from the linear mixed model for both biomarkers with and without LOD. Our method is applied to the Genetic and Inflammatory Markers of Sepsis (GenIMS) study examining biomarker's discrimination power for the 90-day mortality for patients with community-acquired pneumonia.

4.2 METHOD

4.2.1 Best linear combination of longitudinal biomarker data

If biomarker concentrations are measured over time from the same individual, multiple within-person level data are collected. Suppose the biomarker Y^0 from non-disease group with n_0 subjects and Y^1 from disease group with n_1 subjects are expanded to p-dimensional vectors. Assume $Y^0 = (y_1^0, \cdots, y_p^0)^t \sim MVN_p(\mu^0, \Sigma_0)$ and $Y^1 = (y_1^1, \cdots, y_p^1)^t \sim MVN_p(\mu^1, \Sigma_1)$, where Σ_1 and Σ_0 are $p \times p$ positive definite matrices. With a linear combination coefficient $\lambda = (\lambda_1, \cdots, \lambda_n)$, we can make one-dimensional scores $\lambda^t Y^0 \sim N(\lambda^t \mu^0, \quad \lambda^t \Sigma_0 \lambda)$ and $\lambda^t Y^1 \sim N(\lambda^t \mu^1, \quad \lambda^t \Sigma_1 \lambda)$. Now, the condensed measures $\lambda^t Y^0$ and $\lambda^t Y^1$ take a role of continuous test results in ROC analysis. The AUC is calculated based on the distributional assumption :

$$AUC = Pr(\lambda^t Y^0 < \lambda^t Y^1) = \Phi\left(\frac{\lambda^t(\mu^1 - \mu^0)}{\sqrt{\lambda^t(\Sigma_0 + \Sigma_1)\lambda}}\right),$$

where Φ is standard normal cumulative distribution function. Our objective is to find the combination coefficient λ which maximizes the AUC. Since Φ is a strictly increasing function, the maximization of AUC is equivalent to

$$\max_{\lambda}\left[\frac{\lambda^t(\mu^1 - \mu^0)(\mu^1 - \mu^0)^t\lambda}{\lambda^t(\Sigma_0 + \Sigma_1)\lambda}\right]. \tag{4.1}$$

The combination coefficient which maximizes (4.1) is an eigenvector of $(\Sigma_0 + \Sigma_1)^{-1}(\mu^1 - \mu^0)(\mu^1 - \mu^0)^t$. What we actually care about λ is only a direction, not a magnitude. For example, λ=(1, 2, 3) gives the same AUC with λ=(2, 4, 6). In other words, AUC is maximized when λ is proportional to $(\Sigma_0 + \Sigma_1)^{-1}(\mu^1 - \mu^0)$. This is the same result given by Su and Liu [1] and Liu et al. [71].

4.2.2 AUC estimation for the best linear combination

Based on the best linear combination coefficient, we can obtain a smooth ROC curve, that is,

$$Sensitivity = \Phi\left(\frac{\lambda^t(\mu^1-\mu^0)+\Phi^{-1}(1-Specificity)\sqrt{\lambda^t\Sigma_0\lambda}}{\sqrt{\lambda^t\Sigma_1\lambda}}\right),$$

where $\lambda = (\Sigma_0+\Sigma_1)^{-1}(\mu^1-\mu^0)$. The corresponding point estimate of the optimum AUC is,

$$\widehat{AUC}_{opt} = \Phi\left(\frac{\hat{u}}{\sqrt{1+\hat{v}^2}}\right), \quad \text{where} \quad u = \frac{\lambda^t(\mu^1-\mu^0)}{\sqrt{\lambda^t\Sigma_1\lambda}}, \quad v = \frac{\sqrt{\lambda^t\Sigma_0\lambda}}{\sqrt{\lambda^t\Sigma_1\lambda}}. \tag{4.2}$$

To estimate parameters μ^0, μ^1, Σ_0 and Σ_1, we fit the linear mixed model accounting for correlations between repeated measures. Suppose the subject i ($i = 1,\cdots, N$) has $p\times 1$ vector of longitudinal biomarker measurements $Y_i = (Y_{i1},\cdots, Y_{ip})^t$. The linear mixed model we consider is

$$Y_i = X_i\beta + Z_i\gamma_i + e_i,$$

where X_i is an $p\times r$ matrix of fixed effect, β is a $r\times 1$ parameter vector, and Z_i is an $p\times q$ matrix of random effect. Random error e_i and random effect γ_i are independent and normally distributed. If a biomarker measurement is not censored due to LOD, its contribution to the likelihood function is through the normal density function $f(Y_{ij}|\gamma_i)$. Otherwise, the cumulative distribution function $F(\tau_{lo}|\gamma_i)$ or $1-F(\tau_{up}|\gamma_i)$ are used for left or right censored parts, where τ_{lo} and τ_{up} are lower and upper detection limit, respectively. The likelihood function for the covariance parameter vector θ and coefficient parameter vector β is given by

$$L(\beta,\theta;Y) = \prod_{i=1}^{N}\left[\int_{R^q}\prod_{j=1}^{p}\{f(Y_{ij}|\gamma_i)^{I(d_{ij}=0)}F(\tau_{lo}|\gamma_i)^{I(d_{ij}=1)}\left(1-F(\tau_{up}|\gamma_i)\right)^{I(d_{ij}=2)}\}f(\gamma_i)d\gamma_i\right],$$

$$d_{ij} = \begin{cases} 0 & \text{if } Y_{ij} \text{ is completely observed} \\ 1 & \text{if } Y_{ij} \text{ is left censored at } \tau_{lo} \\ 2 & \text{if } Y_{ij} \text{ is right censored at } \tau_{up}, \end{cases}$$

$I(\cdot)$ is an indicator function.

The confidence interval for AUC is constructed using asymptotic normality property of maximum likelihood estimate. The 100(1 - α)% two-sided confidence interval is $\widehat{AUC} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\widehat{AUC})}$. The variance of $\widehat{AUC}$ is estimated by [60]

$$\widehat{Var}(\widehat{AUC}) = \left(\frac{\partial\widehat{AUC}}{\partial\Delta}\right)^2 \widehat{Var}(\hat{\Delta}) + \left(\frac{\partial\widehat{AUC}}{\partial\sigma_0^2}\right)^2 \widehat{Var}(\hat{\sigma_0^2}) + \left(\frac{\partial\widehat{AUC}}{\partial\sigma_1^2}\right)^2 \widehat{Var}(\hat{\sigma_1^2}),$$

where $\Delta = \lambda^t\mu^1 - \lambda^t\mu^0$, $\sigma_k^2 = \lambda^t\Sigma_k\lambda$ $(k = 0, 1)$, $N = n_0 + n_1$ and

$$\frac{\partial\widehat{AUC}}{\partial\Delta} = \frac{e^{-\hat{u}^2/(2+2\hat{v}^2)}}{\sqrt{2\pi(1+\hat{v}^2)\hat{\sigma_1^2}}}$$

$$\frac{\partial\widehat{AUC}}{\partial\sigma_0^2} = -\frac{\hat{u}\hat{v}e^{-\hat{u}^2/(2+2\hat{v}^2)}}{2\hat{\sigma}_0\hat{\sigma}_1\sqrt{2\pi(1+\hat{v}^2)^3}}$$

$$\frac{\partial\widehat{AUC}}{\partial\sigma_1^2} = -\left(\frac{\hat{u}}{2\hat{\sigma}_1}\right)\left(\frac{\partial\widehat{AUC}}{\partial\Delta}\right) - \hat{v}^2\left(\frac{\partial\widehat{AUC}}{\partial\sigma_0^2}\right)$$

$$\widehat{Var}(\hat{\Delta}) = \frac{\hat{\sigma_1^2}}{n_0} + \frac{\hat{\sigma_0^2}}{n_1}$$

$$\widehat{Var}(\hat{\sigma_0^2}) = \frac{2\hat{\sigma_0}^4}{n_0 - 1}$$

$$\widehat{Var}(\hat{\sigma_1^2}) = \frac{2\hat{\sigma_1}^4}{n_1 - 1}.$$

In practice, the value of AUC is bounded by 0.5 and 1.0. If a true optimum AUC approaches 0.5 or 1.0, the distribution of estimated AUC becomes skewed, especially under a small sample size. It is well known that Fisher's Z-transformation improves the behavior of confidence intervals when assessed by coverage rate. Even though the transformed AUC does not strictly follow normal distribution, it tends to become normal rapidly with a small increase of a sample size. The 100(1 - α)% confidence interval for Z-transformed AUC_z is $\widehat{AUC}_z \pm z_{\alpha/2}\sqrt{\widehat{Var}(\widehat{AUC}_z)}$, where

$$AUC_z = ln\left(\frac{1+AUC}{1-AUC}\right) \quad , \quad Var(\widehat{AUC}_z) = \frac{4}{(1-AUC^2)^2}Var(\widehat{AUC}),$$

and $z_{\alpha/2}$ is (1 - $\alpha/2$) percentile of the standard normal distribution. The back transformation $(e^{AUC_z}-1)/(e^{AUC_z}+1)$ enables us to construct the confidence interval for AUC in the original scale [70] [60].

We can notice that the AUC evaluating the performance of linear discriminant analysis for longitudinal biomarker data is equivalent to the optimum AUC derived from linear combination coefficient under a certain condition. Suppose the covariance matrices for two groups are same, i.e. $\Sigma_0 = \Sigma_1 = \Sigma$. In section 3.2.3, we have shown that AUC from the linear discriminant analysis is

$$Pr\left[\left\{\Sigma^{-1}(\mu^1 - \mu^0)\right\}^t Y^1 > \left\{\Sigma^{-1}(\mu^1 - \mu^0)\right\}^t Y^0\right]. \tag{4.3}$$

Notice that the linear combination coefficient λ is proportional to $\Sigma^{-1}(\mu^1 - \mu^0)$. We can rewrite (4.3) in terms of λ as $Pr\left[\lambda^t Y^1 > \lambda^t Y^0\right] = \Phi\left(\frac{\lambda^t(\mu^1-\mu^0)}{\sqrt{\lambda^t(2\Sigma)\lambda}}\right)$, which has the same form as that of the maximum AUC (4.2). Because this AUC depends only on the mean and variance for each group, it remains intact for the censored biomarkers as long as the parameter estimates are consistent.

4.3 SIMULATION STUDY

The linear combination coefficient and the corresponding AUC depend on μ^0, μ^1, Σ_0 and Σ_1. When biomarkers are subject to LOD, conventional approach (so called substitution method) have been widely used, which substitutes censored data to LOD or LOD/2 level. However, we recommend to estimate parameters using the modified likelihood function. In this simulation study, we compare the performance of our method and substitution method by calculating the bias, standard error and coverage probability.

We simulate longitudinal biomarker data from two different random coefficient models to alter the covariance matrix. The data for the first simulation are generated from the random intercept model,

$$Y_{ij} = \varphi_1 + \varphi_2 X_i + \varphi_3 T_{ij} + \varphi_4 X_i \times T_{ij} + a_i + e_{ij}, \quad \text{where} \quad e_{ij} \sim N(0, \sigma^2) \quad \text{and} \quad a_i \sim N(0, \sigma_a^2),$$

where X_i is the categorical variable of group membership (0 or 1) and T_{ij} is follow up time point (T_{ij} = 1,2,3,4). The correlation between two measurements is $\rho = \sigma_a^2/(\sigma^2 + \sigma_a^2)$. The parameters are fixed at φ_1 = 1.0, φ_2 = 0.9, φ_3 = 0.1, φ_4 = -0.2, σ^2 = 1.0. The σ_a^2 are

Table 6: Optimum AUC for the best linear combination of longitudinal biomarker measurements that are generated from the random intercept model.

Pr(censor)	ρ	AUC_{true}	AUC_{opt}	SE	Std	CP	Coefficient
Proposed method							
0.1	0.2	0.695	0.693	0.016	0.016	0.950	(0.25, 0.15, 0.05, -0.05)
	0.5	0.657	0.656	0.017	0.017	0.935	(0.19, 0.09, -0.01, -0.11)
	0.8	0.635	0.633	0.017	0.017	0.940	(0.16, 0.06, -0.04, -0.14)
0.2	0.2	0.695	0.693	0.016	0.016	0.930	(0.25, 0.15, 0.05, -0.05)
	0.5	0.657	0.656	0.017	0.017	0.930	(0.19, 0.09, -0.01, -0.11)
	0.8	0.635	0.634	0.017	0.017	0.945	(0.16, 0.06, -0.04, -0.14)
0.3	0.2	0.695	0.693	0.016	0.016	0.945	(0.25, 0.15, 0.05, -0.05)
	0.5	0.657	0.656	0.017	0.017	0.935	(0.19, 0.09, -0.01, -0.11)
	0.8	0.635	0.634	0.017	0.018	0.935	(0.16, 0.06, -0.04, -0.14)
0.4	0.2	0.695	0.693	0.016	0.017	0.940	(0.25, 0.15, 0.05, -0.05)
	0.5	0.657	0.656	0.017	0.018	0.910	(0.19, 0.09, -0.01, -0.11)
	0.8	0.635	0.634	0.017	0.019	0.890	(0.16, 0.06, -0.04, -0.14)
LOD/2							
0.1	0.2	0.695	0.691	0.017	0.016	0.970	(0.27, 0.16, 0.06, -0.05)
	0.5	0.657	0.652	0.017	0.016	0.950	(0.20, 0.10, -0.01, -0.11)
	0.8	0.635	0.626	0.016	0.017	0.895	(0.16, 0.06, -0.04, -0.14)
0.2	0.2	0.695	0.689	0.017	0.016	0.940	(0.27, 0.16, 0.06, -0.05)
	0.5	0.657	0.650	0.017	0.017	0.930	(0.21, 0.10, -0.01,-0.11)
	0.8	0.635	0.623	0.015	0.017	0.845	(0.17, 0.07, -0.04,-0.14)
0.3	0.2	0.695	0.687	0.017	0.016	0.930	(0.27, 0.17, 0.06, -0.05)
	0.5	0.657	0.648	0.016	0.016	0.910	(0.21, 0.10, -0.01,-0.11)
	0.8	0.635	0.620	0.015	0.017	0.795	(0.17, 0.07, -0.04,-0.14)
0.4	0.2	0.695	0.683	0.016	0.016	0.890	(0.27, 0.17, 0.06, -0.04)
	0.5	0.657	0.644	0.016	0.017	0.865	(0.21, 0.10, -0.04,-0.11)
	0.8	0.635	0.616	0.014	0.017	0.725	(0.17, 0.07, -0.04,-0.14)

varied so that ρ attains the value 0.2, 0.5 and 0.8. We generate 200 datasets, each having 500 diseased subjects and 500 normal subjects. A subject is assumed to have left censored longitudinal biomarker measurements with censoring rates from 10% to 40%.

Table 6 summarizes the estimation results of optimum AUC (AUC_{opt}) averaged over 200 datasets from the proposed method and LOD/2 substitution. The AUC_{true} is calculated using the true parameter values from which the data are generated. The SE indicates the average of the estimated standard error. As an empirical standard error (Std), the standard deviation of $\widehat{AUC}_{opt}$, is presented. The estimated optimum AUC from our method is very close to the true AUC, and the standard error is comparable to the empirical standard error. Up to the censoring proportion of 30%, empirical coverage probabilities (CP) are near to 0.95. The bias of AUC from the LOD/2 method gets larger with the increase of censoring proportion and the coverage probabilities are generally worse than the proposed method. Irrespective of the methods, the measurement at the first time point has the highest weight in the combination. This is not surprising because the trajectory pattern we simulated for the two groups starts far apart each other and becomes narrower in the later time points. Besides, the correlation between two consecutive measurements are same over time. Accordingly, the contribution of biomarkers at the first time point to the discrimination ability is the most important.

In the second simulation, we simulate the data from the random intercept and slope model,

$$Y_{ij} = \varphi_1 + \varphi_2 X_i + \varphi_3 T_{ij} + a_i + b_i T_{ij} + e_{ij},$$
$$\text{where} \quad e_{ij} \sim N(0, \sigma^2) \quad \text{and} \quad \begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_a^2 & \omega\sigma_a\sigma_b \\ \omega\sigma_a\sigma_b & \sigma_b^2 \end{pmatrix} \right].$$

In this model, correlations between two measurements are not fixed, but different depending on the time points. The covariance parameters σ_a^2 and σ_b^2 are set to 1.0 with other coefficient parameters being remained same as before. We change the correlation between the random intercept and random slope ω from -0.8 to 0.8. Total 100 datasets are generated with 200 subjects in each group.

Table 7: Optimum AUC for the best linear combination of longitudinal biomarker measurements that are generated from the random intercept and slope model.

Pr(censor)	ω	AUC_{true}	AUC_{opt}	SE	Std	CP	Coefficient
Proposed method							
0.1	-0.8	0.811	0.810	0.021	0.020	0.960	(0.41, 0.24, 0.06,-0.11)
	0.0	0.728	0.728	0.024	0.022	0.950	(0.25, 0.13, 0.01,-0.11)
	0.8	0.723	0.723	0.025	0.023	0.950	(0.29, 0.14, 0.00,-0.16)
0.2	-0.8	0.811	0.810	0.021	0.020	0.950	(0.41, 0.24, 0.06,-0.11)
	0.0	0.728	0.729	0.024	0.023	0.970	(0.25, 0.13, 0.01,-0.11)
	0.8	0.723	0.722	0.025	0.024	0.930	(0.29, 0.14, 0.00,-0.16)
0.3	-0.8	0.811	0.810	0.021	0.022	0.950	(0.41, 0.24, 0.06,-0.11)
	0.0	0.728	0.728	0.024	0.023	0.970	(0.25, 0.13, 0.01,-0.11)
	0.8	0.723	0.724	0.025	0.025	0.930	(0.29, 0.14,-0.01,-0.16)
0.4	-0.8	0.811	0.810	0.021	0.023	0.920	(0.41, 0.24, 0.07,-0.11)
	0.0	0.728	0.728	0.024	0.026	0.940	(0.25, 0.13, 0.01,-0.11)
	0.8	0.723	0.723	0.025	0.026	0.810	(0.29, 0.14,-0.01,-0.16)
LOD/2							
0.1	-0.8	0.811	0.808	0.023	0.020	0.980	(0.37, 0.22, 0.06,-0.10)
	0.0	0.728	0.728	0.028	0.023	1.000	(0.25, 0.13, 0.01,-0.11)
	0.8	0.723	0.723	0.026	0.023	0.970	(0.29, 0.14,-0.01,-0.16)
0.2	-0.8	0.811	0.805	0.023	0.021	0.980	(0.30, 0.18, 0.05,-0.07)
	0.0	0.728	0.721	0.030	0.024	0.980	(0.18, 0.09, 0.01,-0.08)
	0.8	0.723	0.720	0.028	0.025	0.980	(0.27, 0.13,-0.01,-0.16)
0.3	-0.8	0.811	0.800	0.023	0.024	0.950	(0.25, 0.15, 0.05,-0.05)
	0.0	0.728	0.721	0.030	0.024	0.980	(0.18, 0.09, 0.01,-0.08)
	0.8	0.723	0.717	0.029	0.026	0.990	(0.23, 0.11,-0.01,-0.13)
0.4	-0.8	0.811	0.792	0.024	0.025	0.860	(0.20, 0.12, 0.04,-0.04)
	0.0	0.728	0.713	0.031	0.026	0.950	(0.14, 0.07, 0.01,-0.06)
	0.8	0.723	0.706	0.031	0.028	0.930	(0.17, 0.08,-0.01,-0.10)

From the estimates of AUC, SE, and coverage probability in Table 7, we observe similar results as we see for the mixed model with random intercept only. We can notice that the linear combination coefficients are not the same across different time points. The highest and lowest absolute weights are assigned to the measurements from the first and third time points, respectively.

4.4 APPLICATION TO GENIMS STUDY

The Genetic and Inflammatory Markers of Sepsis (GenIMS) study is a multicenter cohort study of patients admitted to the emergency department with community acquired pneumonia (CAP) between 2001 and 2003. The patients with CAP often experience severe sepsis and infection related death, which costs $8.4 billion each year in the United States [74] [75]. Because these patients are reported to exhibit abnormal levels in biomarkers of inflammation and coagulation, further investigation in the biomarkers helps physicians to efficiently manage CAP. The study enrolled 2320 patients and their biomarker levels were measured daily during the first seven days of hospitalization. The secondary objective of the study is to identify biomarkers which predict patient's death by 90 days after hospitalization. We select the pro-inflammatory marker, interleukin-6 (IL-6) and the coagulation marker, D-dimer for illustration. All biomarker measurements are transformed in a log scale to normalize the distribution. The IL-6 has two lower detection limits at 2 pg/mL and 5 pg/mL. Censoring proportions from day 1 to day 7 are 15.1%, 22.3%, 29.6%, 33.2%, 34.4%, 35.1%, and 34.2%. Boxplots are presented in Figure 5 to show a group-specific trajectory. Although the boxplots overlap between survival and death groups, IL-6 profile for the death group is higher than that for the survival group. In both groups, IL-6 levels fall rapidly from day 1 to day 2, but drop slowly after day 3. Based on this profile, we include the variable 'Day after hospitalization' as a reciprocal form. The estimated linear mixed model for log(IL-6) is

$$Y_{ij} = 1.463 + 1.104X_i + 2.139T_{ij} + a_i + b_iT_{ij} + e_{ij},$$

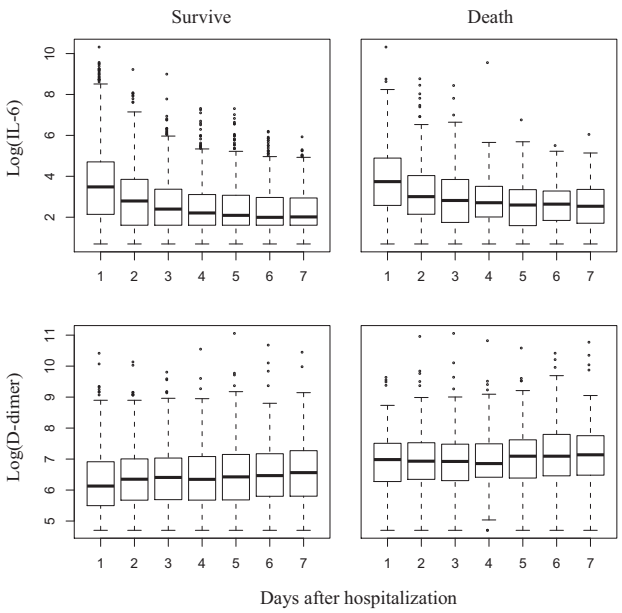

Figure 5: Boxplots for log transformed IL-6 and D-dimer by survival and mortality groups

where

$$e_{ij} \sim N(0, 0.551) \quad \text{and} \quad \begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2.127 & -1.389 \\ -1.389 & 5.031 \end{pmatrix} \right]$$

$$X_i = \begin{cases} 0 & \text{if subject } i \text{ was not dead at 90 days after hospitalization} \\ 1 & \text{if subject } i \text{ was dead at 90 days after hospitalization} \end{cases}$$

$$T_{ij} = \frac{1}{\text{Day after hospitalization}_{ij}}$$

The mean vectors and covariance matrices for the survival and death groups are calculated from the parameter estimates of the linear mixed model. The best linear combination coefficient for each time point is (0.016, 0.038, 0.045, 0.048, 0.050, 0.052, 0.053) showing that the most recent measurement has the highest weight. The optimum AUC and 95% confidence interval for IL-6 are 0.718 (0.683, 0.750). Next, the lower detection limit for D-dimer is 110

ng/mL and its censoring proportions are lower than IL-6 with 5.3%, 4.4%, 5.0%, 5.4%, 4.5%, 3.7%, and 4.0% over 7 days. Logarithm of D-dimer level for death group is a bit higher than that of the survival group (Figure 5). The estimated linear mixed model for log(D-dimer) is

$$Y_{ij} = 6.215 + 0.729X_i + 0.021T_{ij} + a_i + b_iT_{ij} + e_{ij}$$

$$e_{ij} \sim N(0, 0.113) \quad \text{and} \quad \begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.063 & -0.061 \\ -0.061 & 0.023 \end{pmatrix}\right]$$

$$X_i = \begin{cases} 0 & \text{if subject } i \text{ was not dead at 90 days after hospitalization} \\ 1 & \text{if subject } i \text{ was dead at 90 days after hospitalization} \end{cases}$$

$$T_{ij} = \text{Day after hospitalization}_{ij}$$

The linear combination coefficient for each time point is (0.144, 0.070, 0.045, 0.033, 0.026, 0.021, 0.017). Contrary to the IL-6, the weights of the measurements for D-dimer decrease over time. The optimum AUC for D-dimer is 0.695 (0.641, 0.741), which is smaller than the AUC for IL-6.

4.5 DISCUSSION

High dimensionality and censoring issue are often a problem in the evaluation of biomarkers. Motivated by the GenIMS study, we propose the method to find the best linear combination coefficient for longitudinal biomarker measurements. By combining data from multiple time points, we can reduce the time dimension of longitudinal data into one and capture different evolution patterns. The combination coefficient enables us to evaluate biomarker's maximum classification power via AUC. Investigators can also predict the relative importance of each time point to the discrimination ability. The method is applicable for both censored and non-censored measurements. It is noted that the AUC calculated from the linear discriminant

analysis is mathematically equal to the optimum AUC from the best linear combination coefficient. One of the limitations of our method is that it only considers the group level parameters such as mean and variance for event and non-event groups, it's impossible to know individual's classification result based on each subject's biomarker records. However, the method is practically useful in terms of short computational time to provide a summary statistic for biomarker's discrimination ability, especially when the biomarker is repeatedly measured over several time points and censored due to LOD.

5.0 DISCRIMINATION MEASURE OF CENSORED BIOMARKER FOR SURVIVAL OUTCOME

Censoring due to a detection limit is an increasingly common and challenging problem in biomarker studies. When a biomarker is used to predict survival outcome, one of the important problems due to censored biomarker data is corrupted evaluation measures. Biomarker's discrimination potential for survival outcome is frequently evaluated by C-index and time dependent ROC curve. In this chapter, we extend these two methods to left-censored baseline biomarker data. To incorporate the censored biomarker measurements, we use the joint likelihood based method. We derive the analytic form of C-index and time dependent AUC, which is a function of parameters in the joint likelihood function. Simulation study shows that the proposed method outperforms over simple substitution methods in terms of parameter estimation, resulting in less biased evaluation measures. We provide the application with the dataset from biomarker study for acute kidney injury patients.

5.1 INTRODUCTION

The ROC curve has been widely used in the evaluation of diagnostic accuracy for dichotomous outcomes. In a prospective study when the disease onset is observed over a continuous follow-up time, the essential outcome of interest is not only the occurrence of disease (binary) but also the time to event (continuous). In the context of survival outcome, it is more appropriate to consider event time in the calculation of ROC curve. A common example of the extended version of ROC curve to survival outcome is C-index, so called a global accuracy summary measure. The C-index measures how well a biomarker correctly ranks

patients by their survival time throughout the whole study period. When a priori time t is specified, however, time-dependent ROC curve is more desired. The time-dependent ROC curve summarizes the discriminant potential of a biomarker for cumulative events occurred by time t. In this way, the ROC curve is expressed as a function of time. In the Biological Markers of Recovery for the Kidney (BioMaRK) study, one of the scientific hypotheses is that urinary biomarkers can predict the recovery outcome of renal function for critically ill patients with acute kidney injury (AKI). It derives the consequent questions that (1) which biomarker has the best discrimination ability for the time to recovery overall and (2) how well the biomarker can distinguish patients who will recover with those who will not by the follow-up time of t. The first question can be answered by the C-index because scientific interest is more on the biomarker's global discrimination ability for the whole study period. Cumulative/dynamic ROC curve can be applied for the analysis of second question, which focuses more on case or control group on the basis of their vital status at t.

The C-index and time dependent ROC curve have been successful evaluation methods for survival outcome. The C-index and its property were studied by Harrell et al. [32], Pencina and D'Agostino [34], and investigated further to obtain a stable estimation in the presence of tied pairs or censoring in the survival outcome [35] [36]. Later, Gönen and Heller [76] derived analytical expression of the concordance probability in the Cox proportional hazard model. The standard ROC curve has been extended to time to event outcome by Heagerty et al. [39] in the form of cumulative/dynamic time dependent ROC. They introduced a new definition of sensitivity, specificity, and ROC curve in a time dependent manner. Chambless and Diao [77] developed two different estimation method for time dependent AUC as a summary measure of time dependent ROC. However, all of these methods have a limitation to be directly applied to the BioMaRK study because they were developed based on fully detectable data. In our application example, some of urinary biomarker measurements are not observed due to detection of limit. To incorporate censored biomarker data as a covariate in the analysis of survival outcome, D'Angelo et al. [78] presented an index approach in the Cox proportional hazard model. When estimating model, they replaced the censored observations with conditional expectation given fully observed covariates. In this chapter, we will propose a modification of the C-index and time dependent ROC for censored biomarker data. We

start from the joint likelihood based approach to analyze survival outcome in the presence of censored covariate. We briefly review the existing estimation methods for the C-index and time dependent ROC and explain our remedy for censored biomarker measurements. In section 5.3, we examine the performance of the proposed method using simulation study. In section 5.4, the proposed methods are applied to the BioMaRK study in order to evaluate the inflammatory marker interleukin-6 (IL-6) and IL-18 in the prediction of time to recovery outcome.

5.2 METHOD

5.2.1 Survival model with censored covariates

Survival data are commonly fitted by a density function of parametric distribution or by a semiparametric Cox proportional hazard model. While many researches have been done for censored outcome, a model for both censored outcome and censored covariate has not been extensively investigated yet. In this section, we consider the joint likelihood based approach to handle the censored covariates in the survival model.

Parametric distribution of survival time

First, we assume the parametric form for the distribution of survival time. One of frequently used distributions for survival time is exponential distribution, which has the constant hazard during the whole study period. Denote by Z_i the survival time for the i^{th} subject with baseline biomarker value Y_i $(i = 1 \cdots N)$. The relationship between Y_i and Z_i is determined through the exponential distribution function, $h(z_i|y_i) = \lambda_i exp(-\lambda_i z_i)$, where $\lambda_i = exp(\beta y_i)$. For each subject, what we actually observe for survival time is $T_i = \min(Z_i, C_i)$, where C_i represents a censoring time. Let δ_i be the censoring indicator, $\delta_i = 1$ if the subject develops an event within a study period, i.e. $T_i = Z_i$, and $\delta_i = 0$ for the subject who is either loss of follow up or event free at the end of a study, i.e. $T_i = C_i$. The baseline biomarker Y_i has a density function $f_\theta(\cdot)$, which is often assumed to follow (log)normal distribution. We specify

the censoring indicator for a biomarker with lower detection limit at τ as follows.

$$\omega_i = \begin{cases} 1 & \text{if} \quad Y_i \leq \tau \\ 0 & \text{if} \quad Y_i > \tau \end{cases}$$

To incorporate left censored biomarker data in the estimation of a set of parameters $\{\theta, \beta\}$, we use the join likelihood based method, similar to Lyles et al.[19]. The likelihood function for the observed data (t, δ, y)=$(t_i, \delta_i, y_i, i = 1 \cdots N)$ is given by

$$L(\theta, \beta; t, \delta, y) = \prod_{i=1}^{N} \left[\{ f_\theta(y_i) h(t_i|y_i)^{\delta_i} \{1 - H(t_i|y_i)\}^{1-\delta_i} \}^{1-\omega_i} \right. \tag{5.1}$$

$$\left. \{ \int_{-\infty}^{\tau} f_\theta(y_i) h(t_i|y_i)^{\delta_i} \{1 - H(t_i|y_i)\}^{1-\delta_i} dy_i \}^{\omega_i} \right], \tag{5.2}$$

where $H(\cdot)$ denotes the cumulative distribution function for survival time.

Cox proportional hazard model of survival time

The parametric distribution for survival outcome is not always useful, especially in case of survival time after a major surgery. Semiparametric methods are more appropriate in such a situation. The Cox proportional hazard model has been the most widely used procedure in biomedical survival analysis. We can generalize our survival model to Cox proportional hazard model. The Cox proportional hazard model is defined as $\lambda_i(z) = \lambda_0(z)exp(\beta y_i)$, where $\lambda_i(\cdot)$ is the hazard function given y_i, $\lambda_0(\cdot)$ is the baseline hazard function and β is regression parameter. With left-censored covariates, the likelihood function is modified as

$$L(\beta, \theta; t, \delta, y) = \prod_{i=1}^{N} \left[\{ f_\theta(y_i) \ (\lambda_0(t_i)exp(\beta y_i))^{\delta_i} \ exp\left(- \int_0^{t_i} \lambda_0(u)exp(\beta y_i)du \right) \}^{1-\omega_i} \right.$$

$$\left. \times \{ \int_{-\infty}^{\tau} f_\theta(y_i) \ (\lambda_0(t_i)exp(\beta y_i))^{\delta_i} \ exp\left(- \int_0^{t_i} \lambda_0(u)exp(\beta y_i)du \right) dy_i \}^{\omega_i} \right]. \tag{5.3}$$

Rather than specifying a certain distribution for the baseline hazard, we adopt the piecewise constant baseline hazard function. We divide a follow-up period until last event into 20 intervals, which are found out to give stable parameter estimates in our simulation study. The piecewise constant baseline is simple but powerful function in terms of flexibility and practical applicability. Another advantage of using it in the estimation process is that we can avoid computational difficulties by employing Gaussian quadrature techniques. It is

pointed out that Gaussian quadrature technique for the Cox proportional hazard model assuming piecewise constant baseline hazards yields satisfactory parameter estimates [79]. We implement the Gaussian quadrature technique using SAS `Proc nlmixed` procedure.

5.2.2 C-index

The C-index measures biomarker's discrimination ability for survival outcome over the whole study period. Let's denote the actual survival time for i^{th} subject as Z_i, predicted survival time as U_i, predicted probability of survival at time t as W_i, and time-invariant biomarker measurement as Y_i $(i = 1 \cdots N)$. Harrell et al. [33] defined the C-index as $Pr(U_i < U_j | Z_i < Z_j)$, the probability that the person with a shorter event time has a shorter predicted survival time assuming two persons are randomly selected from a cohort. It is pointed out that U_i and W_i are exchangeable if they have one-to-one correspondence. Furthermore, under the proportional hazards assumption, the C-index can be rewritten as $Pr(W_i < W_j | Z_i < Z_j) = Pr(\beta Y_i > \beta Y_j | Z_i < Z_j)$ [34].

There are two ways to estimate the C-index. Without a distributional assumption on biomarker data, the C-index (C_n) is estimated by

$$\hat{C}_n = \frac{1}{Q} \sum_{(i,j) \in R} c_{ij},$$

where R is a set of all usable pairs, Q is the total number of usable pairs in R, and

$$c_{ij} = \begin{cases} 1 & \text{if } (Z_i < Z_j \text{ and } \beta Y_i > \beta Y_j) \text{ or } (Z_i > Z_j \text{ and } \beta Y_i < \beta Y_j) \\ 0 & \text{if } (Z_i < Z_j \text{ and } \beta Y_i < \beta Y_j) \text{ or } (Z_i > Z_j \text{ and } \beta Y_i > \beta Y_j). \end{cases}$$

Alternately, distributional assumptions on biomarker data enable us to calculate the C-index. Suppose $g(z, y)$ denote the joint density function of actual survival time and biomarker. For illustration, we assume that the subject with higher risk of event has larger biomarker measurements, i.e. $\beta > 0$. Then the C-index can be calculated as

$$C_p = Pr(\beta Y_i > \beta Y_j | Z_i < Z_j) = \frac{Pr(Y_i > Y_j, Z_i < Z_j)}{Pr(Z_i < Z_j)}$$

$$= \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{y_i} \int_{0}^{\infty} \int_{0}^{z_j} g(z_i, y_i) g(z_j, y_j) \, dz_i dz_j dy_j dy_i}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{0}^{\infty} \int_{0}^{z_j} g(z_i, y_i) g(z_j, y_j) \, dz_i dz_j dy_j dy_i}.$$

Because we derive the C-index, C_p, directly from the joint likelihood function, it only depends on the parameters in the distribution function. Therefore, C_p can be estimated correctly even in the presence of censored covariates and censored outcomes if we can obtain the unbiased parameter estimates.

Estimation under parametric distribution

Going back to the parametric distribution for survival data, we additionally assume that biomarker Y_i is independently, identically and normally distributed with mean μ and variance σ^2. We can rewrite the joint density function $g(z, y)$ as conditional density of survival time, $h(z|y)$, multiplied by the marginal density of biomarker, $f(y)$. Then C_p is expressed numerically by

$$C_p = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{y_i}\int_{0}^{\infty}\int_{0}^{z_j} \; h(z_i|y_i)h(z_j|y_j) \; f(y_i)f(y_j) \; dz_i dz_j dy_j dy_i}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{z_j} \; h(z_i|y_i)h(z_j|y_j)f(y_i)f(y_j) \; dz_i dz_j dy_j dy_i}$$

$$= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{y_i}\int_{0}^{\infty}\int_{0}^{z_j} \lambda_i exp(-\lambda_i z_i) \; \lambda_j exp(-\lambda_j z_j) \; exp\left(\frac{-(y_i-\mu)^2}{2\sigma^2}\right) \; exp\left(\frac{-(y_j-\mu)^2}{2\sigma^2}\right) \; dz_i dz_j dy_j dy_i}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{z_j} \lambda_i exp(-\lambda_i t_i) \; \lambda_j exp(-\lambda_j t_j) \; exp\left(\frac{-(y_i-\mu)^2}{2\sigma^2}\right) \; exp\left(\frac{-(y_j-\mu)^2}{2\sigma^2}\right) \; dz_i dz_j dy_j dy_i},$$

where μ, σ^2 and β are estimated by maximizing the joint likelihood function (5.1).

Estimation under Cox proportional hazard model

In the framework of Cox proportional hazard model and $Y_i \sim \mathrm{N}(\mu, \sigma^2)$, the C-index is formulated even more simpler. Rather than starting directly from the joint distribution, we express C_p in terms of $Pr(Z_i < Z_j|Y_i = y_i, Y_j = y_j)$ as follows :

$$C_p = \frac{Pr(Y_i > Y_j, Z_i < Z_j)}{Pr(Z_i < Z_j)} = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{y_i} Pr(Z_i < Z_j|y_i, y_j)f(y_i)f(y_j) \; dy_j dy_i}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Pr(Z_i < Z_j|y_i, y_j)f(y_i)f(y_j) \; dy_j dy_i}.$$

Because

$$Pr(Z_i < Z_j|Y_i = y_i, Y_j = y_j) = \int_0^{\infty} S(z|y_i)dS(z|y_j) = \frac{1}{1 + exp(\beta y_j - \beta y_i)},$$

where $S(.)$ denotes a survival function, C_p can be calculated using only regression parameter β, distributional parameter μ and σ^2 estimated from the joint likelihood function (5.3) [76]. The analytic form of C_p is given by

$$C_p = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{y_i} exp\left(\frac{-(y_i-\mu)^2}{2\sigma^2}\right)\ exp\left(\frac{-(y_j-\mu)^2}{2\sigma^2}\right)\ /\,(1+exp(\beta y_j-\beta y_i))\,dy_j dy_i}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} exp\left(\frac{-(y_i-\mu)^2}{2\sigma^2}\right)\ exp\left(\frac{-(y_j-\mu)^2}{2\sigma^2}\right)\ /\,(1+exp(\beta y_j-\beta y_i))\,dy_j dy_i}.$$

5.2.3 Cumulative/dynamic time dependent ROC

If the biomarker's discriminant ability throughout the time interval (0, t] is of clinical interest, cumulative/dynamic time dependent ROC can be used as a standard summary of accuracy. If Y denotes the baseline biomarker measurement with higher value being more indicative of event and Z is actual survival time, sensitivity and specificity are specified as $Sensitivity(t) = Pr(Y > c|Z \leq t)$ and $Specificity(t) = Pr(Y \leq c|Z > t)$. In other words, sensitivity is evaluated using events which occur in the time interval (0, t], whereas specificity is calculated based on events after the time t. The AUC at time t is defined as $AUC(t) = Pr(Y_i > Y_j|Z_i \leq t, Z_j > t)$, the probability that subjects with an event by time t has a higher biomarker level than those without an event.

Heagerty et al. [39] showed nonparametric estimation method for $AUC(t)$ employing the Kaplan-Meier estimator for survival function and empirical distribution function for biomarker measurements. The estimator for sensitivity and specificity at time t is given by

$$Sensitivity(t) = \frac{\{1-\hat{S}(t|Y>c)\}\{1-\hat{K}(c)\}}{1-\hat{S}(t)}$$
$$Specificity(t) = \frac{\hat{S}(t|Y\leq c)\hat{K}(c)}{\hat{S}(t)},$$

where $\hat{K}(c) = \sum_{i=1}^{N} I(Y_i \leq c)/N$ and $\hat{S}(t|Y \geq c)$ is the Kaplan-Meier estimator based on the subset of subjects with $\{Y \geq c\}$. Using the $sensitivity(t)$ and $specificity(t)$ over all possible values of c, AUC can be estimated by the trapezoidal rule. As an alternative estimation approach, Chambless and Diao [77] derived the direct formula for $AUC(t)$ in terms of a survival function and density function for biomarker. Denote the conditional

density function of biomarker as $q(y|z)$ and marginal density function of biomarker as $f(y)$. Then,

$$AUC(t) = Pr(Y_i > Y_j | Z_i \leq t, Z_j > t) = \int_{-\infty}^{\infty} \int_{y_j}^{\infty} q(y_i | Z_i \leq t) q(y_j | Z_j > t) \; dy_i dy_j.$$

Because

$$q(y_i | Z_i \leq t) = \frac{Pr(Z_i \leq t | Y_i = y_i) f(y_i)}{Pr(Z_i \leq t)} = \frac{(1 - S(t|y_i)) f(y_i)}{\int_{-\infty}^{\infty} (1 - S(t|y_i)) f(y_i) dy_i} = \frac{(1 - S(t|y_i)) f(y_i)}{E\left[(1 - S(t|Y_i))\right]},$$

where E indicates the expectation with respect to Y_i. Therefore, the $AUC(t)$ can be written as

$$\int_{-\infty}^{\infty} \int_{y_j}^{\infty} \frac{(1 - S(t|y_i)) \, f(y_i) S(t|y_j) f(y_j) dy_i dy_j}{E\left[1 - S(t|Y_i)\right] E\left[S(t|Y_j)\right]} = \frac{E\left[(1 - S(t|Y_i)) S(t|Y_j) I(Y_i > Y_j)\right]}{E\left[1 - S(t|Y_i)\right] E\left[S(t|Y_j)\right]}.$$

Estimation under parametric distribution

We first assume the parametric distributions for both survival time and biomarker data censored by a lower detection limit as follows.

$$\begin{cases} Y_i = \mu + e_i \quad , \quad e_i \sim N(0, \sigma^2) \\ Z_i | Y_i \sim \text{EXP}(\lambda_i) \quad \text{,where} \quad \lambda_i = exp(\beta Y_i). \end{cases}$$

Under this assumption, the $AUC(t)$ is given by

$$AUC(t) = \frac{\int_{-\infty}^{\infty} \int_{y_j}^{\infty} (1 - exp(-\lambda_i t)) \; exp(-\lambda_j t) \; exp\left(\frac{-(y_i - \mu)^2}{2\sigma^2}\right) \; exp\left(\frac{-(y_j - \mu)^2}{2\sigma^2}\right) \; dy_i dy_j}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (1 - exp(-\lambda_i t)) \; exp(-\lambda_j t) \; exp\left(\frac{-(y_i - \mu)^2}{2\sigma^2}\right) \; exp\left(\frac{-(y_j - \mu)^2}{2\sigma^2}\right) \; dy_i dy_j},$$

where $\hat{\mu}$, $\hat{\sigma^2}$, $\hat{\lambda_i}$ and $\hat{\lambda_j}$ are maximum likelihood estimates obtained from (5.1).

Estimation under Cox proportional hazard model

Everything being same as before, we use the Cox proportional hazard model $\lambda_i(z) = \lambda_0(z) exp(\beta y_i)$ for survival outcome. In this case, the survival function at time z is specified as $S(z|y_i) = S_0(z)^{exp(\beta y_i)}$, where baseline survival function $S_0(z)$ is given by $exp(-\int_0^z \lambda_0(u) du)$. The C-index depends only on the regression coefficient β in the Cox proportional hazard

model, so the estimation of baseline hazard is unnecessary. However, we cannot estimate $AUC(t)$ without obtaining baseline hazard function. In the semiparametric model, the baseline hazard is usually unspecified and it can have any form. By assuming the piecewise constant hazard, however, we can circumvent this problem. Suppose we divide the follow-up period into 20 intervals and denote I_k (k=1,$\cdots$,20) as every 5^{th} quantile. The λ_{0k} is the piecewise constant baseline hazard in each interval, i.e. $\lambda_0(t) = \lambda_{0k}$ for $[I_{k-1} < t \leq I_k]$. Then, the cumulative baseline hazard is [79]

$$\int_0^t \lambda_0(u)du = \sum_{k=1}^{20} \lambda_{0k} \text{ max } (0,\ \text{min } (I_k - I_{k-1}, t - I_{k-1})).$$

Now the $AUC(t)$ can be obtained by using the parameter estimates from the likelihood function (5.3) and the estimate of cumulative baseline hazard,

$$AUC(t) = \frac{\int_{-\infty}^{\infty}\int_{y_j}^{\infty} \left(1 - S_0(t)^{exp(\beta y_i)}\right)\ S_0(t)^{exp(\beta y_j)}\ exp\left(\frac{-(y_i-\mu)^2}{2\sigma^2}\right)\ exp\left(\frac{-(y_j-\mu)^2}{2\sigma^2}\right)\ dy_i dy_j}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(1 - S_0(t)^{exp(\beta y_i)}\right)\ S_0(t)^{exp(\beta y_j)}\ exp\left(\frac{-(y_i-\mu)^2}{2\sigma^2}\right)\ exp\left(\frac{-(y_j-\mu)^2}{2\sigma^2}\right)\ dy_i dy_j}.$$

5.3 SIMULATION

In the first simulation, we present the point estimates of C-index calculated from the proposed method as well as those from the substitution methods, in which censored biomarker measurements are replaced by LOD and LOD/2. True value of the C-index is obtained from a numerical integration using the known parameter values. For data generation, time-invariant biomarker measurements are simulated from $Y \sim N(\mu, \sigma^2)$. The actual survival time Z is assumed to follow the exponential distribution $Z|Y \sim EXP(\lambda)$, where $\lambda = exp(\beta_0 + \beta_1 y)$. We simulate censoring time C from the uniform distribution $U(0,\ 4)$ and take the observed time as T=min(Z, C). The C-index is expected to be larger when biomarker data are widely distributed so that biomarker level from a subject with shorter survival time shows big difference from that from a subject who lives longer. Correspondingly, the variability of Y, σ^2, is varied at the level of 0.1, 1.0, and 5.0. Other parameters are fixed at μ=0.0, β_0=0.0 and β_1=1.0. Under this setup, a subject with higher biomarker value has higher risk of event.

The lower detection limit for biomarker Y is chosen empirically so that the censoring rate is set to 0.2 and 0.4. Total 500 observations are generated for each dataset. For evaluation, estimated C-indices and parameter estimates are averaged over 100 datasets. Later, we generalize the distributional assumption for survival time and use the Cox proportional hazard model. Fixing other simulation setup the same as before, we regenerate the actual survival time Z based on the hazard function $\lambda(z) = \lambda_0(z)exp(\beta y)$, where $\lambda_0(z) = 2z$ and β=1.

Tables 8 and 9 summarize the simulation results under the assumption of exponential distribution and Cox proportional hazard model, respectively. The parameter estimates from the proposed method are close to the true values, while larger biases are shown in the substitution methods as censoring proportion becomes higher. The C-index from substitution methods tends to be biased downward. Our method produces C-index that is much closer to the true value, which is competently similar to the omniscient estimate obtained from complete data.

The second part of the simulation study is designed to compare the performance of our method in estimating $AUC(t)$ to that of the substitution methods. In this simulation, baseline biomarker measurements are generated from $Y \sim N(\mu, \sigma^2)$ with $\mu = 0$ and $\sigma^2 = 4$ and 9. We simulate actual survival time Z from exponential distribution, $EXP[exp(\beta_0 + \beta_1 y)]$ with $\beta_0 = 0$ and $\beta_1 = 0.1$. The censoring time C follows uniform distribution $U(0, 4)$. For the Cox proportional hazard model, we generate the actual survival time Z with exponential survival for $S_0(Z)$ and $\beta_1 = 0.1$. Tables 10 and 11 present the results averaged over 100 datasets. The simulation results show that the proposed method produces comparable $AUC(t)$ estimates to omniscient and true values. However, the biases of both parameter and $AUC(t)$ estimates from the substitution methods increase with the censoring proportion.

5.4 APPLICATION TO BIOMARK STUDY

Acute kidney injury (AKI) is the most common problem in the intensive care unit, which affects negatively on patient's quality of life and causes subsequent health care cost [80]. Conducted as a part of the Acute Renal Failure Trial Network study (ATN study), the

Table 8: Comparison of C-index estimated from the proposed method (PM) and substitution methods (LOD, LOD/2) assuming exponential distribution for survival time ($\mu = 0$, $\beta_0 = 0$, $\beta_1 = 1$)

σ^2	Pr(censor)	Method	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	C	$\hat{C}(SD)$
0.1	0.2	Omni	-0.001	0.100	0.004	1.017	0.586	0.588(0.015)
		LOD	0.034	0.069	-0.037	1.167		0.584(0.014)
		LOD/2	0.061	0.056	-0.073	1.236		0.580(0.013)
		PM	0.001	0.098	0.002	1.032		0.588(0.014)
	0.4	Omni	-0.001	0.100	0.004	1.017	0.586	0.588(0.015)
		LOD	0.089	0.044	-0.118	1.349		0.578(0.014)
		LOD/2	0.105	0.040	-0.146	1.399		0.577(0.013)
		PM	0.004	0.097	0.000	1.039		0.588(0.014)
1.0	0.2	Omni	-0.004	1.004	0.007	1.004	0.725	0.726(0.010)
		LOD	0.108	0.695	-0.090	1.110		0.713(0.009)
		LOD/2	0.193	0.561	-0.200	1.184		0.707(0.009)
		PM	-0.001	0.994	0.006	1.008		0.725(0.010)
	0.4	Omni	-0.004	1.004	0.007	1.004	0.725	0.726(0.010)
		LOD	0.281	0.448	-0.329	1.282		0.702(0.010)
		LOD/2	0.333	0.396	-0.415	1.333		0.699(0.010)
		PM	0.001	0.993	0.004	1.009		0.725(0.010)
5.0	0.2	Omni	-0.009	5.020	0.003	1.003	0.801	0.800(0.004)
		LOD	0.243	3.481	-0.096	1.050		0.801(0.003)
		LOD/2	0.432	2.809	-0.277	1.112		0.799(0.003)
		PM	-0.005	4.990	0.002	1.004		0.800(0.004)
	0.4	Omni	-0.009	5.020	0.003	1.003	0.801	0.800(0.004)
		LOD	0.628	2.239	-0.510	1.188		0.795(0.004)
		LOD/2	0.745	1.977	-0.684	1.235		0.793(0.004)
		PM	-0.007	5.006	0.002	1.004		0.800(0.004)

Table 9: Comparison of C-index estimated from the proposed method (PM) and substitution methods (LOD, LOD/2) assuming Cox proportional hazard model for survival time ($\mu = 0$, $\beta = 1$)

σ^2	Pr(censor)	Method	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\beta}$	C	$\hat{C}(SD)$
0.1	0.2	Omni	-0.001	0.100	0.993	0.586	0.586(0.013)
		LOD	0.034	0.069	1.143		0.582(0.013)
		LOD/2	0.061	0.056	1.208		0.578(0.013)
		PM	0.001	0.099	1.003		0.586(0.014)
	0.4	Omni	-0.001	0.100	0.993	0.586	0.586(0.013)
		LOD	0.089	0.045	1.326		0.577(0.013)
		LOD/2	0.105	0.040	1.374		0.575(0.012)
		PM	0.004	0.097	1.016		0.586(0.014)
1.0	0.2	Omni	-0.004	1.004	1.007	0.725	0.726(0.011)
		LOD	0.108	0.695	1.118		0.714(0.011)
		LOD/2	0.193	0.561	1.165		0.704(0.011)
		PM	-0.001	0.994	1.010		0.726(0.012)
	0.4	Omni	-0.004	1.004	1.007	0.725	0.726(0.011)
		LOD	0.281	0.448	1.263		0.700(0.011)
		LOD/2	0.333	0.396	1.300		0.695(0.011)
		PM	0.001	0.992	1.012		0.726(0.012)
5.0	0.2	Omni	-0.009	5.020	1.008	0.843	0.844(0.008)
		LOD	0.243	3.475	1.081		0.829(0.008)
		LOD/2	0.432	2.809	1.111		0.818(0.008)
		PM	-0.005	4.989	1.004		0.843(0.008)
	0.4	Omni	-0.009	5.020	1.003	0.843	0.844(0.008)
		LOD	0.628	2.239	1.188		0.812(0.009)
		LOD/2	0.745	1.977	1.235		0.807(0.009)
		PM	-0.007	5.006	1.004		0.843(0.008)

Table 10: Comparison of $AUC(t)$ estimated from the proposed method (PM) and substitution methods (LOD, LOD/2) assuming exponential distribution for survival time

Pr(censor)	Method	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\widehat{AUC}(3)(SD)$	$\widehat{AUC}(5)(SD)$
True		0.000	4.000	0.000	0.100	0.663	0.737
0.2	Omni	-0.005	4.001	0.018	0.102	0.665(0.037)	0.738(0.044)
	LOD	0.218	2.770	-0.010	0.119	0.661(0.037)	0.732(0.045)
	LOD/2	0.386	2.239	-0.033	0.127	0.656(0.038)	0.726(0.048)
	PM	-0.002	3.971	0.016	0.103	0.665(0.037)	0.738(0.045)
0.4	Omni	-0.005	4.001	0.018	0.102	0.665(0.037)	0.738(0.044)
	LOD	0.564	1.779	-0.063	0.140	0.653(0.040)	0.723(0.050)
	LOD/2	0.667	1.573	-0.082	0.146	0.650(0.040)	0.719(0.050)
	PM	0.003	3.955	0.016	0.104	0.667(0.039)	0.740(0.046)
True		0.000	9.000	0.000	0.100	0.727	0.803
0.2	Omni	-0.009	9.003	0.015	0.102	0.726(0.029)	0.805(0.029)
	LOD	0.327	6.233	-0.024	0.117	0.719(0.030)	0.798(0.031)
	LOD/2	0.582	5.041	-0.058	0.125	0.713(0.032)	0.791(0.034)
	PM	-0.004	8.943	0.015	0.102	0.726(0.030)	0.804(0.030)
0.4	Omni	-0.009	9.003	0.015	0.102	0.726(0.029)	0.805(0.029)
	LOD	0.848	4.010	-0.104	0.137	0.709(0.033)	0.787(0.034)
	LOD/2	0.999	3.539	-0.129	0.143	0.705(0.033)	0.784(0.036)
	PM	0.001	8.917	0.015	0.103	0.727(0.032)	0.805(0.032)

Table 11: Comparison of $AUC(t)$ estimated from the proposed method (PM) and substitution methods (LOD, LOD/2) assuming Cox proportional hazard model for survival time

Pr(censor)	Method	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\beta}$	$\widehat{AUC}(1)(SD)$	$\widehat{AUC}(2)(SD)$
True		0.000	4.000	0.100	0.588	0.624
0.2	Omni	-0.008	4.017	0.098	0.586(0.023)	0.620(0.031)
	LOD	0.217	2.780	0.113	0.583(0.023)	0.617(0.031)
	LOD/2	0.386	2.247	0.119	0.580(0.023)	0.613(0.031)
	PM	-0.004	3.988	0.099	0.587(0.024)	0.623(0.034)
0.4	Omni	-0.008	4.017	0.098	0.586(0.023)	0.620(0.031)
	LOD	0.562	1.791	0.132	0.580(0.023)	0.614(0.023)
	LOD/2	0.666	1.581	0.136	0.578(0.022)	0.612(0.031)
	PM	-0.003	3.987	0.100	0.587(0.025)	0.623(0.034)
True		0.000	9.000	0.100	0.629	0.677
0.2	Omni	-0.012	9.037	0.099	0.627(0.023)	0.672(0.028)
	LOD	0.325	6.254	0.113	0.623(0.023)	0.668(0.028)
	LOD/2	0.579	5.055	0.119	0.618(0.023)	0.664(0.029)
	PM	-0.007	8.983	0.100	0.628(0.023)	0.675(0.030)
0.4	Omni	-0.012	9.037	0.099	0.627(0.023)	0.672(0.028)
	LOD	0.842	4.030	0.131	0.618(0.023)	0.665(0.030)
	LOD/2	0.999	3.558	0.134	0.615(0.025)	0.662(0.034)
	PM	-0.007	8.987	0.100	0.628(0.024)	0.675(0.031)

Biological Markers of Recovery for the Kidney (BioMARK) study is designed to investigate the role of plasma and urinary biomarkers in prediction of renal outcomes. The objective of our analysis is to measure the biomarker's predictive accuracy for time to recover of renal function. Furthermore, we investigate the biomarker's discrimination power for cumulative recovery events by time t. The baseline urine biomarkers were collected from 76 participants in the intensive monitoring cohort. We select IL-6 and IL-18 as biomarkers of interest for the illustration purpose of our method. The censoring proportions for the baseline IL-6 and IL-18 are 14.9% and 15.5%, respectively. Recovery of renal function is defined by survival and dialysis independence. For this analysis, all deaths are treated as censored at the end of the follow-up, 60 days. Total 53% subjects have censored survival outcome.

We assume the log-normal distribution for baseline biomarker measurements and use the Cox proportional hazard model for time to recover data. The estimated value of the C-index for IL-6 is 0.579. The LOD and LOD/2 substitutions produce the C-indices of 0.570 and 0.575. As expected, the substitution methods report lower discrimination power for time to recovery. For IL-18, we calculate $AUC(t)$ at t=20 and t=40. The discrimination potential for IL-18 is pretty low; 0.508 at 20 days and 0.509 at 40 days. The $AUC(t)$ from substitution method (LOD/2) is 0.501 for both time points.

5.5 DISCUSSION

In this chapter, we extend the C-index and time dependent ROC methods to time-invariant censored biomarker data. In the estimation procedure, both methods require to fit a survival model including biomarker measurements as a covariate. In order to reduce a bias caused by censored covariates in the survival model, we use joint likelihood approach. The simulation study shows that our approach provides improved estimates than the LOD or LOD/2 substitution methods for the considered scenarios. Better parameter estimation enables us to calculate the C-index and time dependent AUC correctly because those measures depend on the parameters in the joint likelihood function.

In the survival analysis, all the subjects in the cohort are followed in a given study period. At the end of the study, subject can have an event of interest, does not experience an event, or can be lost during the follow-up without an event. The proposed estimation method considers all of these cases in the model. Thus C-index and time-dependent AUC are not sensitively affected by censoring proportion of survival outcome.

6.0 CONCLUSION

The main objective of this dissertation is to develop discriminant models and corresponding evaluation methods for censored biomarker data when the outcome is either binary or time to event data. For the binary outcome, we propose a new discriminant analysis method. We use the likelihood based method to account for the censoring due to detection limit. The classification is based on the newly calculated risk scores that are derived using a linear mixed model. Through the simulation, we point out that the linear mixed model should be carefully specified by comparing fit statistics. The discrimination power is evaluated by AUC. Furthermore, we assess the biomarker's predictive capacity by predictiveness curve.

As an alternative classification method, joint modeling approach is desirable. The basic idea under the joint modeling is that the linear mixed model for censored longitudinal biomarker data and logistic regression model for binary outcome can be linked via shared random effect parameters. In this case, posterior probability of event takes a role of the risk score from our discriminant analysis. Patients can be classified into two groups based on the the posterior probability. One of the difficulties in this approach is the estimation of individual's random effects, which are not actually observed. Further research should be carried out to calculate individual's posterior probability which is a function of random effect parameters.

Regarding the evaluation of biomarker performance, we investigate the AUC calculation for biomarkers with large time dimensions. Rather than directly handling longitudinal data, some researchers want to use condensed measure which contains all the information in the longitudinal data. We derive the best linear combination of time points that maximizes the AUC for both censored and non-censored biomarkers. Our methodology is easy and straightforward to implement with standard statistical software and provides satisfactory evaluation

results. Although we introduce the optimum AUC for single longitudinal biomarker here, the extension to the multiple longitudinal biomarkers would be possible.

For the survival outcome, we develop two evaluation methods for time-invariant censored biomarker data. Previous studies have introduced the new definition of sensitivity and specificity so that a biomarker's performance can be evaluated in a time dependent manner. In case a *priori* time point is not specified for evaluation, a global summary measure of discrimination can be used. Based on these concepts, we develop the estimation methods for a time dependent AUC and C-index by using a joint likelihood function of time to event and censored biomarker data. The time to event data are fitted by either parametric model or semi-parametric Cox proportional hazard model.

All the proposed methods are based on the assumption of normal distribution. However, the biomarker data are often highly skewed and may not satisfy the normality assumption. Appropriate treatment such as Box-Cox transformation can be considered to make the distribution of data close to normal. We mainly use the likelihood-based method to handle the censored data. As an alternative, multiple imputation method is increasingly applied to the analysis of censored data. Once the data set is completed by the imputed values, we can utilize the evaluation methods for biomarker which have been developed for fully observed data.

Our methodology can be widely applied to clinical decision-making when it is necessary to handle below or above the detection limit values. The proposed methods are useful to improve the quality of clinical decision making and facilitate health policy formation so that patients can get better treatment with less health care cost.

BIBLIOGRAPHY

[1] Su JQ, Liu JS. Linear Combination of Multiple Diagnostic Markers. Journal of the American Statistical Association. 1993;88:1350–1355.

[2] Network TVARFT. Intensity of Renal Support in Critically Ill Patients with Acute Kidney Injury. The New England Journal of Medicine. 2008;359:7–20.

[3] Authority EFS. Management of Left-Censored data in Dietary Exposure Assessment of Chemical Substances. EFSA Journal. 2010;8(3):1–96.

[4] Hughes JP. Mixed Effects Models with Censored Data with Application to HIV RNA Levels. Biometrics. 1999;55:625–629.

[5] Helsel DR. Insider Censoring: Distortion of Data with Nondetects. Human and Ecological Risk Assessment: An International Journal. 2005;11:1127–1137.

[6] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer; 2001.

[7] McIntosh MW, Pepe MS. Combining Several Screening Tests: Optimality of the Risk Score. Biometrics. 2002;58:657–664.

[8] Gao H, Davis JW. Why Direct LDA is Not Equivalent to LDA. Pattern Recognition. 2006;39:1002–1006.

[9] Hornung R, Reed L. Estimation of Average Concentration in the Presence of Nondetectable Values. Appl Occup Environ Hyg. 1990;5:46–51.

[10] Garland M, Morris JS, Rosner BA, Stampfer MJ, Spate VL, Baskett CJ, et al. Toenail Trace Element Levels as Biomarkers: Reproducibility over a 6-Year Period. Cancer Epidemiol Biomarkers Prev. 1993;2:493–497.

[11] Helsel DR. Less than Obvious - Statistical Treatment of Data Below the Detection Limit. Environmental Science & Technology. 1990;24:1766–1774.

[12] Lubin JH, Colt JS, Camann D, Davis S, Cerhan JR, Severson RK, et al. Epidemiologic Evaluation of Measurement Data in the Presence of Detection Limits. Environ Health Perspect. 2004;112:1691–1696.

[13] Croghan C, Egeghy PP. Methods of Dealing With Values Below the Limit of Detection Using SAS. Presented at Southern SAS User Group, St Petersburg, FL. 2003;.

[14] Baccarelli A, Pfeiffera R, Consonnib D, Pesatorib AC, Bonzinib M, Jr DGP, et al. Handling of Dioxin Measurement Data in the Presence of Non-detectable values: Overview of Available Methods and Their Application in the Seveso Chloracne Study . Chemosphere. 2005;60:898–906.

[15] Chen H, Quandt SA, Grzywacz JG, Arcury TA. A Distribution-Based Multiple Imputation Method for Handling Bivariate Pesticide Data with Values below the Limit of Detection. Environ Health Perspect. 2011;119:351–356.

[16] Jin Y, Hein MJ, Deddens JA, Hines CJ. Analysis of Lognormally Distributed Exposure Data with Repeated Measures and Values below the Limit of Detection Using SAS. The Annals of Occupational Hygiene. 2011;55:97–112.

[17] Tobin J. Estimation of Relationships for Limited Dependent Variables. Econometrica. 1958;26:24–36.

[18] Thiébaut R, Jacqmin-Gadda H. Mixed Models for Longitudinal Left-censored Repeated Measures. Computer Methods and Programs in Biomedicine. 2004;74:255–260.

[19] Lyles RH, Lyles CM, Taylor DJ. Random Regression Models for Human Immunodeficiency Virus Ribonucleic Acid data subject to Left censoring and Informative drop-outs. Applied Statistics. 2000;49:485–497.

[20] Hewett P, Ganser GH. A Comparison of Several Methods for Analyzing Censored Data. The Annals of Occupational Hygiene. 2007;51:611–632.

[21] Huang Y, Pepe MS, Feng Z. Evaluating the Predictiveness of a Continuous Marker. Biometrics. 2007;63:1181–1188.

[22] Pepe MS. Evaluating Technologies for Classification and Prediction in Medicine. Statistics in Medicine. 2005;24:3687–3696.

[23] Bamber D. The Area above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characteristic Graph. Journal of Mathematical Psychology. 1975;12:387–415.

[24] Hanley JA, McNeil BJ. The Meaning and Use of the Area under a Receive Operating Characterist(ROC) Curve. Radiology. 1982;143:29–36.

[25] 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.

[26] Hanley JA. The Use of The 'binormal' Model for Parametric ROC Analysis of Quantitative Diagnostic Tests. Statistics of Medicine. 1996;15:1575–1585.

[27] Perkins NJ, Schisterman EF, Vexler A. Receiver Operating Characteristic Curve Inference from a Sample with a Limit of Detection. American Journal of Epidemiology. 2007;165:325–333.

[28] Perkins NJ, Schisterman EF, Vexler A. Generalized ROC Curve Inference for a Biomarker subject to a Limit of Detection and Measurement Error. Statistics in Medicine. 2009;28:1841–1860.

[29] Vexler A, Liu A, Eliseeva E, Schisterman EF. Maximum Likelihood Ratio Tests for Comparing the Discriminatory Ability of Biomarkers Subject to Limit of Detection. Biometrics. 2008;64:895–903.

[30] Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, et al. Integrating the Predictiveness of a Marker with Its Performance as a Classifier. American Journal of Epidemiology. 2008;167:362–368.

[31] Gu W, Pepe M. Measures to Summarize and Compare the Predictive Capacity of Markers. The International Journal of Biostatistics. 2009;5:Article27.

[32] Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. JAMA. 1982;247:2543–2546.

[33] Harrell FE, Lee KL, Mark DB. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. Statistics in Medicine. 1996;15:361–387.

[34] Pencina MJ, D'Agostino RB. Overall C as a Measure of Discrimination in Survival Analysis: Model Specific Population Value and Confidence Interval Estimation. Statistics in Medicine. 2004;23:2109–2123.

[35] Yan G, Greene T. Investigating the Effects of Ties on Measures of Concordance. Statistics in Medicine. 2008;27:4190–4206.

[36] Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival data. Statistics in Medicine. 2011;30:1105–1117.

[37] DeLong ER, Vernon WB, Bollinger RR. Sensitivity and Specificity of a Monitoring Test. Biometrics. 1985;41:947–958.

[38] Parker CB, DeLong ER. ROC Methodology within a Monitoring Framework. Statistics in Medicine. 2003;22:3473–3488.

[39] Heagerty PJ, Lumley T, Pepe MS. Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. Biometrics. 2000;56:337–344.

[40] Zheng Y, Heagerty PJ. Prospective Accuracy for Longitudinal Markers. Biometrics. 2007;63:332–341.

[41] Saha P, Heagerty PJ. Time-Dependent Predictive Accuracy in the Presence of Competing Risks. Biometrics. 2010;66:999–1011.

[42] Heagerty PJ, Zheng Y. Survival Model Predictive Accuracy and ROC Curves. Biometrics. 2005;61:92–105.

[43] Etzioni R, Pepe M, Longton G, Hu C, Goodman G. Incorporating the Time Dimension in Receiver Operating Characteristic Curves: A Case Study of Prostate Cancer. Medical Decision Making. 1999;19:242–251.

[44] Slate EH, Turnbull BW. Statistical Models for Longitudinal Biomarkers of Disease Onset. Statistics in Medicine. 2000;19:617–637.

[45] Zheng Y, Heagerty PJ. Semiparametric Estimation of Time-Dependent ROC Curves for Longitudinal Marker Data. Biostatistics. 2004;5:615–632.

[46] Cai T, Pepe MS, Zheng Y, Lumley T, Jenny NS. The Sensitivity and Specificity of Markers for Event Times. Biostatistics. 2006;7:182–197.

[47] Pepe MS, Zheng Y, Jin Y, Huang Y, Parikh CR, Levy WC. Evaluating the ROC Performance of Markers for Future Events. Lifetime Data Analysis. 2008;14:86–113.

[48] Flier WVD, Scheltens P. Knowing the Natural Course of Biomarkers in AD: Longitudinal MRI, CSF and PET data. The Journal of Nutrition, Health and Aging. 2009;13:353–355.

[49] de Leon MJ, DeSanti S, Zinkowski R, Mehta PD, Pratico D, Segala S, et al. Longitudinal CSF and MRI Biomarkers Improve the Diagnosis of Mild Cognitive Impairment. Neurobiology of Aging. 2006;27:394–401.

[50] Bagshaw SM. Short- and Long-term Survival after Acute Kidney Injury. Nephrology Dialysis Transplantation. 2008;23:2126–2128.

[51] Tomasko L, Helms RW, Snapinn SM. A Discriminant Analysis Extension to Mixed Models. Statistics in Medicine. 1999;18:1249–1260.

[52] Marshall G, Barón AE. Linear Discriminant Models for Unbalanced Longitudinal data. Statistics in Medicine. 2000;19:1969–1981.

[53] Wernecke KD, Kalb G, Schink T, Wegner B. A Mixed Model Approach to Discriminant Analysis with Longitudinal Data. Biometrical Journal. 2004;46:246–254.

[54] Roy A, Khattree R. On Discrimination and Classification with Multivariate Repeated Measures data. Journal of Statistical Planning and Inference. 2005;134:462–485.

[55] Marshall G, la Cruz-Mesía RD, Quintana FA, Barón AE. Discriminant Analysis for Longitudinal Data with Multiple Continuous Responses and Possibly Missing Data. Biometrics. 2009;65:69–80.

[56] Kohlmann M, Held L, Grunert VP. Classification of Therapy Resistance Based on Longitudinal Biomarker Profiles. Biometrical Journal. 2009;51:610–626.

[57] Langdon MJ, Taylor CC, West RM. Classification of type I-censored bivariate data. Computational Statistics & Data Analysis. 2007;51:4562–4576.

[58] Cnaan A, Laird NM, Slasor P. Tutorial in Biostatistics Using the General Linear Mixed Model to Analyse Unbalanced Repeated Measures and Longitudinal Data. Statistics in Medicine. 1997;16:2349–2380.

[59] Littell R, Milliken G, Stroup W, Wolfinger R, Schabenberger O. SAS for Mixed Models. SAS Institute Inc.; 2006.

[60] McClish DK. Analyzing a Portion of the ROC Curve. Medical Decision Making. 1989;9:190–195.

[61] Palevsky PM, Zhang JH, OConnor TZ, Chertow GM, Crowley ST, Choudhury D, et al. Intensity of Renal Support in Critically Ill Patients with Acute Kidney Injury. The new england journal of medicine. 2008;359:7–20.

[62] Whitcomb BW, Schisterman EF. Assays with Lower Detection Limits: Implications for Epidemiological Investigations. Paediatric and Perinatal Epidemiology. 2008;22:597–602.

[63] Han C, Kronmal R. Box-Cox Transformation of Left-Censored Data with Application to the Analysis of Coronary Artery Calcification and Pharmacokinetic Data. Statistics in Medicine. 2004;23:3671–3679.

[64] Hamasakia T, Kim SY. Box and Cox Power-Transformation to Confined and Censored Nonnormal Responses in Regression. Computational Statistics & Data Analysis. 2007;51:3788–3799.

[65] He X, Ng KW, Shi J. Marginal versus Joint BoxCox Transformation with Applications to Percentile Curve Construction for IgG subclasses and Blood Pressures. Statistics in Medicine. 2003;22:397–408.

[66] Gurka MJ, Edwards LJ, Muller KE, Kupper LL. Extending the BoxCox Transformation to the Linear Mixed Model. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2006;169:273–288.

[67] Emir B, Wieand S, Su JQ, Cha S. Analysis of Repeated Markers used to Predict Progression of Cancer. Statistics in Medicine. 1998;17:2563–2578.

[68] Emir B, Wieand S, Jung SH, Ying Z. Comparison of Diagnostic Markers with Repeated Measurements: A Non-Parametric ROC Curve Approach. Statistics in Medicine. 2000;19:511–523.

[69] Lin H, McCulloch CE, Turnbull BW, Slate EH, Clark LC. A Latent Class Mixed Model for Analysing Biomarker Trajectories with Irregularly Scheduled Observations. Statistics in Medicine. 2000;19:1303–1318.

[70] Xiong C, Jr DWM, Miller JP, Morris JC. Combining Correlated Diagnostic Tests: Application to Neuropathologic Diagnosis of Alzheimer's Disease. Medical Decision Making. 2004;24:659–669.

[71] Liu A, Schisterman EF, Zhu Y. On Linear Combinations of Biomarkers to Improve Diagnostic Accuracy. Statistics in Medicine. 2005;24:37–47.

[72] Gao F, Xiong C, Yan Y, Yu K, Zhang Z. Estimating Optimum Linear Combination of Multiple Correlated Diagnostic Tests at a Fixed Specificity with Receiver Operating Characteristic Curves. Journal of Data Science. 2008;6:1–13.

[73] Pepe MS, Thompson ML. Combining Diagnostic Test Results to Increase Accuracy. Biostatistics. 2000;1:123–140.

[74] Huang DT, Weissfeld LA, Kellum JA, Yealy DM, Kong L, Martino M, et al. Risk Prediction With Procalcitonin and Clinical Rules in Community-Acquired Pneumonia. Annals of Emergency Medicine. 2008;52:48–58.

[75] Kellum JA, Kong L, Fink MP, Weissfeld LA, Yealy DM, Pinsky MR, et al. Understanding the Inflammatory Cytokine Response in Pneumonia and Sepsis. Archives of Internal Medicine. 2007;167:1655–1663.

[76] Gőnen M, Heller G. Concordance Probability and Discriminatory Power in Proportional Hazards Regression. Biometrika. 2005;92:965–970.

[77] Chambless LE, Diao G. Estimation of Time-dependent Area Under the ROC Curve for Long-term Risk Prediction. Statistics in Medicine. 2006;25:3474–3486.

[78] D'Angelo G, Weissfeld L, on behalf of the GenIMS Investigators. An Index Approach for the Cox model with Left Censored Covariates. Statistics in Medicine. 2008;27:4502–4514.

[79] Liu L, Huang X. The Use of Gaussian Quadrature for Estimation in Frailty Proportional Hazards Models. Statistics in Medicine. 2008;27:2665–2683.

[80] Manns B, Doig CJ, Lee H, Dean S, Tonelli M, Johnson D, et al. Cost of Acute Renal Failure requiring Dialysis in the Intensive Care Unit: Clinical and Resource Implications of Renal recovery. Critical Care Medicine. 2003;31:449–455.