JOINT MODELING OF LONGITUDINAL AND SURVIVAL OUTCOMES USING GENERALIZED ESTIMATING EQUATIONS

Mengjie Zheng

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the Department of Biostatistics,

Indiana University

July 2018

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Sujuan Gao, Ph.D., Chair

Doctoral Committee

Huiping Xu, Ph.D.

Jianjun Zhang, Ph.D

May 7, 2018

Ying Zhang, Ph.D.

DEDICATION

My humble effort I dedicate to my advisor, teachers, my family and friends, and everyone who helped and encouraged me.

ACKNOWLEDGMENTS

I would like to express sincere gratitude to my advisor Dr. Sujuan Gao for her constant guidance, encouragement and support in my study, my research and my career. Her training will continue benefiting me throughout my journey.

I am thankful to Dr. Ying Zhang, Dr. Huiping Xu and Dr. Jianjun Zhang for kindly serving on my research committee. I am also thankful to Dr. Shanshan Li for valuable comments on my research. I thank my friends Jing Li, Chenghao Chu and others for all the help.

This research is supported by National Institutes of Health (NIH) grant RO1 AG0145350 and NIH grant P30 AG10133.

Mengjie Zheng

JOINT MODELING OF LONGITUDINAL AND SURVIVAL OUTCOMES USING GENERALIZED ESTIMATING EQUATIONS

Joint models for longitudinal and time-to-event data has been introduced to study the association between repeatedly measured exposures and the risk of an event. The use of joint models allows a survival outcome to depend on some characteristic functions from the longitudinal measures. Current estimation methods include a two-stage approach, Bayesian and maximum likelihood estimation (MLEs) methods. The two-stage method is computationally straightforward but often yields biased estimates. Bayesian and MLE methods rely on the joint likelihood of longitudinal and survival outcomes and can be computationally intensive.

In this work, we propose a joint generalized estimating equation framework using an inverse intensity weighting approach for parameter estimation from joint models. The proposed method can be used to longitudinal outcomes from the exponential family of distributions and is computationally efficient. The performance of the proposed method is evaluated in simulation studies. The proposed method is used in an aging cohort to determine the relationship between longitudinal biomarkers and the risk of coronary artery disease.

Sujuan Gao, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

In many cohort studies or clinical trials, participants are routinely followed with regularly scheduled evaluation to collect new symptoms and biological samples until a designated study endpoint is reached. The longitudinal data collected during study follow-up are of great interest in modeling event time. For example, in AIDS studies, it is often of interest to study the association between repeated CD4 counts and the risk of death for HIV-infected patients. In longitudinal dementia research, subjects are evaluated at frequent time intervals to measure potential risk factors. The objective is to identify the relationship between the changes in these exposures and the clinical onset of the disease. The correlation between longitudinal responses and event times also occurs in many other disciplines. For example, in civil engineering, structural integrity indicators are recorded regularly to evaluate the risk of failure of the structure (Rizopoulos, 2012).

When the longitudinal and survival outcomes are correlated, modeling the longitudinal and survival outcome seperately can lead to biased estimates (Ibrahim et al., 2010). Joint modeling of longitudinal and time-to-event outcomes utilizes informaiton collected during the course of study, which can lead to reduced bias and impovements in the efficiencey of statistical inference (Hogan and Laird, 1998). Comprehensive overview of joint models can be found in Tsiatis and Davidian (2004) and Wu et al. (2011). Tsiatis and Davidian (2004) focused on the early work of

joint models while Wu et al. (2011) included more recent methods. In a typical model setting, the longitudinal process is assumed to follow a mixed effect model, and the time-to-event process is captured by a Cox's proportional hazard model or an Accelerated Failure Time (AFT) model, with the longitudinal and time-to-event sharing common random effects. There are two general approaches used for estimations in joint models: a two-stage approach and a likelihood based method.

A naive two-stage method conducts estimation in two separate steps: in the first stage, a linear mixed effects (LME) model is used for the longitudinal data without regard to the time-to-event process; in the second stage, a survival model for time-to-event is used with the predicted longitudinal responses from the first stage as independent variables (Ye et al., 2008a). An advantage of the naive two-stage approach is that by separating out the two models, various methods can be used for one model without affecting the other model. In addition, it can be implemented with standard statistical software without complex estimation techniques (Wu et al., 2011). However, the two-stage method can introduce estimation biases and a lose of efficiency because the longitudinal model fitted at the first stage can be biased without accounting for the survival outcomes (Ghisletta et al., 2006; Faucett and Thomas, 1996; Albert and Shih, 2010; Sweeting and Thompson, 2011; Murawska et al., 2012), and the estimation of the risk coefficient in the survival model does not consider the uncertainty form the predicted longitudinal outcomes.

The two-stage approach has been considered by many authors and there have been attempts at correcting for the biases. Self and Pawitan (1992) used the least square method to estimate the longitudinal covariate trajectories; Tsiatis et al. (1995), Bycott and Taylor (1998), Dafni and Tsiatis (1998) considered an approximation of

the hazard function to account for measurement error in the observed longitudinal markers; Ye et al. (2008b) proposed two approaches: the ordinary regression calibration (ORC) approach that fits the LME model with all observed longitudinal measures, and the risk regression calibration (RRC) method that fits the LME model by removing subjects who have an event before a time point; Albert and Shih (2010) simulated complete longitudinal pseudo-measurements and used the "complete data" in the second stage. Even though various modifications have been proposed to address the informative dropout and measurement error issues, the two-stage approach may still yield biased estimates (Wu et al., 2011).

Rather than using the two-stage approximation, many authors have taken the likelihood approach. In a likelihood based appraoch, subject specific effects are used to model the longitudinal outcomes and predicted individual responses from the longitudinal models are used as covariates in the survival model for time-to-event. Joint likelihood functions from both the longitudinal and survival models are used to obtain parameter estimates and for the statistical inference. Likelihood based approach includes maximum likelihood using EM algorithm and the Bayesian approach. De Gruttola and Tu (1994), Wulfsohn and Tsiatis (1997), Lin et al. (2002), Tseng et al. (2005), Rizopoulos et al. (2009), Wu et al. (2010), Choi et al. (2015) and Yang et al. (2016) considered the likelihood approach and applied the EM algorithm for estimation. Monte Carlo methods or Laplace approximations can be applied for the E step (Rizopoulos et al., 2009; Wu et al., 2010), and Newton-Raphson method is often used for the M step (Yang et al., 2016). Various researchers have also used the Bayesian approach in joint models, including Faucett and Thomas (1996), Xu and Zeger (2001a), Xu and Zeger (2001b), Wang and Taylor (2001), Law et al. (2002),

R Brown and G Ibrahim (2003), Brown and Ibrahim (2003), Ibrahim et al. (2004), Chi and Ibrahim (2006), Rizopoulos and Ghosh (2011), Sweeting and Thompson (2011), Huang et al. (2011), Hatfield et al. (2012) Lawrence Gould et al. (2015) and He and Luo (2016). Bayesian methods can be implemented via Markov Chain Monte Carlo (MCMC) techniques. Both the maximum likelihood approach using EM and Bayesian method produce valid and efficient inference if the longitudinal and survival models are correctly specified. However, the biggest challenge to the wide application of the likelihood based approach for joint models is computational complexity (Wu et al., 2011).

Many previous works in the joint modeling literature focused on a single longitudinal outcome that is associated with the time-to-event model. However, in practice, it is common for studies to collect multiple longitudinal outcomes and aim to determine the association between the longitudinal outcomes and time-to-event. Most current methodologies use joint modeling of linear mixed effect models for multivariate longitudinal outcomes and proportional hazard models for the survival model while assuming multivariate normal distributions for random effects (Hickey et al., 2016). Taking into account the correlations among multiple longitudinal markers can improve the predictive ability of the joint models (Lin et al., 2002; Brown et al., 2005; Chi and Ibrahim, 2006; Gueorguieva and Sanacora, 2006; Fieuws et al., 2008; McCulloch, 2008; Proust-Lima et al., 2009). The longitudinal outcomes may also be a mixture of discrete and continuous measures that can be potentially predictive of the risk of time-to-event. Rizopoulos and Ghosh (2011) proposed a Bayesian framework for this setting.

In this work, we focus on joint modeling with multiple longitudinal outcomes that are associated with the risk of an event. We are considering a joint generalized estimating equation framework where inverse intensity weighting is used for the longitudinal models to adjust for informative truncation by the event, and a partial likelihood is used for the event time data. Such an approach can allow a straightforward extension of models from normally distributed data to non-normal longitudinal outcomes. In addition, this approach is computationally efficient to carry out by avoiding high dimensional integration. Moreover, we use a partial likelihood for the time-to-event data to avoid estimation of the baseline hazard function, which further simplifies the estimation.

Chapter 2

Joint Models for Multiple Longitudinal Continuous Outcomes and Time-to-Event Data

In this chapter, we introduce joint models for analyzing the association between multiple continuous longitudinal measures and a time-to-event outcome. In section 2.2, we define the notation and model formulation of longitudinal and survival processes. We consider multivariate mixed effects models for longitudinal biomarkers without the assumption that the random effects and random errors are normally distributed. The survival model is assumed to follow the Cox proportional hazard model with a function of the true longitudinal measures as time-dependent covariates. We review the background on generalized estimating equations and describe our estimating approach in section 2.1. The proposed method is applied to data from a longitudinal cohort in section 2.5.

2.1 Introduction

Generalized estimating equations (GEE) offers a robust estimation method to capture the characteristics of correlated data from repeated measures (Liang and Zeger, 1986). There are many advantages of GEE: GEE is closely related to quasi-likelihood (Wedderburn, 1974), which requires only the relationship between the expected value of outcome variable and covariates, and the relationship between the conditional mean and variance of response variable; GEE allows the specification of a working corre-

lation matrix that accounts for within-subject correlation of responses on dependent variables; asymptotic consistency of GEE estimators holds even with the misspecification of the working correlation structure; when the working correlation is correctly specified, GEE estimators are efficient within the linear estimating function family; in addition, estimates using GEE are computationally efficient (Liang and Zeger, 1986; Zeger et al., 1988; Liang et al., 1992; Zorn, 2001; McCulloch and Neuhaus, 2001).

The GEE approach can also be used in data with multiple longitudinal outcomes. Rochon (1996) considered a GEE approach for bivariate repeated measures of discrete and continuous outcomes. Shelton et al. (2004) described a SAS macro for the analysis of multivariate longitudinal binary outcomes with GEE. Lipsitz et al. (2009) proposed a joint estimation of the marginal models by using a single modified GEE for longitudinal binary outcomes with missing data. Asar and Ilk (2013) extended GEE to multivariate longitudinal binary, continuous and count responses and developed the R package mmm.

Longitudinal measures related to a time-to-event outcome could be considered as a case of missing data with outcome-dependent follow-up. That is, subject's data are observed at certain intervals until an event occurs. Robins et al. (1995) proposed a class of weighted estimating equations that lead to consistent and asymptotically normal estimators for monotone missing data patterns. This technique weighted a subject's contribution to the estimating equation by the inverse of the conditional probability of being observed. The most recent review of inverse probability weighting (IPW) for dealing with missing data is provided by Seaman and White (2013). To handle irregular visits and outcome-dependent follow-up, Lin et al. (2004) proposed an inverse intensity of visit process weighted GEEs that can handle arbitrary pat-

terns of missing data when longitudinal responses are observed in continuous times. However, this method requires an estimation of baseline intensity of the observation process. Buzkova and Lumley (2007) proposed a similar approach, inverse intensity rate ratio (IIRR) weighted estimating equation, that avoids the smoothing of the baseline hazard, and the estimation could be easily implemented in standard statistical software.

In joint models, the estimated true longitudinal measures are considered as time-dependent covariates in the survival model. The uncertainty of the predicted random effects in longitudinal covariates needs to be taken into account. In the survival model, the adjustment of the variability of the random effects can be considered as a setting of measurement error models. Several methods have been developed to deal with measurement error starting with Prentice (1982). Augustin and Schwarz (2002) provided a review and comparison of methods for Cox's proportional hazards model with covariate measurement errors. Nakamura (1990), Nakamura (1992), Kong and Gu (1999), Kulich and Lin (2000), Hu and Lin (2002), Zucker and Spiegelman (2008) have adopted the method of correcting the partial likelihood score function. The idea is to construct a corrected score function such that the expectation with respect to the measurement error distribution is equal to the original score function based on the unknown true variables. Pepe et al. (1989), Hughes (1993), Wang et al. (1997), Xie et al. (2001) and Spiegelman et al. (2001) etc. have considered the regression calibration method, which replaces the unobserved true variables by their expectation given the observed variables. The regression calibration method is the most well-known tool for measurement error correction, the main advantage is its easy implementation and the estimates can be obtained by standard software

(Augustin and Schwarz, 2002). Some other approaches including Huang and Wang (2000) who had rewritten the partial score function as a function of empirical process, Gorfine et al. (2004) considered weighted estimating equations method, Song et al. (2002) and Li and Ryan (2006) treated measurement error as missing data problem and used multiple imputation for measurement error correction, and Zucker (2005) presented a pseudo-partial likelihood procedure.

In this work, first, we use the weighted GEE to reduce biases in estimating the longitudinal model parameters. Second, in the survival model, we account for the measurement error in the predicted subject-specific random effects in order to further reduce biases in estimating risk coefficients.

2.2 Model Formulation

2.2.1 Longitudinal Models

Let $\mathbf{Y}_i = (\mathbf{y}_{i1}^T, \cdots, \mathbf{y}_{iL}^T)^T$ denotes L-variate response vector for the i-th subject, $\mathbf{y}_{il} = (y_{il}(t_{i1}), \cdots, y_{il}(t_{in_i}))^T$ is a $n_i \times 1$ vector of longitudinal response at different time points, where $i = 1, \cdots, N$ and $l = 1, \cdots, L$. $y_{il}(t_{ij})$ is the l-th response collected on subject i at time point t_{ij}, where $j = 1, \cdots, n_i$. Let $\mathbf{t}_i = (t_{i1}, \cdots, t_{in_i})^T$ denotes time points for subject i. Assuming the l-th longitudinal outcome has the set of covariates $\mathbf{X}_{il} = (\mathbf{X}_{il}^T(t_{i1}), \cdots, \mathbf{X}_{il}^T(t_{in_i}))^T$, where $\mathbf{X}_{il}^T(t_{ij}) = (1, x_{il1}(t_{ij}), \cdots, x_{ilp_l}(t_{ij}))^T$ is a $(p_l + 1) \times 1$ vector of covariates for the l-th response of subject i at time t_{ij}. The l-th response for i-th subject is modeled by the following mixed effects model:

$$\begin{aligned} y_{il}(t) &= y_{il}^*(t) + \epsilon_{il}(t) \\ &= \mathbf{X}_{il}^T(t)\boldsymbol{\beta}_l + \mathbf{Z}_{il}^T(t)\mathbf{b}_{il} + \epsilon_{il}(t), \end{aligned} \tag{2.1}$$

where $\mathbf{X}_{il}^T(t)$ and $\mathbf{Z}_{il}^T(t)$ denote row vectors of the design matrices for the fixed and random effects corresponding to the l-th longitudinal outcome for the i-th subject at time t; $\boldsymbol{\beta}_l$ is the corresponding fixed effects; $\mathbf{b}_{il}$ is the vector of random effects with mean $\mathbf{0}$ and variance $\mathbf{G}_l$; $\epsilon_{il}(t)$ is a random error term with mean 0 and variance σ_l^2.

To take into account the correlation for the multiple longitudinal markers, the responses for the i-th subject can be modeled by the following mixed effects model:

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{2.2}$$

where $\mathbf{X}_i = \begin{pmatrix} \mathbf{X}_{i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{i2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{X}_{iL} \end{pmatrix}_{(\sum_{j=1}^{L}(p_j+1))\times(\sum_{j=1}^{L}(p_j+1))}$,

$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_L \end{pmatrix}_{(\sum_{j=1}^{L}(p_j+1))\times 1}$, $\mathbf{Z}_i = \begin{pmatrix} \mathbf{Z}_{i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{i2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{Z}_{iL} \end{pmatrix}$,

$\mathbf{b}_i = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_L \end{pmatrix}$ with covariane matrix $\mathbf{G} = \begin{pmatrix} \mathbf{G}_1 & \mathbf{G}_{12} & \cdots & \mathbf{G}_{1L} \\ \mathbf{G}_{21} & \mathbf{G}_2 & \cdots & \mathbf{G}_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}_{L1} & \cdots & \mathbf{G}_{L(L-1)} & \mathbf{G}_L \end{pmatrix}$,

$\boldsymbol{\epsilon}_i = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_L \end{pmatrix}$ represent independent random errors, with covariance matrix

$\boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma_1^2\mathbf{I}_{n_i} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sigma_2^2\mathbf{I}_{n_i} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \sigma_L^2\mathbf{I}_{n_i} \end{pmatrix}$, where $\mathbf{I}_{n_i}$ is $n_i \times n_i$ identity matrix.

Assume $\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i$ are mutually independent, then the variance covariance of $\mathbf{Y}_i$ is $\mathbf{V}_i = Cov(\mathbf{Y}_i) = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^T + \boldsymbol{\Sigma}_i$ and $E(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta}$. Note that $\mathbf{Y}_i$'s are observed until an time-to-event occurs. That is, longitudinal data are unbalanced follow-ups with irregular time points of observation and also with right "censoring" missing caused by the survival outcome.

2.2.2 The Survival Model

For a time-to-event outcome, let T_i^* be the true event time for subject i. Some subjects may not experience the event at the end of study, or they may drop out during the study, so their event times are right censored. Let C_i denote the censoring time for the i-th subject. The observed event time $T_i = \min(T_i^*, C_i)$. Define the event indicator as $\delta_i = I(T_i^* \leq C_i)$, here $I(\cdot)$ represents indicator function. Assume that censoring times and event times are independent. The effects of longitudinal

outcomes on survival times are captured by a hazard model:

$$h_i(t) = h_0(t) \exp \left\{ \mathbf{u}_i^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l \left(y_{il}^*(t) \right) \right\}, \tag{2.3}$$

where $h_0(t)$ is the baseline hazard; $\mathbf{u}_i$ denotes a vector of baseline covariates with corresponding regression coefficient vector $\boldsymbol{\gamma}$; $f_l(\cdot)$ is a known function of the true longitudinal profile for the l-th outcome; and α_l denotes a parameter measuring the effect of the l-th longitudinal process on the time to event outcome. Parameters α_l's $(l = 1, 2, \cdots, L)$ are of primary interest in this work.

Function $f_l(\cdot)(l = 1, 2, \cdots, L)$ are assumed to be known and reflect the association between the longitudinal outcomes and the event process. The identity function quantifies the effect of the true current underlying longitudinal outcomes on the hazard for an event; a first derivative with respect to time t implies that the risk for an event at time t depends on the slope of the true longitudinal trajectory at time t; an integration function over time indicates that the event risk of depends on the cumulative history of the longitudinal biomarkers over time.

2.3 Estimation Method

Three broad approaches have been used for parameter estimation for the joint models defined above. The two-stage method derives predicted longitudinal outcomes from mixed effects models fitted to the longitudinal data without considering the survival model at the first stage. Cox's models are used at the second stage using the predicted longitudinal outcomes as covariates without accounting for the prediction errors. The two stage method has been shown to result in biased results (Wu et al.,

2011). Maximum likelihood approaches and Bayesian methods that utilize the full likelihood based on the joint models have been proposed. However, these methods can be computationally intensive as the likelihood function or the posterior distribution function involves integration over multiple random effects.

In this section, we present the joint generalized estimating equation methodology for parameter estimation of joint models for multiple longitudinal biomarkers and time-to-event outcome.

2.3.1 The Joint Generalized Estimating Equation Method

Hypothetically, if we knew parameters in the survival model, a weight can be calculated using the inverse visit intensity by Lin et al. (2004) and the weight can be used in a weighted GEE to obtain parameter estimates for the fixed effect, $\boldsymbol{\beta}$, in the longitudinal model. Individual random effects can then be obtained by minimizing the squared "distance" between observed individual responses and predicted population averages. The new estimates can be used in the survival models to provide the next round of parameter estimates. The process iterates until converges.

Let $\mathbf{W}_i$ be the diagonal weight matrix, $\mathbf{U}(\boldsymbol{\beta})$, $\mathbf{D}(\mathbf{b})$, $L(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ are objective functions for solving $\boldsymbol{\beta}$, $\mathbf{b}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ respectively. We propose a joint generalized estimating equation (JGEE) approach for estimation, which is to solve the following systematic estimating equations:

Estimate $\boldsymbol{\beta}$ from:

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{W}_i (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta}) = \mathbf{0}. \tag{2.4}$$

Predict random effects $\mathbf{b}$ from:

$$\mathbf{b} = \underset{\mathbf{b}}{\operatorname{argmin}} \sum_{i=1} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i). \tag{2.5}$$

Estimate risk coefficients $\boldsymbol{\alpha}$ from:

$$\boldsymbol{\alpha} = \underset{\boldsymbol{\alpha},\boldsymbol{\gamma}}{\operatorname{argmax}} \prod_{i=1}^{N} \left[\frac{e^{\mathbf{u}_i^T\boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l(y_{il}^*(T_i))}}{\sum_{j=1}^{N} I(T_j \geq T_i) e^{\mathbf{u}_j^T\boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l(y_{jl}^*(T_j))}} \right]^{\delta_i}. \tag{2.6}$$

Equation (2.4) is a weighted GEE for longitudinal outcomes. In equation (2.4), $\mathbf{Y}_i$ is the multivariate response of the i-th subject with dimension $(L \times n_i) \times 1$; $\mathbf{X}_i$ is the design matrix with dimension $(\sum_{j=1}^{L}(p_j + 1)) \times (\sum_{j=1}^{L}(p_j + 1))$, and $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients for the fixed effects with dimension $(\sum_{j=1}^{L}(p_j + 1)) \times L) \times 1$. $\mathbf{V}_i$ is the variance covariance matrix of $\mathbf{Y}_i$ with dimension $(L \times n_i) \times (L \times n_i)$.

$\mathbf{W}_i$ is a $(L \times n_i) \times (L \times n_i)$ diagonal weight matrix with the j-th diagonal element as

$$W(t_{ij}; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = \frac{c(t_{ij}; \mathbf{u}_i, \boldsymbol{\gamma}_0)}{h_0(t_{ij}) \exp\left\{ \mathbf{u}_i^T\boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l(y_{il}^*(t_{ij}-)) \right\}},$$

where $y_{il}^*(t_{ij}-)$ denotes all the past observations until time t_{ij}. Note that $\mathbf{W}$ can use $\mathbf{Y}$'s collected right up to time t_{ij}. This weight is called inverse visit intensity weight motivated by Lin et al. (2004) and Buzkova and Lumley (2007), and it is used to account for missing caused by time-to-event. $c(\cdot)$ is a pre-specified function of the time independent covriates $\mathbf{u}$ and has the potential to stabilize the influence of small values

in the denominator. In addition, if we choose $c(\cdot)$ as $h_0(t_{ij}) \exp\left\{\mathbf{u}_i^T \boldsymbol{\gamma}_0\right\}$, when $y_i(t)$'s have no influence on the risk of event at time t, the weight becomes identity matrix reducing to the regular GEE for equation (2.4). Note that longitudinal measures at different time points will assign different weights, but multiple longitudinal outcomes will share the same weight at the same time point. The solution to the weighted GEEs (2.4) can be found by applying the Fisher Scoring method (Nelder and Baker, 1972). The algorithm is to get initial estimates of $\boldsymbol{\beta}$, $\mathbf{W}_i$ and $\mathbf{V}_i$, update $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_{r+1} = \boldsymbol{\beta}_r + \left[\sum_{i=1}^N \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{W}_i \mathbf{X}_i\right]^{-1} \left[\sum_{i=1}^N \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{W}_i (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta}_r)\right]$ until converge.

Equation (2.5) minimizes the squared distance between an individual's response measures to the population average. $\mathbf{b}$ can be predicted as minimizing the squared length of residual vector:

$$\mathbf{b} = \underset{\mathbf{b}}{\operatorname{argmin}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})^T \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}).$$

Thus $\mathbf{b}$ is the generalized least squares estimates with explicit form

$$\hat{\mathbf{b}} = \left(\mathbf{Z}^T \mathbf{V}^{-1} \mathbf{Z}\right)^{-1} \mathbf{Z}^T \mathbf{V}^{-1} \left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right),$$

with $\boldsymbol{\beta}$ is replaced by the consistent estimate $\hat{\boldsymbol{\beta}}$.

Equation (2.6) is to maximize the partial likelihood function for Cox proportional hazard model with unobserved longitudinal measures replaced by estimated longitudinal measures from equations (2.4) and (2.5):

$$\hat{y}_{il}^*(t_{ij}) = \mathbf{X}_i^T(t_{ij}) \hat{\boldsymbol{\beta}}_l + \mathbf{Z}_i^T(t_{ij}) \hat{\mathbf{b}}_{il}^*,$$

where $\hat{\mathbf{b}}_{il}^*$ is an adjusted predicted random effect given by equation (2.5). The corresponding log partial likelihood function is:

$$l(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \sum_{i=1}^{N} \delta_i \left[\mathbf{u}_i^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f(\hat{y}_{il}^*(T_i)) \right. \\ \left. - \log \sum_{j=1}^{N} I(T_j \geq T_i) \exp\{\mathbf{u}_j^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f(\hat{y}_{jl}^*(T_i))\} \right]. \tag{2.7}$$

Coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ can be obtained by maximizing this log partial likelihood.

Note that the weight $\mathbf{W}_i$'s in equation (2.4), and $y_{il}^*(T_i)$'s in equation (2.6) depend on the true model parameters, which are unknown in reality. Equations (2.4),(2.5) and (2.6) can be solved iteratively with $\mathbf{W}_i$, $y_{il}^*(T_i)$ replaced by their estimators. The initial value of $\mathbf{W}_i$ can be obtained from a naive two-stage method; equation (2.4) can provide an estimator of fixed effects $\hat{\boldsymbol{\beta}}$; given $\hat{\boldsymbol{\beta}}$, equation (2.5) can provide a prediction of random effects; from equation (2.4) and (2.5), an estimator y^* can be obtained and $f_l(\hat{y}_{il}^*(t)) = f_l(\mathbf{X}_{il}^T(t)\hat{\boldsymbol{\beta}}_l + \mathbf{Z}_{il}^T(t)\hat{\mathbf{b}}_{il}^*)$ can be plugged into equation (2.6). The three equations can be solved iteratively.

2.3.2 Variance Estimates

For general estimating equations, Liang and Zeger (1986) have proposed a sandwich variance covariance matrix estimator for $\hat{\boldsymbol{\beta}}$ the robust variance covariance estimator. Following the sandwich variance covariance of GEEs, the robust sandwich variance covariance matrix of $\hat{\boldsymbol{\beta}}$ from weighted GEEs will be calculated by:

$$\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{W}\mathbf{X}\right)^{-1} \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{W}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{W}\mathbf{V}^{-1}\mathbf{X}\right) \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{W}\mathbf{X}\right)^{-1}, \tag{2.8}$$

with $\mathbf{V}$, $\boldsymbol{\beta}$ and $\mathbf{W}$ are replaced by estimates.

The variance estimate of $\hat{\boldsymbol{\alpha}}$ will be obtained by inverse of the information matrix of the log partial likelihood:

$$V(\hat{\boldsymbol{\alpha}}) = -\left[\frac{\partial^2 l(\hat{\boldsymbol{\alpha}})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T}\right]^{-1}. \tag{2.9}$$

In Chapter 4, we will provide theoretical support of the asymptotic properties for the estimators obtained from the proposed iterative joint GEE approach. The proposed approach uses an iterative computational approach and avoids high dimensional integration required in likelihood or Bayesian methods. The proposed method can be implemented in standard statistical software. Because the standard errors of fixed effects in the longitudinal model are calculated based on modified GEE as if the weights are known, and the standard errors of the risk coefficient estimators are obtained through partial likelihood assuming that all the true covariate values are known, estimated standard errors are likely to be underestimating the true variance. In the following section, we investigate the finite sample performance of the proposed method in simulation studies.

2.4 Simulation Study

In this section, we perform simulation studies to assess the performance of the proposed method. Simulation results are compared with those obtained using the two-stage method and the GEE approach. The GEE method is solving equations (2.4), (2.5) and (2.6) with $\mathbf{W}$ be identity matrix in (2.4). Model specification and data generation process in the simulation study follows part of Yang's strategy (Yang,

2013). Data are simulated from the joint models with two correlated normally distributed longitudinal outcome and time-to-event outcome. Similar fixed and random effects structures are considered for two longitudinal outcomes, where the fixed effects include intercept, time, and one baseline binary covariate, and the random effects include random intercept and slope. Two longitudinal outcomes are correlated through the random effects.

Suppose longitudinal outcomes follow the model:

$$y_1(t_{ij}) = y_1^*(t_{ij}) + \epsilon_1(t_{ij}) = \beta_{01} + \beta_{11}t_{ij} + \beta_{21}u_i + b_{01i} + b_{11i}t_{ij} + \epsilon_1(t_{ij}),$$

$$y_2(t_{ij}) = y_2^*(t_{ij}) + \epsilon_2(t_{ij}) = \beta_{02} + \beta_{12}t_{ij} + \beta_{22}u_i + b_{02i} + b_{12i}t_{ij} + \epsilon_2(t_{ij}),$$

where $\epsilon_1(t_{ij}) \sim N(0, \sigma_1^2)$, $\epsilon_2(t_{ij}) \sim N(0, \sigma_2^2)$ and

$$\begin{pmatrix} b_{01i} \\ b_{11i} \\ b_{02i} \\ b_{12i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{01}^2 & \rho_1\sigma_{01}\sigma_{11} & \rho_2\sigma_{01}\sigma_{02} & \rho_1\sigma_{01}\sigma_{12} \\ \rho_1\sigma_{01}\sigma_{11} & \sigma_{11}^2 & \rho_1\sigma_{11}\sigma_{02} & \rho_1\sigma_{11}\sigma_{12} \\ \rho_2\sigma_{01}\sigma_{02} & \rho_1\sigma_{11}\sigma_{02} & \sigma_{02}^2 & \rho_1\sigma_{02}\sigma_{12} \\ \rho_1\sigma_{01}\sigma_{12} & \rho_1\sigma_{11}\sigma_{12} & \rho_1\sigma_{02}\sigma_{12} & \sigma_{12}^2 \end{pmatrix}\right).$$

Assume the time-to-event outcomes follow Cox Proportional Hazard (PH) Model with Weibull baseline function:

$$h(t) = abt^{b-1}\exp(\alpha_1 y_1^*(t) + \alpha_2 y_2^*(t)), \tag{2.10}$$

where a is the shape parameter and b is the scale parameter of the Weibull distribution. The Cox PH model is assumed to depend on the current values of the two longitudinal outcomes at event time points. For each subject, a survival probability s_i at the true even time T_i^* is simulated from uniform $(0, 1)$, and the true event time T_i^* is solved from the equation:

$$s_i - \exp\left\{-\int_0^{T_i^*} abt^{b-1}\exp(\alpha_1 y_1^*(t) + \alpha_2 y_2^*(t))dt\right\} = 0.$$

Censoring times are simulated from another uniform distribution independently. The event indicator δ_i for the ith subject is determined by comparing true event time T_i^* and censoring time C_i, $\delta_i = I(T_i^* \leq C_i)$. The observed time $T_i = \min(T_i^*, C_i)$. Four censoring percentages and three different sample sizes are considered. The true parameters used in the simulations are summarized in Table 2.1 and Table 2.2.

The proposed JGEE algorithm will be performed as the following steps:

Step 1: Run native two-stage method to estimate the initial values for all unknown parameters. Specially, the mixed effect model (2.2) with multiple biomarkers is used to calculate the initial values in longitudinal models, including $\hat{\beta}_{01}$, $\hat{\beta}_{11}$, $\hat{\beta}_{02}$, $\hat{\beta}_{02}$, $\hat{\beta}_{12}$ and $\hat{\beta}_{22}$. Then the Cox PH model using estimated longitudinal biomarker measures at the event times as time-dependent covariates is used to obtain the initial values for parameters in the hazard function, including $\hat{\alpha}_1$, $\hat{\alpha}_2$.

Step 2: Obtain initial values of weights using estimated parameters, specially, diagonal values of the weight matrix have the form:

$$W_i(t_{ij}) = \frac{1}{\exp\left\{\hat{\alpha}_1\left(\hat{\beta}_{01} + \hat{\beta}_{11}t_{ij} + \hat{\beta}_{21}u_i\right) + \hat{\alpha}_2\left(\hat{\beta}_{02} + \hat{\beta}_{12}t_{ij} + \hat{\beta}_{22}u_i\right)\right\}}.$$

Step 3: Solve longitudinal fixed effect parameters from weighted joint generalized estimation equations (2.4):

$$\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{W}_i (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta}) = \mathbf{0},$$

where

$$\mathbf{X}_i^T = \begin{bmatrix} 1 & t_{i1} & u_i & 0 & 0 & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 1 & t_{in_i} & u_i & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & t_{i1} & u_i \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & 1 & t_{in_i} & u_i \end{bmatrix}^T,$$

Denote estimated parameters from the weighted GEE in (2.4) as $\hat{\beta}_{01}^*$, $\hat{\beta}_{11}^*$, $\hat{\beta}_{02}^*$, $\hat{\beta}_{02}^*$, $\hat{\beta}_{12}^*$ and $\hat{\beta}_{22}^*$.

Step 4: Obtain individual random effects $\hat{\mathbf{b}}$ form equation (2.5) by solving random effects model $\mathbf{Y} - \mathbf{X}^T \hat{\boldsymbol{\beta}} = \mathbf{Z}^T \mathbf{b} + \boldsymbol{\epsilon}$. We provide a correction to account for the variability in estimated $\hat{\mathbf{b}}_i$. Similar to idea of regression calibration in measurement error models (Carroll et al., 1995), we use $\hat{b}_{il}^* = \hat{b}_{il} / \sqrt{1 + Var(\hat{\mathbf{b}}_l)}$. Calibrated random effects depend on variation of predicted subject specific random effects; if the predicted random effects have large variance, the calibrated random effects shrink them to return less variable random effects.

Step 5: Estimated longitudinal measures $\hat{y}_1^*(t_{ij}) = \hat{\beta}_{01}^* + \hat{\beta}_{11}^* t_{ij} + \hat{\beta}_{21}^* w_i + \hat{b}_{01i}^* + \hat{b}_{11i}^* t_{ij}$ and $\hat{y}_2^*(t_{ij}) = \hat{\beta}_{02}^* + \hat{\beta}_{12}^* t_{ij} + \hat{\beta}_{22}^* w_i + \hat{b}_{02i}^* + \hat{b}_{12i}^* t_{ij}$ are used as time-varying

covariates in the survival model (2.10). The Cox model procedure is used with the partial likelihood

$$L(\alpha_1, \alpha_2) = \prod_{i=1}^{N} \left[\frac{\alpha_1 \hat{y}^*_{i1}(T_i) + \alpha_2 \hat{y}^*_{i2}(T_i)}{\sum_{j=1}^{N} I(T_j \geq T_i) e^{\alpha_1 \hat{y}^*_{j1}(T_j) + \alpha_2 \hat{y}^*_{j2}(T_j)}} \right]^{\delta_i}$$

to obtain new estimates $\hat{\alpha}^*_1$ and $\hat{\alpha}^*_2$.

Step 6: Repeat step 1 to 5 until converge.

In this simulation, 500 Monte Carlo samples are generated. Longitudinal data are simulated for 200, 500, 1000 subjects with 10 equally spaced bivariate longitudinal observations over 5 year period. Survival data are simulated for 40%, 30%, 10% and 0% of censoring. Longitudinal observations are right censored by time-to-event data. Simulation results are presented in Tables 2.3 to 2.6. Relative bias (defined as $((\hat{\theta} - \theta_{true})/\theta_{true})$, empirical standard errors (Emp. S.E.), model based standard errors (Model S.E.), coverage probability of the 95% confidence intervals (95% CI C.P.) and Mean Square Error (MSE) (defined as $E\left[(\hat{\theta} - \theta)^2\right]$) based on 500 Monte Carlo samples are reported.

From simulation results, it can be observed that for a fixed survival censoring percentage with a given sample size, proposed JGEE method performs better than GEE and two-stage method with smaller relative bias, larger 95% coverage probability and comparable MSE. For a fixed survival censoring percentage and with sample size gets larger, relative bias, standard error, 95% coverage probability and MSE get smaller as expected for proposed JGEE method; and if we compare different methods on relative bias improvement, with sample size gets larger, the improvement gets better (relative bias from two-stage and GEE method tends to get larger while

JGEE method relative bias gets smaller). We also observe that JGEE method tends to underestimate the standard error, this is because the standard errors of β's are calculated based on modified GEE as if the weights are known, and the standard errors of α's are obtained through partial likelihood assuming that all the true covariate values are known. It is also observed that the GEE method always performs worse than the naive Two-stage method, this is because for longitudinal data, missing caused by survival information is not missing completely at random, regular GEE leads to more biased results on longitudinal estimates, thus worse survival estimates. This also reflects the proposed JGEE method adding weight in the GEE method does adjust for missing data.

For a given sample size, it can be observed from simulation results that, with censoring percentage gets smaller, proposed JGEE approach performs better with smaller relative bias and smaller MSE, while GEE and two-stage method get larger relative bias and MSE. This is because in the longitudinal part, missingness is caused by either censoring or time-to-event; missing data come from censoring are non-informative, which do not affect parameter estimation; while missing caused by event is nonignorable missing, the less censoring, the more events have happened, thus more bias would be for longitudinal estimates if the missing is ignored. On the other hand, survival effect depends on true unobserved longitudinal measures, the more bias in longitudinal estimation, the more bias result in survival estimation. Because JGEE method did adjust for missingness related to event by adding weight into estimation equation to acquire better longitudinal estimates, with better longitudinal parameter estimates, proposed JGEE method yield better survival parameter estimates.

Table 2.1: True parameter values for the two longitudinal outcomes and the proportional hazard function.

Outcome 1		Outcome 2		Time to event Outcome	
Parameter	Value	Parameter	Value	Parameter	Value
β_{01}	0.2	β_{02}	1	a	0.005
β_{11}	0.5	β_{12}	0.2	b	1.1
β_{21}	0.2	β_{22}	0.5	α_1	1.0
				α_2	1.5

Table 2.2: True values for random errors, random intercept and correlation for the longitudinal models, and censoring percentage from survival.

Scenario	σ_1	σ_2	σ_{01}	σ_{11}	σ_{02}	σ_{12}	ρ_1	ρ_2	Censor
1	0.5	0.5	0.5	0.2	0.5	0.2	0.2	0.4	40%
2	0.5	0.5	0.5	0.2	0.5	0.2	0.2	0.4	30%
3	0.5	0.5	0.5	0.2	0.5	0.2	0.2	0.4	10%
4	0.5	0.5	0.5	0.2	0.5	0.2	0.2	0.4	0%

Table 2.3: Simulation results for comparing the two-stage approach with the joint generalized estimating equation (JGEE) approach (40% censoring).

		Two-Stage					GEE					JGEE				
Sample Size	True Parm	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE
$n = 200$	$\beta_{01}(0.2)$	0.228	0.063	0.060	87.4%	0.006	0.373	0.070	0.065	75.6%	0.010	-0.101	0.063	0.062	94.0%	0.004
	$\beta_{11}(0.5)$	-0.046	0.027	0.025	83.0%	0.001	-0.110	0.032	0.031	55.0%	0.004	0.035	0.032	0.031	91.0%	0.001
	$\beta_{21}(0.2)$	-0.132	0.082	0.083	95.4%	0.007	-0.233	0.095	0.093	91.6%	0.011	-0.077	0.084	0.086	93.8%	0.007
	$\beta_{02}(1.0)$	0.058	0.060	0.060	85.0%	0.007	0.096	0.062	0.064	68.8%	0.013	0.030	0.061	0.062	90.6%	0.005
	$\beta_{12}(0.2)$	-0.176	0.027	0.024	67.6%	0.002	-0.400	0.029	0.028	20.6%	0.007	-0.008	0.035	0.030	90.4%	0.001
	$\beta_{22}(0.5)$	-0.054	0.083	0.082	93.8%	0.008	-0.104	0.091	0.089	89.4%	0.011	-0.029	0.086	0.085	94.4%	0.008
	$\alpha_1(1.0)$	-0.092	0.179	0.174	90.2%	0.041	-0.110	0.173	0.176	89.6%	0.042	0.070	0.209	0.200	93.6%	0.048
	$\alpha_2(1.5)$	-0.083	0.180	0.191	90.2%	0.048	-0.106	0.178	0.193	86.4%	0.057	0.029	0.205	0.211	96.6%	0.044
$n = 500$	$\beta_{01}(0.2)$	0.238	0.038	0.038	76.6%	0.004	0.377	0.042	0.042	55.8%	0.007	0.103	0.039	0.039	91.0%	0.002
	$\beta_{11}(0.5)$	-0.046	0.017	0.016	68.8%	0.001	-0.113	0.020	0.020	18.4%	0.004	0.035	0.022	0.020	81.8%	0.001
	$\beta_{21}(0.2)$	-0.133	0.054	0.053	90.6%	0.004	-0.220	0.061	0.059	87.8%	0.006	-0.066	0.055	0.055	93.6%	0.003
	$\beta_{02}(1.0)$	0.060	0.038	0.038	61.6%	0.005	0.100	0.041	0.041	31.4%	0.012	0.032	0.039	0.039	87.4%	0.003
	$\beta_{12}(0.2)$	-0.175	0.016	0.015	36.0%	0.001	-0.402	0.018	0.018	1.0%	0.007	0.015	0.022	0.019	91.2%	0.000
	$\beta_{22}(0.5)$	-0.064	0.054	0.052	88.6%	0.004	-0.121	0.059	0.057	79.2%	0.007	-0.038	0.056	0.054	92.2%	0.003
	$\alpha_1(1.0)$	-0.085	0.110	0.107	86.0%	0.019	-0.105	0.105	0.108	83.2%	0.022	0.076	0.128	0.122	90.0%	0.022
	$\alpha_2(1.5)$	-0.098	0.117	0.117	74.2%	0.035	-0.123	0.115	0.118	63.0%	0.047	0.013	0.130	0.129	95.4%	0.017
$n = 1000$	$\beta_{01}(0.2)$	0.237	0.027	0.027	60.0%	0.003	0.375	0.029	0.029	26.6%	0.006	0.103	0.028	0.028	87.0%	0.001
	$\beta_{11}(0.5)$	-0.050	0.012	0.011	40.4%	0.001	-0.113	0.014	0.014	1.2%	0.003	0.034	0.015	0.014	76.2%	0.001
	$\beta_{21}(0.2)$	-0.123	0.038	0.037	88.8%	0.002	-0.216	0.042	0.042	83.0%	0.004	-0.060	0.039	0.039	92.8%	0.002
	$\beta_{02}(1.0)$	0.056	0.026	0.027	46.6%	0.004	0.096	0.029	0.029	8.4%	0.010	0.028	0.027	0.028	84.2%	0.002
	$\beta_{12}(0.2)$	-0.170	0.012	0.011	12.4%	0.001	-0.398	0.014	0.013	0.0%	0.007	0.021	0.016	0.014	91.6%	0.000
	$\beta_{22}(0.5)$	-0.056	0.036	0.037	88.8%	0.002	-0.112	0.040	0.040	70.4%	0.005	-0.031	0.038	0.038	92.0%	0.002
	$\alpha_1(1.0)$	-0.092	0.079	0.075	73.6%	0.015	-0.111	0.076	0.075	65.2%	0.018	0.069	0.091	0.085	84.6%	0.013
	$\alpha_2(1.5)$	-0.105	0.078	0.081	50.6%	0.031	-0.130	0.077	0.082	32.8%	0.044	0.006	0.089	0.090	96.0%	0.008

Table 2.4: Simulation results for comparing the two-stage approach with the joint generalized estimating equation (JGEE) approach (30% censoring).

		Two-Stage					GEE					JGEE				
Sample Size	True Parm	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE
$n = 200$	$\beta_{01}(0.2)$	0.235	0.064	0.060	86.4%	0.006	0.399	0.070	0.065	73.4%	0.011	-0.087	0.063	0.061	94.0%	0.004
	$\beta_{11}(0.5)$	-0.050	0.026	0.024	79.4%	0.001	-0.115	0.030	0.029	49.4%	0.004	0.039	0.031	0.030	90.0%	0.001
	$\beta_{21}(0.2)$	-0.135	0.082	0.083	95.2%	0.007	-0.255	0.096	0.094	90.8%	0.012	-0.072	0.085	0.086	94.2%	0.007
	$\beta_{02}(1.0)$	0.061	0.060	0.060	83.8%	0.007	0.104	0.062	0.064	63.4%	0.015	0.028	0.062	0.061	91.2%	0.005
	$\beta_{12}(0.2)$	-0.187	0.026	0.022	62.6%	0.002	-0.425	0.027	0.026	10.4%	0.008	-0.013	0.033	0.029	90.2%	0.001
	$\beta_{22}(0.5)$	-0.057	0.083	0.082	93.6%	0.008	-0.112	0.090	0.089	89.0%	0.011	-0.028	0.085	0.085	95.2%	0.007
	$\alpha_1(1.0)$	-0.097	0.159	0.155	89.2%	0.035	-0.115	0.155	0.156	87.8%	0.037	0.070	0.188	0.178	93.4%	0.040
	$\alpha_2(1.5)$	-0.089	0.170	0.173	86.2%	0.047	-0.114	0.169	0.175	82.8%	0.057	0.029	0.194	0.191	94.6%	0.040
$n = 500$	$\beta_{01}(0.2)$	0.247	0.038	0.038	74.4%	0.004	0.403	0.042	0.042	50.6%	0.008	0.092	0.0392	0.039	92.0%	0.002
	$\beta_{11}(0.5)$	-0.049	0.017	0.015	62.0%	0.001	-0.118	0.019	0.019	10.4%	0.004	0.039	0.021	0.019	77.6%	0.001
	$\beta_{21}(0.2)$	-0.137	0.054	0.053	90.8%	0.004	-0.238	0.062	0.060	86.8%	0.006	-0.062	0.055	0.054	93.4%	0.003
	$\beta_{02}(1.0)$	0.063	0.037	0.038	60.8%	0.005	0.108	0.041	0.040	24.2%	0.013	0.030	0.039	0.039	88.8%	0.002
	$\beta_{12}(0.2)$	-0.186	0.015	0.014	28.2%	0.002	-0.426	0.017	0.017	0.2%	0.008	0.021	0.021	0.018	91.2%	0.000
	$\beta_{22}(0.5)$	-0.064	0.054	0.052	89.0%	0.004	-0.130	0.059	0.057	78.0%	0.008	-0.037	0.056	0.054	91.6%	0.003
	$\alpha_1(1.0)$	-0.094	0.100	0.095	80.6%	0.019	-0.115	0.099	0.096	75.6%	0.023	0.071	0.117	0.109	90.0%	0.019
	$\alpha_2(1.5)$	-0.104	0.109	0.106	66.4%	0.036	-0.129	0.108	0.107	54.8%	0.049	0.014	0.122	0.117	94.0%	0.015
$n = 1000$	$\beta_{01}(0.2)$	0.243	0.027	0.027	57.0%	0.003	0.402	0.029	0.029	21.0%	0.007	0.090	0.027	0.028	89.0%	0.001
	$\beta_{11}(0.5)$	-0.052	0.012	0.011	34.4%	0.001	-0.120	0.013	0.0131	0.2%	0.004	0.037	0.014	0.013	68.4%	0.001
	$\beta_{21}(0.2)$	-0.124	0.038	0.037	89.4%	0.002	-0.232	0.042	0.042	82.4%	0.004	-0.055	0.039	0.038	93.0%	0.002
	$\beta_{02}(1.0)$	0.058	0.026	0.027	43.0%	0.004	0.103	0.028	0.029	4.8%	0.011	0.026	0.027	0.028	84.6%	0.001
	$\beta_{12}(0.2)$	-0.180	0.011	0.010	6.0%	0.001	-0.419	0.012	0.012	0.0%	0.007	0.026	0.015	0.013	90.2%	0.000
	$\beta_{22}(0.5)$	-0.058	0.036	0.037	87.8%	0.002	-0.121	0.041	0.040	65.8%	0.005	-0.030	0.038	0.038	92.6%	0.002
	$\alpha_1(1.0)$	-0.100	0.070	0.066	63.2%	0.015	-0.120	0.068	0.067	53.2%	0.019	0.065	0.082	0.076	85.0%	0.011
	$\alpha_2(1.5)$	-0.112	0.074	0.074	37.6%	0.034	-0.137	0.074	0.075	21.4%	0.048	0.005	0.085	0.082	94.2%	0.007

Table 2.5: Simulation results for comparing the two-stage approach with the joint generalized estimating equation (JGEE) approach (10% censoring).

		Two-Stage					GEE					JGEE				
Sample Size	True Parm	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE
$n = 200$	$\beta_{01}(0.2)$	0.250	0.062	0.059	85.8%	0.006	0.447	0.068	0.065	71.6%	0.013	0.068	0.062	0.061	94.0%	0.004
	$\beta_{11}(0.5)$	-0.052	0.023	0.021	75.2%	0.001	-0.124	0.026	0.026	33.2%	0.005	0.045	0.028	0.027	86.0%	0.001
	$\beta_{21}(0.2)$	-0.143	0.080	0.082	95.2%	0.007	-0.289	0.096	0.094	89.8%	0.012	-0.066	0.084	0.085	94.6%	0.007
	$\beta_{02}(1.0)$	0.065	0.059	0.059	81.2%	0.008	0.116	0.062	0.063	53.8%	0.017	0.025	0.061	0.061	92.0%	0.004
	$\beta_{12}(0.2)$	-0.203	0.024	0.020	49.0%	0.002	-0.455	0.024	0.023	4.0%	0.009	-0.021	0.031	0.026	90.2%	0.001
	$\beta_{22}(0.5)$	-0.061	0.081	0.081	94.0%	0.007	-0.113	0.088	0.088	88.8%	0.012	-0.026	0.085	0.084	94.6%	0.007
	$\alpha_1(1.0)$	-0.115	0.133	0.128	82.4%	0.031	-0.133	0.129	0.129	79.6%	0.034	0.058	0.159	0.148	92.8%	0.029
	$\alpha_2(1.5)$	-0.110	0.147	0.147	79.2%	0.049	-0.134	0.147	0.149	70.6%	0.062	0.016	0.172	0.164	94.2%	0.030
$n = 500$	$\beta_{01}(0.2)$	0.262	0.038	0.038	70.8%	0.004	0.452	0.042	0.041	40.8%	0.010	0.070	0.039	0.039	92.6%	0.002
	$\beta_{11}(0.5)$	-0.054	0.015	0.014	50.8%	0.001	-0.128	0.017	0.016	3.0%	0.004	0.045	0.019	0.017	72.0%	0.001
	$\beta_{21}(0.2)$	-0.144	0.054	0.052	89.8%	0.004	-0.272	0.062	0.060	85.6%	0.007	-0.054	0.055	0.054	94.0%	0.003
	$\beta_{02}(1.0)$	0.067	0.037	0.037	57.6%	0.006	0.120	0.040	0.040	15.4%	0.007	0.027	0.038	0.038	89.4%	0.002
	$\beta_{12}(0.2)$	-0.200	0.014	0.013	15.2%	0.002	-0.453	0.015	0.015	0.0%	0.008	0.030	0.019	0.016	90.0%	0.000
	$\beta_{22}(0.5)$	-0.069	0.054	0.051	89.0%	0.004	-0.146	0.059	0.056	74.4%	0.009	-0.035	0.055	0.053	92.2%	0.003
	$\alpha_1(1.0)$	-0.113	0.085	0.078	67.0%	0.020	-0.133	0.083	0.079	60.8%	0.025	0.058	0.101	0.091	87.4%	0.014
	$\alpha_2(1.5)$	-0.120	0.093	0.090	48.0%	0.041	-0.145	0.093	0.091	33.4%	0.056	0.005	0.106	0.100	94.0%	0.011
$n = 1000$	$\beta_{01}(0.2)$	0.258	0.026	0.027	52.4%	0.003	0.449	0.029	0.029	12.4%	0.009	0.070	0.027	0.027	91.0%	0.001
	$\beta_{11}(0.5)$	-0.057	0.011	0.010	17.8%	0.001	-0.128	0.012	0.012	0.0%	0.004	0.043	0.013	0.012	54.8%	0.001
	$\beta_{21}(0.2)$	-0.130	0.037	0.037	89.6%	0.002	-0.262	0.042	0.042	77.8%	0.005	-0.047	0.039	0.038	92.8%	0.002
	$\beta_{02}(1.0)$	0.063	0.025	0.026	35.2%	0.005	0.116	0.028	0.028	1.2%	0.014	0.023	0.027	0.027	86.4%	0.001
	$\beta_{12}(0.2)$	-0.196	0.010	0.009	2.0%	0.002	-0.448	0.011	0.011	0.0%	0.008	0.0325	0.014	0.012	88.6%	0.000
	$\beta_{22}(0.5)$	-0.062	0.036	0.036	85.4%	0.002	-0.139	0.041	0.040	57.4%	0.006	-0.030	0.038	0.038	94.0%	0.002
	$\alpha_1(1.0)$	-0.120	0.060	0.055	41.4%	0.018	-0.140	0.058	0.055	31.2%	0.023	0.049	0.072	0.064	85.8%	0.008
	$\alpha_2(1.5)$	-0.127	0.066	0.063	17.2%	0.041	-0.152	0.065	0.064	7.4%	0.056	-0.003	0.075	0.070	93.6%	0.006

Table 2.6: Simulation results for comparing the two-stage approach with the joint generalized estimating equation (JGEE) approach (0% censoring).

		Two-Stage					GEE					JGEE				
Sample Size	True Parm	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE
$n = 200$	$\beta_{01}(0.2)$	0.257	0.061	0.059	85.2%	0.006	0.470	0.067	0.064	67.0%	0.013	0.059	0.062	0.060	94.2%	0.004
	$\beta_{11}(0.5)$	-0.055	0.022	0.021	71.8%	0.001	-0.126	0.025	0.025	26.6%	0.005	0.048	0.027	0.026	83.6%	0.001
	$\beta_{21}(0.2)$	-0.144	0.080	0.082	94.8%	0.007	-0.308	0.095	0.094	88.2%	0.013	-0.064	0.083	0.085	94.6%	0.007
	$\beta_{02}(1.0)$	0.067	0.057	0.059	80.2%	0.008	0.121	0.061	0.063	50.0%	0.018	0.024	0.061	0.061	93.4%	0.004
	$\beta_{12}(0.2)$	-0.210	0.022	0.020	44.0%	0.002	-0.466	0.022	0.022	1.2%	0.009	-0.022	0.029	0.025	89.8%	0.001
	$\beta_{22}(0.5)$	-0.063	0.080	0.081	93.8%	0.007	-0.136	0.087	0.088	88.0%	0.012	-0.025	0.084	0.084	95.0%	0.007
	$\alpha_1(1.0)$	-0.124	0.121	0.119	80.2%	0.030	-0.142	0.119	0.120	75.6%	0.034	0.052	0.146	0.138	91.6%	0.024
	$\alpha_2(1.5)$	-0.117	0.135	0.138	72.8%	0.049	-0.142	0.136	0.140	65.2%	0.064	0.009	0.158	0.154	94.8%	0.025
$n = 500$	$\beta_{01}(0.2)$	0.269	0.038	0.037	68.2%	0.004	0.473	0.042	0.041	35.6%	0.011	0.061	0.039	0.038	93.2%	0.002
	$\beta_{11}(0.5)$	-0.056	0.015	0.013	45.2%	0.001	-0.130	0.016	0.016	1.8%	0.004	0.047	0.018	0.016	67.4%	0.001
	$\beta_{21}(0.2)$	-0.147	0.053	0.052	89.8%	0.004	-0.290	0.06	0.060	84.0%	0.007	-0.051	0.055	0.054	94.4%	0.003
	$\beta_{02}(1.0)$	0.069	0.037	0.037	55.2%	0.006	0.125	0.040	0.040	11.6%	0.017	0.026	0.038	0.038	90.8%	0.002
	$\beta_{12}(0.2)$	-0.206	0.013	0.012	11.0%	0.002	-0.462	0.015	0.014	0.0%	0.009	0.031	0.018	0.016	89.4%	0.000
	$\beta_{22}(0.5)$	-0.070	0.053	0.051	87.8%	0.004	-0.153	0.059	0.056	70.6%	0.009	-0.034	0.055	0.053	92.8%	0.003
	$\alpha_1(1.0)$	-0.120	0.081	0.073	59.6%	0.021	-0.140	0.079	0.074	50.0%	0.026	0.053	0.097	0.085	86.8%	0.012
	$\alpha_2(1.5)$	-0.128	0.087	0.085	39.8%	0.044	-0.152	0.087	0.086	25.2%	0.060	-0.000	0.101	0.095	93.4%	0.010
$n = 1000$	$\beta_{01}(0.2)$	0.265	0.026	0.026	50.0%	0.003	0.468	0.029	0.029	10.6%	0.010	0.061	0.027	0.027	91.2%	0.001
	$\beta_{11}(0.5)$	-0.058	0.010	0.009	14.0%	0.001	-0.131	0.012	0.011	0.0%	0.004	0.045	0.013	0.012	48.0%	0.001
	$\beta_{21}(0.2)$	-0.132	0.037	0.037	89.2%	0.002	-0.276	0.042	0.042	75.6%	0.005	-0.043	0.039	0.038	93.0%	0.002
	$\beta_{02}(1.0)$	0.065	0.025	0.026	30.8%	0.005	0.121	0.028	0.028	0.4%	0.016	0.022	0.027	0.027	87.0%	0.001
	$\beta_{12}(0.2)$	-0.202	0.010	0.009	1.0%	0.002	-0.459	0.010	0.010	0.0%	0.009	0.035	0.013	0.011	87.4%	0.000
	$\beta_{22}(0.5)$	-0.064	0.036	0.036	84.2%	0.002	-0.146	0.040	0.040	54.0%	0.007	-0.029	0.038	0.038	94.4%	0.002
	$\alpha_1(1.0)$	-0.127	0.057	0.051	33.2%	0.019	-0.147	0.055	0.052	22.2%	0.025	0.045	0.068	0.060	84.6%	0.007
	$\alpha_2(1.5)$	-0.134	0.063	0.059	10.0%	0.044	-0.158	0.063	0.060	4.4%	0.060	-0.007	0.073	0.066	93.2%	0.005

2.5 Data Application

In this section, the proposed method is applied to data from a longitudinal cohort to examine the association of longitudinally collected low density lipoprotein (LDL), high density lipoprotein (HDL) measures and their association with the risk of coronary artery disease (CAD).

It has been reported that high LDL, low HDL cholesterol levels are associated with increased risk of coronary artery disease (CAD) (Wilson, 1990; Mertens and Holvoet, 2001). However, many observational studies have questioned the relevance of lipids in relation to cardiovascular disease risk (Weverling-Rijnsburger et al., 1997; Schatz et al., 2001). There are very few studies on LDL, especially HDL cholesterol levels in relation to risk of CAD in old adults (Weverling-Rijnsburger et al., 2003). Existing studies on the association of lipids and risks of CAD have typically utilized either baseline measurements or mean values from multiple measurements over time. And so far, no study have looked at the association between longitudinal LDL and HDL measures together and the risk of CAD.

2.5.1 The Longitudinal Cohort

The study population comes from the Indianapolis-Ibadan Dementia Project (IIDP). The IIDP is a 20-year study of dementia in elderly African Americans living in Indianapolis, Indiana, and elderly Africans living in Ibadan and Nigeria, all of whom were age 65 or older. The project recruitment was conducted in two phases: for the first recruitment, 2212 African Americans age 65 or older who live in Indianapolis were enrolled in 1992; for the second enrollment in 2001, additional 1893 community-dwelling

African American with age 70 years and older were recruited. Detailed description of the IIDP cohort can be found at Hendrie et al. (2001) and Hall et al. (2009).

This data application merged electronic medical records (EMRs) with data collected from the Indianapolis-Ibadan Dementia project (IIDP). Of the 4105 participants enrolled in IIDP, 3778 (92%) were identified in EMR using social security numbers, name, gender, and date of birth. For each individual, HDL and LDL measures associated with outpatient visits were retrieved from EMR data. The event of CAD was defined as the diagnosis of CAD using ICD-9 codes or cause of death by CAD. For patients with an event, age at CAD diagnosis is used as the event time; for subjects without CAD diagnosis, age at last evaluation time in IIDP or age of death other than CAD is used as the right censoring point.

The current work focuses on a subset of 979 female patients, with a total number of 3728 observations. During the study follow-up, 343(35.0%) of the subjects were diagnosed with CAD by the end of study period. The average number of longitudinal lipid measurements is 3.81(sd 4.00). The average length of follow up is 2.21 years (range 0 to 21.67). Demographic characteristics of the study population including baseline age, year of education, baseline smoking status and diabetes condition are summarized in Table 2.7. CAD and non-CAD patients are not different in baseline age, education years, baseline smoking status and baseline LDL levels, but they are different in percent with diabetes and mean baseline HDL levels.

Figure 2.1 and 2.2 plot annualized longitudinal LDL and HDL measures over time. The blue dots are observed lipids for Non-CAD group, and red dots are observed lipids for CAD group. The blue and red lines represent fitted population mean profiles for Non-CAD and CAD group respectively. From Figure 2.1, population mean LDL

over time for patients in CAD and Non-CAD group are both decreasing. From Figure 2.2, population mean HDL over time for patients in Non-CAD group stays flat, while population mean HDL decreases over time for patients in CAD group. Figure 2.1 and 2.2 indicates potential association between the risk of CAD and longitudinal LDL and HDL over time.

Table 2.7: Comparisons of baseline demographic of female subjects between CAD and non-CAD groups ($n = 979$)

Baseline Characters	All (n=979)	CAD (n=343)	Non-CAD (n=636)	p-value
Age at baseline, mean(sd)	77.02(6.71)	76.53(6.54)	77.28(6.80)	0.096
Year of Education, mean(sd)	11.09(2.52)	11.15(2.57)	10.98(2.41)	0.327
Smoking, n (%)	479(48.93%)	182(53.06%)	297(46.70%)	0.057
Diabetes, n (%)	276(28.19%)	118(34.40%)	158(24.84%)	0.002
Baseline LDL, mean(sd)	132.48(42.28)	134.56(43.35)	131.35(41.68)	0.258
Baseline HDL, mean(sd)	58.79(16.70)	56.34(15.39)	60.11(17.23)	0.001
Number of measures, mean(sd)	3.81(4.00)	3.28(3.20)	4.09(4.34)	0.003
Years of follow up, median(range)	2.21(0, 21.67)	1.48(0, 17.52)	2.80(0, 21.67)	<.0001

Figure 2.1: Observed annualized longitudinal LDL measures over time and fitted population mean curves for the CAD and Non-CAD group.

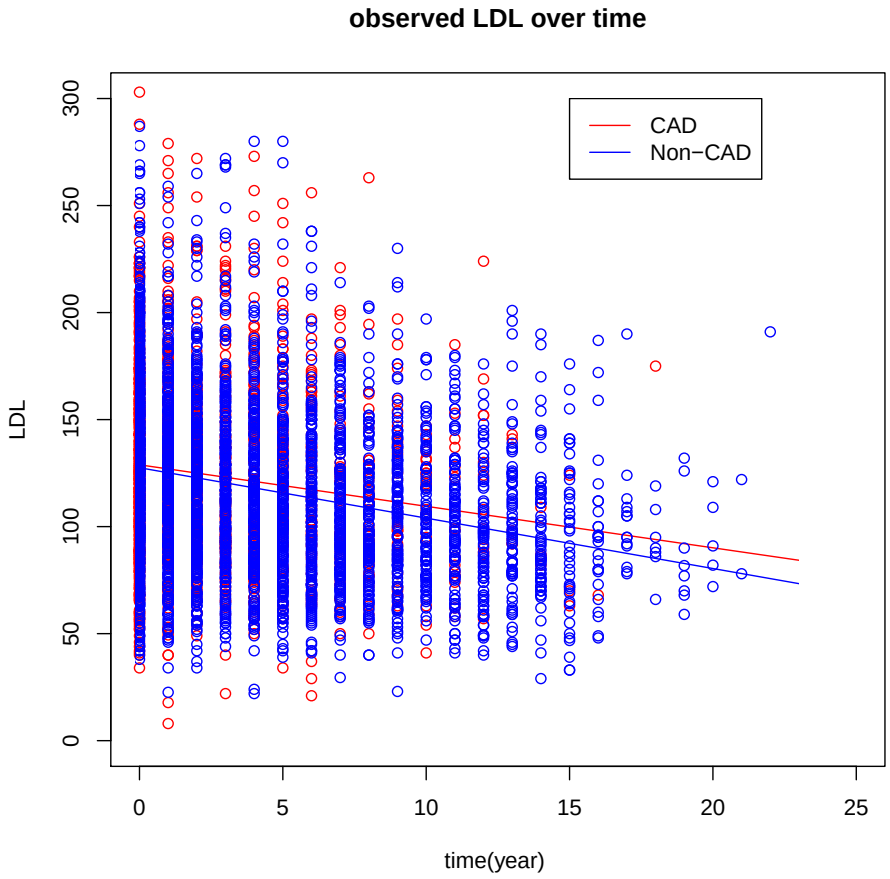
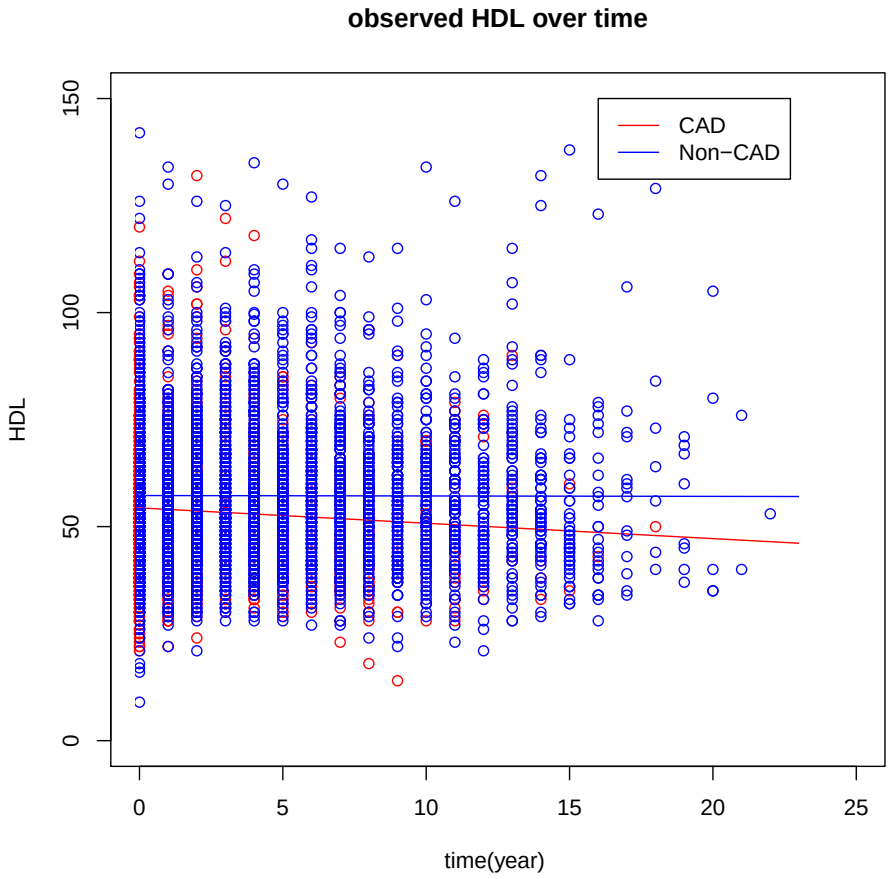


Figure 2.2: Observed annualized longitudinal HDL measures over time and fitted population mean curves for the CAD and Non-CAD group.



2.5.2 Application to IIDP Data

The proposed JGEE method is applied to the longitudinal cohort with 979 female patients. Four different sets of joint models using the proposed JGEE method are fitted. The best set of models are determined using the AIC. The four sets of joint models are outlined as follows:

Joint model 1: Consider the following longitudinal model:

$$\begin{aligned} y_l(t_{ij}) &= y_l^*(t_{ij}) + \epsilon_{ijl}(t_{ij}) \\ &= \beta_{0l} + \beta_{1l}t_{ij} + \beta_{2l}\text{grade}_i + \beta_{3l}\text{baseage}_i + b_{0li} + b_{1li}t_{ij} + \epsilon_{ijl}(t_{ij}), \end{aligned}$$

where y_l, $l = 1, 2$ denote the observed longitudinal LDL and HDL measures, y_l^* denote the true unobserved longitudinal outcomes. Grade denotes year of education, and baseage denotes age at baseline. $\mathbf{b}_i = (b_{01i}, b_{11i}, b_{02i}, b_{12i})^T$ is random effect with mean zero and variance covariance matrix $\mathbf{G}$. ϵ_{ij1} and ϵ_{ij2} are independent and normally distributed error term with mean 0 and variance parameters σ_1^2 and σ_2^2. ϵ_{ij1} and ϵ_{ij2} are independent of $\mathbf{b}_i$.

The survival model is assumed to follow the Cox PH model with the form:

$$h(t) = h_0(t)\exp\{\gamma_1\text{baseage}_i + \gamma_2\text{somke}_i + \gamma_3\text{diabetes}_i + \alpha_1 y_{i1}^*(t) + \alpha_2 y_{i2}^*(t)\}.$$

This hazard function assumes that the hazard of CAD depends on baseline age, smoking status, history of diabetes and the current value of LDL and HDL measures on the log scale.

Joint model 2: The longitudinal model:

$$\begin{aligned} y_l(t_{ij}) &= y_l^*(t_{ij}) + \epsilon_{ijl}(t_{ij}) \\ &= \beta_{0l} + \beta_{1l}t_{ij} + \beta_{2l}\text{grade}_i + \beta_{3l}\text{baseage}_i + b_{0li} + b_{1li}t_{ij} + \epsilon_{ijl}(t_{ij}). \end{aligned}$$

The survival model:

$$\begin{aligned} h(t) = h_0(t)\exp\{&\gamma_1\text{baseage}_i + \gamma_2\text{somke}_i + \gamma_3\text{diabetes}_i \\ &+ \alpha_1(\beta_{11} + b_{11i}) + \alpha_2(\beta_{12} + b_{12i})\}. \end{aligned}$$

Joint model 2 is the same as joint model 1 for longitudinal sub-models, while joint model 2 assumes the hazard function depends on the slope of LDL and HDL measures over time.

Joint model 3: The longitudinal model:

$$\begin{aligned} y_l(t_{ij}) &= y_l^*(t_{ij}) + \epsilon_{ijl}(t_{ij}) \\ &= \beta_{0l} + \beta_{1l}t_{ij} + \underline{\beta_{2l}t_{ij}^2} + \beta_{3l}\text{grade}_i + \beta_{4l}\text{baseage}_i + b_{0li} + b_{1li}t_{ij} + \epsilon_{ijl}(t_{ij}). \end{aligned}$$

The survival model:

$$h(t) = h_0(t)\exp\{\gamma_1\text{baseage}_i + \gamma_2\text{somke}_i + \gamma_3\text{diabetes}_i + \alpha_1 y_{i1}^*(t) + \alpha_2 y_{i2}^*(t)\}.$$

Joint model 3 is the same as joint model 1 in the survival sub-model, while the longitudinal linear mixed effect models are assumed to depend on a quadratic fixed time effect in joint model 3.

Joint model 4: The longitudinal model:

$$y_l(t_{ij}) = y_l^*(t_{ij}) + \epsilon_{ijl}(t_{ij})$$
$$= \beta_{0l} + \beta_{1l}t_{ij} + \underline{\beta_{2l}t_{ij}^2} + \beta_{3l}\text{grade}_i + \beta_{4l}\text{baseage}_i + b_{0li} + b_{1li}t_{ij} + \epsilon_{ijl}(t_{ij}).$$

The survival model:

$$h(t) = h_0(t)\exp\{\gamma_1\text{baseage}_i + \gamma_2\text{somke}_i + \gamma_3\text{diabetes}_i$$
$$+ \alpha_1(\beta_{1l} + 2\beta_{21}t_{ij}) + \alpha_2(\beta_{12} + 2\beta_{22}t_{ij})\}$$

Joint model 4 is the same as joint model 3 for longitudinal sub-model, while that joint model 4 assumes the hazard function depends on the slope of LDL and HDL measures over time.

Estimated parameters applying proposed JGEE method for the four sets of joint models are presented in Tables 2.8, 2.9, 2.10 and 2.11 respectively. Models are compared according to the AIC on survival models. Joint model 1 is the best with smallest AIC(AIC=3867.824) followed by the joint model 3(AIC=3871.145), the joint model 4 (AIC=3886.694) and the joint model 2(AIC=3895.248). As a result, we focus on the joint model 1 for associating longitudinal LDL and HDL on the risk of CAD.

Results from the joint model 1 are presented in Table 2.8. It can be seen that on population average, both LDL and HDL are decreasing over time. It can be observed that LDL is not statistically significantly associated with the risk of CAD development, while lower HDL measures are significantly associated with the risk of developing CAD. With one unit increase of the current value of HDL, the hazard of developing CAD decrease by 4%($e^{\alpha_1} = 0.96$). The fitted model also identified several

other risk factors for CAD, participants with older age, smoking and with diabetes have a higher risk of CAD.

Four different methods are compared for estimated associations between the longitudinal LDL, HDL and CAD: Cox PH model using observed baseline LDL and HDL measures as time-independent covariates; Cox PH model using observed time-dependent LDL and HDL measures; joint model 1 with two-stage approach; joint model 1 with proposed JGEE method. Estimated association parameters α_1 (LDL) and α_2 (HDL) of four different methods are presented in Table 2.12 : Cox PH model with baseline LDL and HDL measures as time independent covariates, along with baseline covariates baseline age, smoking indicator and diabetes indicator; Cox PH model considers longitudinal LDL and HDL as time dependent covariates, adjusting for baseline age, smoking indicator and diabetes indicator; the naive two-stage approach with the joint model 1; the proposed JGEE method for the joint model 1. Figure 2.3 and 2.4 plots the estimated association parameters of LDL and HDL and the corresponding 95% CI from the four different methods respectively. All methods demonstrate that LDL is not statistically significantly associated with the risk of CAD development. It can be observed that lower HDL measures are significantly associated with the risk of developing CAD. From Cox PH with baseline, to Cox PH with time-dependent covariates, to joint models with two-stage, to joint models with JGEE, the effect of HDL on CAD risk getting larger, with the proposed method has the largest effect estimator for the association between HDL and risk of CAD.

Cox PH model uses only baseline measures or considers longitudinal measures as time-dependent covariates are commonly used approaches in practice. As longitudinal measures are collected intermittently with measurement errors at a few time

points for each subject; use only baseline measures can not catch possible survival risk effects; use longitudinal measures as time-dependent covariates fail to take measurement error into account and those measures right ahead of the event make a larger contribution to survival parameter estimation. Joint modeling framework can take into account the longitudinal measurement error and link true longitudinal covariates to the risk effect. Naive two-stage method tends to underestimate the risk effects refer to simulation studies. The proposed method adopts weighted estimating equation idea justifies missing caused by the time-to-event. This can be an explanation why the proposed approach obtains larger association parameters than the native two-stage method.

Table 2.8: Parameter estimates, standard errors and 95% CI using the Joint Model 1. α_1 and α_2 are association estimates between the risk of CAD and the current value of LDL, HDL at event time point, respectively.

Parameter	Estimate	StdErr	Lower 95% CI	Upper 95% CI
Longitudinal LDL				
Intercept	232.570	31.142	171.533	293.608
Time	-3.173	0.511	-4.173	-2.172
Year of Education	-0.810	0.723	-2.227	-0.607
Age at baseline	-1.275	0.354	-1.969	-0.580
Longitudinal HDL				
Intercept	47.994	12.703	23.097	72.891
Time	-0.608	0.635	-1.852	0.637
Year of Education	0.016	0.318	-0.607	0.640
Age at baseline	0.091	0.151	-0.205	0.386
Time to CAD				
Age at baseline	0.078	0.014	0.050	0.106
Smoke	0.298	0.109	0.085	0.512
Diabetes	0.341	0.114	0.118	0.564
α_1 (LDL)	-0.001	0.002	-0.005	0.003
α_2 (HDL)	-0.041	0.008	-0.057	-0.026

Table 2.9: Parameter estimates, standard errors and 95% CI using the Joint Model 2. α_1 and α_2 are association estimates between the risk of CAD and the slope of LDL, HDL at event time point, respectively.

Parameter	Estimate	StdErr	Lower 95% CI	Upper 95% CI
Longitudinal LDL				
Intercept	214.146	30.123	155.107	273.185
Time	-2.448	0.500	-3.427	-1.469
Year of Education	-0.729	0.692	-2.086	-0.627
Age at baseline	-1.066	0.345	-1.743	-0.390
Longitudinal HDL				
Intercept	52.799	13.677	25.993	79.605
Time	0.137	0.280	-0.412	0.685
Year of Education	-0.172	0.332	-0.822	0.479
Age at baseline	0.077	0.165	-0.245	0.400
Time to CAD				
Age at baseline	0.066	0.014	0.038	0.094
Smoke	0.283	0.109	0.070	0.495
Diabetes	0.356	0.114	0.133	0.579
α_1 (LDL)	0.024	0.026	-0.028	0.075
α_2 (HDL)	-1.212	0.488	-2.168	-0.255

Table 2.10: Parameter estimates, standard errors and 95% CI using the Joint Model 3. α_1 and α_2 are association estimates between the risk of CAD and the current value of LDL, HDL at event time point, respectively.

Parameter	Estimate	StdErr	Lower 95% CI	Upper 95% CI
Longitudinal LDL				
Intercept	237.395	32.468	173.758	301.032
Time	-4.413	1.144	-6.655	-2.171
Time^2	0.128	0.072	-0.013	0.269
Year of Education	-0.812	0.721	-2.226	0.602
Age at baseline	-1.326	0.369	-2.049	-0.603
Longitudinal HDL				
Intercept	44.647	13.126	18.921	70.373
Time	0.312	0.521	-0.709	1.333
Time^2	-0.080	0.045	-0.168	0.008
Year of Education	-0.007	0.318	-0.630	0.615
Age at baseline	0.132	0.155	-0.171	0.435
Time to CAD				
Age at baseline	0.075	0.014	0.048	0.103
Smoke	0.300	0.109	0.088	0.513
Diabetes	0.346	0.113	0.123	0.568
α_1 (LDL)	-0.001	0.002	-0.005	0.003
α_2 (HDL)	-0.039	0.008	-0.055	-0.024

Table 2.11: Parameter estimates, standard errors and 95% CI using the Joint Model 4. α_1 and α_2 are association estimates between the risk of CAD and the slope of LDL, HDL at event time point, respectively.

Parameter	Estimate	StdErr	Lower 95% CI	Upper 95% CI
Longitudinal LDL				
Intercept	215.883	31.787	153.582	278.184
Time	-2.583	0.893	-4.332	-0.833
Time^2	0.061	0.055	-0.047	0.169
Year of Education	-0.564	0.685	-1.906	0.779
Age at baseline	-1.155	0.361	-1.863	-0.447
Longitudinal HDL				
Intercept	53.519	14.455	25.188	81.850
Time	0.205	0.474	-0.724	1.134
Time^2	-0.029	0.028	-0.084	0.026
Year of Education	-0.216	0.326	-0.855	0.422
Age at baseline	0.069	0.171	-0.267	0.405
Time to CAD				
Age at baseline	0.094	0.015	0.064	0.124
Smoke	0.280	0.109	0.067	0.493
Diabetes	0.345	0.109	0.121	0.569
α_1 (LDL)	0.008	0.114	-0.029	0.045
α_2 (HDL)	-0.778	0.196	-1.163	-0.393

Table 2.12: Association parameter estimates, standard errors, hazard ratio (HR) and 95% confidence interval of hazard ratio (HR) using four different methods: Cox PH model with baseline LDL, HDL as time-independent covariates; Cox PH model with observed LDL and HDL as time dependent covariates; the two-stage method for the Joint model 1; the JGEE method for the joint model 1

Parameter	Estimate	StdErr	HR for CAD (95% CI)
Cox PH baseline			
α_1 (LDL)	-0.001	0.001	0.999 (0.996, 1.002)
α_2 (HDL)	-0.011	0.004	0.989 (0.982, 0.996)
Cox PH time-dependent			
α_1 (LDL)	-0.001	0.001	0.999 (0.996, 1.002)
α_2 (HDL)	-0.014	0.004	0.986 (0.978, 0.993)
Joint models two-stage			
α_1 (LDL)	-0.002	0.002	0.998 (0.994, 1.002)
α_2 (HDL)	-0.028	0.007	0.972 (0.958, 0.987)
Joint models JGEE			
α_1 (LDL)	-0.001	0.002	0.999 (0.995, 1.003)
α_2 (HDL)	-0.041	0.008	0.959 (0.945, 0.974)

Figure 2.3: Comparison of estimated associations between the longitudinal LDL and CAD from four methods, the blue solid dots are estimated $\hat{\alpha}_1$ from four methods. The upper and lover bars are 95% confidence interval of parameter estimates.

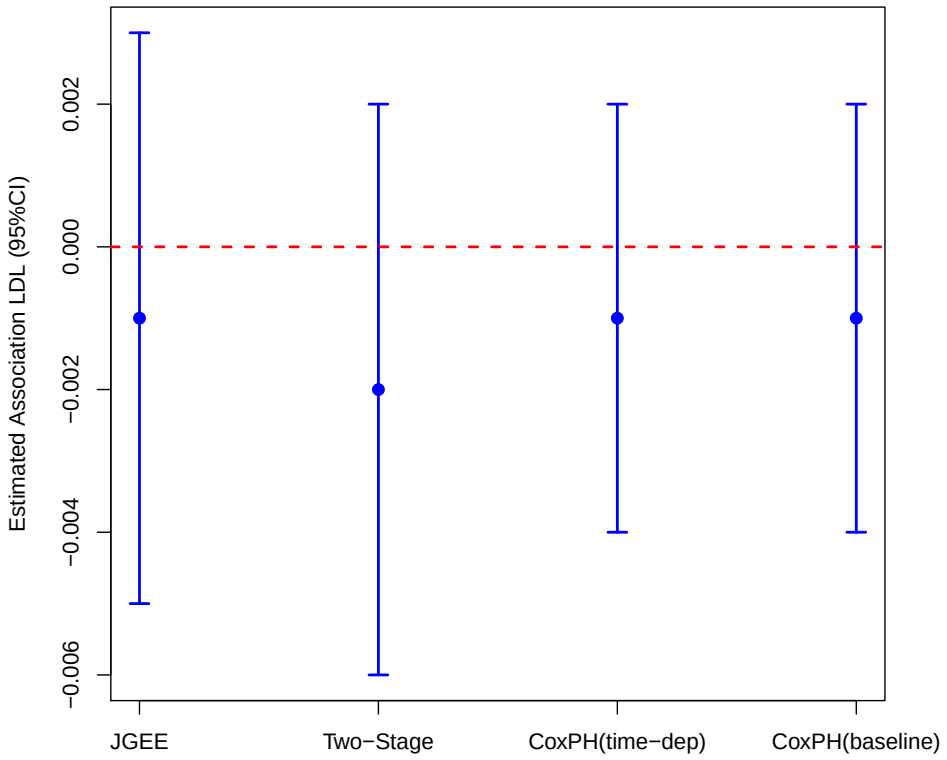
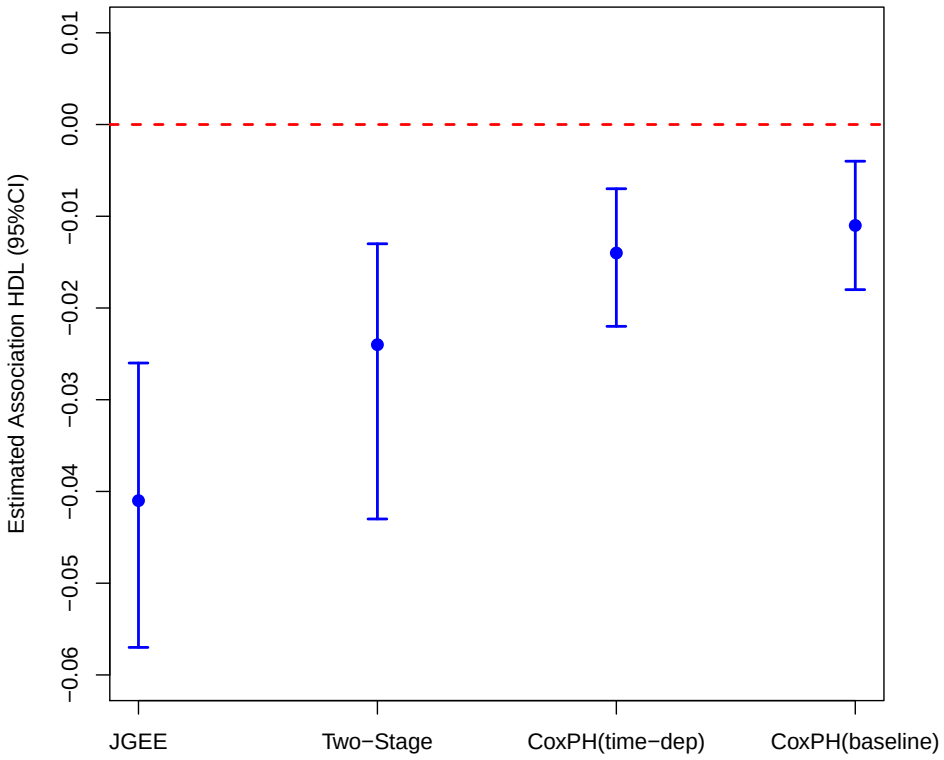


Figure 2.4: Comparison of estimated associations between the longitudinal HDL and CAD from four methods, the blue solid dots are estimated $\hat{\alpha}_2$ from four methods. The upper and lover bars are 95% confidence interval of parameter estimates.



2.6 Discussion

In this work, joint generalized estimating equation approach (JGEE) has been proposed for parameter estimation in joint modeling of multiple longitudinal measures and a time-to-event outcome. The performance of the proposed approach have been evaluated through extensive simulation studies. We have illustrated the proposed method using data from a longitudinal cohort study.

The JGEE approach does not require complete specification of the joint distribution of the longitudinal measures and survival outcomes, but only the first two moments of the longitudinal outcomes. Thus it can be easily extended to non-normal longitudinal, or mixtures of different types of longitudinal processes. In comparison, maximum likelihood with EM algorithm and Bayesian approach can be computationally extremely intensive because of the integration over multiple random effects, while the proposed JGEE method is computationally fast and feasible, and can be implemented in standard software.

There are a few limitations of the proposed methodology. On one hand, JGEE method loses the features of limiting properties likelihood estimation possesses such as straight consistency and asymptotic normality. Theoretical justification is needed for JGEE method on large sample properties. On the other hand, JGEE method first obtains the fixed effect estimators. During the process, subject specific random effects have been predicted and added back into the longitudinal outcome estimation. A correction has been provided for predicting random effects borrowing regression calibration method from measurement error problem. However, regression calibration methods work only for additive measurement error situation. In our case, measure-

ment error in the predicted random effects is not necessary additive or normally distributed. Thus the random effects is still not perfectly predicted, and the estimators are not completely unbiased. Moreover, by adding a weight into GEE function, it can not fully accommodate the non-ignorable missing caused by the occurrence of an event.

Joint modeling of multiple longitudinal outcomes and time-to-event data are applicable to many clinical trials and observational studies, when the interest lies on an association between longitudinal measures and the risk of an event. It may be more clinical meaningful and accurate to predict patients' risk effect based on all available longitudinal biomarker history. Multivariate longitudinal responses are not necessarily normal, they could belong to different response families, which is common in medical science. Thus it is worthwhile to study general multivariate longitudinal processes and time-to-event outcomes.

Chapter 3

Joint Models for Multiple Longitudinal Poisson Outcomes and Time-to-Event Data

3.1 Introduction

In a previous chapter we introduced the framework of joint models for multiple continuous longitudinal outcomes and time-to-event data, and we showed how models of this type can be solved with generalized estimating equation approach. In that work, we focused on multiple continuous longitudinal outcomes and one survival endpoint. The majority of the research in joint models has focused on continuous longitudinal responses. However, longitudinal outcomes may not be continuous in some biomedical studies. For example, increases in the number of skin tumors each week can have an influence on the time of death; the longitudinal outcome could be categorical such as test results positive or negative that are relates to the onset of an event.

Extensions of joint models to deal with different types of longitudinal and survival data are limited in literature. Huang et al. (2001) considered bivariate binary longitudinal processes and bivariate event times with Expectation-Maximization (EM) algorithm for estimation. Wang et al. (2002) accomplished Markov Chain Monte Carlo methods for analysing data from ordinal multiple longitudinal outcomes when death is present. Dunson and Herring (2005) considered Poisson model for univaraite longitudinal tumor counts response and gamma frailty models for survival following a Bayesian approach to inference. Rizopoulos and Ghosh (2011) proposed a Bayesian

flexible multivariate joint model that relates different types of longitudinal outcomes to a time-to-event. Luo (2014) proposed a Bayesian approach to joint analysis of multivariate longitudinal data and parametric accelerated failure time for survival. Choi et al. (2015) considered simultaneously modeling the survival time with single longitudinal categorical responses with Expectation-Maximization (EM) algorithm.

The aim of this work is to present one of the extensions, namely joint models with multiple count longitudinal responses and one time-to-event outcome, and how parameters in such models can be estimated using joint generalized estimating equation(JGEE) approach. This structure results in a log-log of expected counts to the risk of events. Multivariate longitudinal Poisson data will be modeled by generalized linear mixed models (GLMMs). And Cox's Proportional Hazards (PH) semi-parametric model will be used to capture the time-to-event process. GLMMs are an extension of linear mixed models to allow response variables from different distributions, such as binary or count responses. In contrast to linear mixed models, fitting GLMMs under maximum likelihood is computationally much more challenging. In a linear mixed model, the log likelihood is well defined, while in GLMM, it is possible that no valid joint distribution can be constructed; numerical integration is much more complicated with the number of random effects increase.

The remainder of this chapter is organized as follows. In section 3.1, we introduce the joint models for analyzing the association between multiple longitudinal counts measures and a time-to-event outcome. We define the notation and model formulation of longitudinal and survival processes. In section 3.2, we introduce the proposed estimation methods. Simulation studies and results are summarized in

section 3.3. The proposed method is applied to data from a longitudinal cohort in section 3.4.

3.2 Model Formulation

The joint models for time-to-event data and multivariate longitudinal count data comprise two submodels: a multivariate longitudinal data model and a time-to-event data model depend on the true longitudinal process.

3.2.1 Longitudinal Models

Following the same notation as in Section 2.2, let $y_{il}(t_{ij})$ denote the j-th observed value of the l-th longitudinal outcome for subject i, measured at time t_{ij}, where $i = 1, \cdots, N$; $l = 1, \cdots, L$ and $j = 1, \cdots, n_i$. $\mathbf{Y}_i = (\mathbf{y}_{i1}^T, \cdots, \mathbf{y}_{iL}^T)^T$ denotes L-variate response vector for the i-th subject, $\mathbf{y}_{il} = (y_{il}(t_{i1}), \cdots, y_{il}(t_{in_i}))^T$ is a vector of longitudinal response at different time points. The set of covariates for the l-th longitudinal outcome are denoted as $\mathbf{X}_{il} = (\mathbf{X}_{il}^T(t_{i1}), \cdots, \mathbf{X}_{il}^T(t_{in_i}))^T$, where $\mathbf{X}_{il}^T(t_{ij}) = (1, x_{il1}(t_{ij}), \cdots, x_{ilp_l}(t_{ij}))^T$ is a vector of covariates for the l-th response of subject i at time t_{ij}. The longitudinal data are modeled by the following generalized linear mixed effect model:

$$y_{il}(t_{ij})|\mathbf{b}_{il} \sim \text{Poisson}(E(y_{il}(t_{ij})|\mathbf{b}_{il})),$$

$$\log(E(y_{il}(t_{ij})|\mathbf{b}_{il})) = m_{il}(t_{ij}) = \mathbf{X}_{il}^T(t_{ij})\boldsymbol{\beta}_l + \mathbf{Z}_{il}^T(t_{ij})\mathbf{b}_{il}, \tag{3.1}$$

where $\mathbf{X}_{il}^T(t_{ij})$ and $\mathbf{Z}_{il}^T(t_{ij})$ are row vectors of covariates for subject i associated with fixed and random effects respectively; $\boldsymbol{\beta}_l$ is a vector of fixed effects parameters for the l-th outcome; $\mathbf{b}_{il}$ is a vector of subject specific random effects for the lth outcome.

For multiple longitudinal markers, the responses for the i-th subject can be modeled by the following generalized mixed effects model:

$$\log(E(\mathbf{Y}_i|\mathbf{b}_i)) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \tag{3.2}$$

where $\mathbf{X}_i = \begin{pmatrix} \mathbf{X}_{i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{i2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{X}_{iL} \end{pmatrix}_{(\sum_{j=1}^{L}(p_j+1))\times(\sum_{j=1}^{L}(p_j+1))}$,

$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_L \end{pmatrix}_{(\sum_{j=1}^{L}(p_j+1))\times 1}$, $\mathbf{Z}_i = \begin{pmatrix} \mathbf{Z}_{i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{i2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{Z}_{iL} \end{pmatrix}$,

$\mathbf{b}_i = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_L \end{pmatrix}$ with covariane matrix $\mathbf{G} = \begin{pmatrix} \mathbf{G}_1 & \mathbf{G}_{12} & \cdots & \mathbf{G}_{1L} \\ \mathbf{G}_{12} & \mathbf{G}_2 & \cdots & \mathbf{G}_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}_{1L} & \cdots & \mathbf{G}_{(L-1)L} & \mathbf{G}_L \end{pmatrix}$.

Thus $E(y_{il}|\mathbf{b}_{il}) = Var(y_{il}|\mathbf{b}_{il}) = e^{\mathbf{X}_{il}^T\boldsymbol{\beta}_l+\mathbf{Z}_{il}^T\mathbf{b}_{il}}$ and $E(\mathbf{Y}_i|\mathbf{b}_i) = Var(\mathbf{Y}_i|\mathbf{b}_i) = e^{\mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_i\mathbf{b}_i}$.

The Marginal Model

The mean value, variance and covariance of the $\mathbf{Y}$ marginals can be computed. For mean of y_{il}:

$$\mu_{il} = E(y_{il}) = E_{\mathbf{b}_{il}}(E(y_{il}|\mathbf{b}_{il})) = E(e^{\mathbf{X}_{il}^T\boldsymbol{\beta}_l+\mathbf{Z}_{il}^T\mathbf{b}_{il}}) = e^{\mathbf{X}_{il}^T\boldsymbol{\beta}_l}E(e^{\mathbf{Z}_{il}^T\mathbf{b}_{il}}).$$

Then

$$\log(E(y_{il})) = \mathbf{X}_{il}^T\boldsymbol{\beta}_l + \log\left(E(e^{\mathbf{Z}_{il}^T\mathbf{b}_{il}})\right).$$

Thus in the marginal model, $\log\left(E(e^{\mathbf{Z}_{il}^T\mathbf{b}_{il}})\right)$ can be considered as an offset variable. For variance of y_{il}:

$$\begin{aligned}
Var(y_{il}) &= Var\left[E(y_{il}|\mathbf{b}_{il})\right] + E\left[Var(y_{il}|\mathbf{b}_{il})\right] \\
&= Var\left[e^{\mathbf{X}_{il}^T\boldsymbol{\beta}_l+\mathbf{Z}_{il}^T\mathbf{b}_{il}}\right] + E\left[e^{\mathbf{X}_{il}^T\boldsymbol{\beta}_l+\mathbf{Z}_{il}^T\mathbf{b}_{il}}\right] \\
&= e^{\mathbf{X}_{il}^T\boldsymbol{\beta}_l}e^{\mathbf{X}_{il}^T\boldsymbol{\beta}_l}Var[e^{\mathbf{Z}_{il}^T\mathbf{b}_{il}}] + E(y_{il}) \\
&= E(y_{il})\left[e^{\mathbf{X}_{il}^T\boldsymbol{\beta}_l}\frac{Var[e^{\mathbf{Z}_{il}^T\mathbf{b}_{il}}]}{E(e^{\mathbf{Z}_{il}^T\mathbf{b}_{il}})} + 1\right].
\end{aligned}$$

Thus variance of y_{il} can be written as $Var(y_{il}) = \phi_{il}E(y_{il})$, where

$$\phi_{il} = e^{\mathbf{X}_{il}^T\boldsymbol{\beta}_l}\frac{Var[e^{\mathbf{Z}_{il}^T\mathbf{b}_{il}}]}{E(e^{\mathbf{Z}_{il}^T\mathbf{b}_{il}})} + 1.$$

For the covariance, with $l \neq k$,

$$\begin{aligned} Cov(y_{il}, y_{ik}) &= Cov\left[E(y_{il}|\mathbf{b}_{il}), E(y_{ik}|\mathbf{b}_{ik})\right] + E\left[Cov(y_{il}|\mathbf{b}_{il}, y_{ik}|\mathbf{b}_{ik})\right] \\ &= Cov\left[e^{\mathbf{X}_{il}^T\boldsymbol{\beta}_l + \mathbf{Z}_{il}^T\mathbf{b}_{il}}, e^{\mathbf{X}_{ik}^T\boldsymbol{\beta}_k + \mathbf{Z}_{ik}^T\mathbf{b}_{ik}}\right] \\ &= e^{\mathbf{X}_{il}^T\boldsymbol{\beta}_l + \mathbf{X}_{ik}^T\boldsymbol{\beta}_k} Cov(e^{\mathbf{Z}_{il}^T\mathbf{b}_{il}}, e^{\mathbf{Z}_{ik}^T\mathbf{b}_{ik}}). \end{aligned}$$

That is the correlation among different responses are introduced by correlation among random effects $\mathbf{b}_i$.

3.2.2 The Survival Model

For a time-to-event outcome, let T_i^* be the true event time for subject i. Some subjects may not experience the event at the end of study, or they may drop out during the study, so their event times are right censored. Let C_i denote the censoring time for the i-th subject. The observed event time $T_i = \min(T_i^*, C_i)$. Define the event indicator as $\delta_i = I(T_i^* \leq C_i)$, here $I(\cdot)$ represents indicator function. Assume that censoring times and event times are independent. Suppose time-to-event sub-model follows Cox's Proportional Hazards model. The hazard function for subject i at time t is given by:

$$h_i(t) = h_0(t) \exp\left\{\mathbf{u}_i^T\boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l\left(m_{il}(t)\right)\right\}, \tag{3.3}$$

where $h_0(t)$ denotes the baseline hazard function; $\mathbf{u}_i^T$ is a vector of baseline covariates; $\boldsymbol{\gamma}$ is a vector of fixed effects parameters; $m_{il}(t)$ is the longitudinal profile as defined in (3.1); function $f_l(\cdot)$ captures the association between the longitudinal measurements and the event process; and α_l denotes a parameter measuring the effect of the l-th longitudinal process on the time to event outcome. Parameters α_l's $(l = 1, 2, \cdots, L)$

are of primary interest in this work. Note that multiple longitudinal outcomes are assumed to have additive effects on the survival times.

3.3 Estimation Method

Different approaches can be used for parameter estimation for the joint models defined above, the two-stage method can result in biased results; maximum likelihood based approach and Bayesian method utilizes the full likelihood based on the joint models can be computationally intensive as the likelihood function or the posterior distribution function involves integration over multiple random effects. In this section, we present the joint generalized estimating equation methodology for parameter estimation of joint models for multiple longitudinal counts biomarkers and time-to-event outcome.

Let $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_{i1}^T, \cdots, \boldsymbol{\mu}_{iL}^T)^T$ be the mean vector corresponding to the vector of measurements $\mathbf{Y}_i$ on the i-th subject. Let $\mathbf{V}_i$ denote the variance covariance matrix of $\mathbf{Y}_i$. Let $\mathbf{W}_i$ be the diagonal weight matrix, $\mathbf{U}(\boldsymbol{\beta})$, $\mathbf{D}(\mathbf{b})$, $L(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ are objective functions for solving $\boldsymbol{\beta}$, $\mathbf{b}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ respectively. We propose a joint generalized estimating equation (JGEE) approach for estimation, which is to solve the following estimating equations:

Estimate $\boldsymbol{\beta}$ from:

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^T \mathbf{V}_i^{-1} \mathbf{W}_i (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}. \tag{3.4}$$

Predict random effects $\mathbf{b}$ from:

$$\mathbf{b} = \underset{\mathbf{b}}{\operatorname{argmin}} \sum_i \left\{ \mathbf{Y}_i \log \frac{\mathbf{Y}_i}{e^{\mathbf{X}_i^T\boldsymbol{\beta}+\mathbf{Z}_i^T\mathbf{b}_i}} - (\mathbf{Y}_i - e^{\mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_i^T\mathbf{b}_i}) \right\}. \tag{3.5}$$

Estimate risk coefficients $\boldsymbol{\alpha}$ from:

$$\boldsymbol{\alpha} = \underset{\boldsymbol{\alpha},\boldsymbol{\gamma}}{\operatorname{argmax}} \prod_{i=1}^{N} \left[\frac{e^{\mathbf{u}_i^T\boldsymbol{\gamma}+\sum_{l=1}^{L}\alpha_l f_l(m_{il}(T_i))}}{\sum_{j=1}^{N} I(T_j \geq T_i) e^{\mathbf{u}_j^T\boldsymbol{\gamma}+\sum_{l=1}^{L}\alpha_l f_l(m_{jl}(T_j))}} \right]^{\delta_i}. \tag{3.6}$$

Equation (3.4) is the weighted GEE for longitudinal outcomes. $\mathbf{W}_i$ is an $(L \times n_i) \times (L \times n_i)$ diagonal weight matrix with the j-th diagonal element as

$$W(t_{ij}; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = \frac{c(t_{ij}; \mathbf{u}_i, \boldsymbol{\gamma}_0)}{h_0(t_{ij}) \exp\left\{ \mathbf{u}_i^T\boldsymbol{\gamma} + \sum_{l=1}^{L}\alpha_l f_l(m_{il}(t_{ij}-)) \right\}},$$

where $m_{il}(t_{ij}-)$ denotes log of the mean of the past observations until time t_{ij}. Note that $\mathbf{W}$ can use $\mathbf{Y}$'s collected right up to time t_{ij}. This weight is called inverse visit intensity weight motivated by Lin et al. (2004) and Buzkova and Lumley (2007), and it is used to account for missing caused by time-to-event. $c(\cdot)$ is a pre-specified function of the time independent covriates $\mathbf{u}$ and has the potential to stabilize the influence of small values in the denominator. In addition, if we choose $c(\cdot)$ as $h_0(t_{ij}) \exp\left\{ \mathbf{u}_i^T\boldsymbol{\gamma}_0 \right\}$, when $y_i(t)$'s have no influence on the risk of event at time t, the weight becomes identity matrix reducing to the regular GEE for equation (3.4). Note that longitudinal measures at different time points will assign different weights, but multiple longitudinal outcomes will share the same weight at the same time point. The solution to the weighted GEEs (3.4) can be

found by applying the Fisher Scoring method (Nelder and Baker, 1972). The algorithm is to get initial estimates of $\boldsymbol{\beta}$, $\mathbf{W}_i$ and $\mathbf{V}_i$, update $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_{r+1} = \boldsymbol{\beta}_r + \left[\sum_{i=1}^{N}(e^{\mathbf{X}_i\boldsymbol{\beta}_r}\mathbf{X}_i)^T\mathbf{V}_i^{-1}\mathbf{W}_i e^{\mathbf{X}_i\boldsymbol{\beta}_r}\mathbf{X}_i\right]^{-1}\left[\sum_{i=1}^{N}(e^{\mathbf{X}_i\boldsymbol{\beta}_r}\mathbf{X}_i)^T\mathbf{V}_i^{-1}\mathbf{W}_i(\mathbf{Y}_i - e^{\mathbf{X}_i\boldsymbol{\beta}_r})\right]$ until converge.

Equation (3.5) is to minimize the deviance residuals between observation and the estimated population average, which provides a prediction of random effects. For subject i, let

$$\begin{aligned} D &= \sum_i \left\{ y_i \log \frac{y_i}{e^{X_i\hat{\beta}+Z_ib_i}} - \left(y_i - e^{X_i\hat{\beta}+Z_ib_i}\right)\right\} \\ &= \sum_i \left\{ y_i \log y_i - y_i(X_i\hat{\beta} + Z_ib_i) - \left(y_i - e^{X_i\hat{\beta}+Z_ib_i}\right)\right\}. \end{aligned} \tag{3.7}$$

For each subject at each time point, there is a random effect Z_ib_i associated with it. Let $Z_b = Z_ib_i$, consider Z_b as an unknown parameter, take derivative of equation (3.7) respect to Z_b:

$$\frac{\partial D}{\partial Z_b} = -y_i + e^{X_i\hat{\beta}}e^{Z_b} = 0. \tag{3.8}$$

Then there is explicit solution for (3.8): $Z_b = \log Y_i - X_i\hat{\beta}$. Thus $\mathbf{Zb} = \log\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. $\mathbf{b}$ can be predicted from the generalized least squares estimates with explicit form

$$\hat{\mathbf{b}} = \left(\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{V}^{-1}\left(\log\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right).$$

Equation (3.6) is the partial likelihood function for Cox Proportional Hazard model with the true unobserved counts replaced by the estimated rates from equation (3.4) and (3.5):

$$\hat{m}_{il}(t_{ij}) = \mathbf{X}_i^T(t_{ij})\hat{\boldsymbol{\beta}}_l + \mathbf{Z}_i^T(t_{ij})\hat{\mathbf{b}}_{il}^*,$$

where $\hat{\mathbf{b}}_{il}^{*}$ is an adjusted random effect based on the estimate from (3.5). The corresponding log partial likelihood function is :

$$l(\boldsymbol{\alpha},\boldsymbol{\gamma}) = \sum_{i=1}^{N} \delta_i \left[\mathbf{u}_i^T\boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f(\hat{m}_{il}(T_i)) \right. \\ \left. - \log \sum_{j=1}^{N} I(T_j \geq T_i) \exp\{\mathbf{u}_j^T\boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f(\hat{m}_{jl}(T_j))\} \right]. \tag{3.9}$$

Note that the weight $\mathbf{W}_i$'s in equation (3.4), and $m_{il}(T_i)$'s in equation (3.6) depend on the true parameters, which are unknown in reality. Equations (3.4),(3.5) and (3.6) can be solved with $\mathbf{W}_i$, $m_{il}(T_i)$ replaced by their estimators. Equation (3.4) can provide an estimator of fixed effects $\hat{\boldsymbol{\beta}}$; given $\hat{\boldsymbol{\beta}}$, equation (3.5) can provide a prediction of random effects; from equation (3.4) and (3.5), an estimator m can be obtained and $f_l(\hat{m}(t)) = f_l(\mathbf{X}_{il}^T(t)\hat{\boldsymbol{\beta}}_l + \mathbf{Z}_{il}^T(t)\hat{\mathbf{b}}_{il}^{*})$ can be plugged into equation (3.6). The three equations can be solved iteratively.

Note that the marginal mean $E(y_{il}) = e^{\mathbf{X}_{il}^T\boldsymbol{\beta}_l} E(e^{\mathbf{Z}_{il}^T\mathbf{b}_{il}})$. In the iteration algorithm, in equation (3.4), we can start without random effect to obtain $\hat{\boldsymbol{\beta}}$ with weighted GEE. But once we have one iteration, we should be able to use the predicted $E(e^{\mathbf{Z}_{il}^T\mathbf{b}_{il}})$ as an offset for the GEE at subsequent iterations to get better estimates $\hat{\boldsymbol{\beta}}$.

Following the sandwich variance covariance of GEEs, the robust sandwich variance covariance matrix of $\hat{\boldsymbol{\beta}}$ from weighted GEEs will be calculated by:

$$\begin{aligned}
&\left(\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T \mathbf{V}^{-1}\mathbf{W}\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^{-1} \\
&\left(\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T \mathbf{V}^{-1}\mathbf{W}(\mathbf{Y}-\boldsymbol{\mu}_i)(\mathbf{Y}-\boldsymbol{\mu}_i)^T\mathbf{W}\mathbf{V}^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right) \\
&\left(\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T \mathbf{V}^{-1}\mathbf{W}\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^{-1},
\end{aligned} \tag{3.10}$$

with $\mathbf{V}$, $\boldsymbol{\beta}$ and $\mathbf{W}$ are replaced by estimates.

The variance estimate of $\hat{\boldsymbol{\alpha}}$ will be obtained by inverse of the information matrix of the log partial likelihood:

$$V(\hat{\boldsymbol{\alpha}}) = -\left[\frac{\partial^2 l(\hat{\boldsymbol{\alpha}})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T}\right]^{-1} \tag{3.11}$$

The proposed approach uses an iterative computational approach and avoids the high dimensional integration required in likelihood or Bayesian methods. The proposed method can be implemented in standard statistical software. Because the standard errors of fixed effects in longitudinal model are calculated based on modified GEE as if the weights are known, and the standard errors of the risk coefficient estimators are obtained through partial likelihood assuming that all the true covariate values are known, estimated standard errors are likely to be underestimating the true variance. In the following section, we investigate the finite sample performance of the proposed method in simulation studies.

3.4 Simulation Study

In this section, we perform simulation studies to assess the performance of the proposed method. Simulation results are compared with those obtained using the two-stage method and the GEE approach. The GEE method is solving equations (3.4), (3.5) and (3.6) with $\mathbf{W}$ be identity matrix in (3.4). Data are simulated form joint models with two longitudinal Poisson outcomes and time-to-event outcome. Similar fixed and random effects structures are considered for two longitudinal outcomes, where the fixed effects include intercept, time, and one baseline binary covariate, and the random effects include random intercept and slope.

Suppose longitudinal outcomes follow the model:

$$y_1(t_{ij})|(b_{01}, b_{11}) \sim \text{Poisson}(\mu_1(t_{ij})),$$

$$\log\left\{\mu_1(t_{ij}))\right\} = m_1(t_{ij}) = \beta_{01} + \beta_{11}t_{ij} + \beta_{21}u_i + b_{01i} + b_{11i}t_{ij},$$

$$y_2(t_{ij})|(b_{02}, b_{12}) \sim \text{Poisson}(\mu_2(t_{ij})),$$

$$\log\left\{\mu_2(t_{ij})\right\} = m_2(t_{ij}) = \beta_{02} + \beta_{12}t_{ij} + \beta_{22}u_i + b_{02i} + b_{12i}t_{ij},$$

where

$$\begin{pmatrix} b_{01i} \\ b_{02i} \\ b_{11i} \\ b_{12i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{01}^2 & \rho_1\sigma_{01}\sigma_{11} & \rho_2\sigma_{01}\sigma_{02} & \rho_1\sigma_{01}\sigma_{12} \\ \rho_1\sigma_{01}\sigma_{11} & \sigma_{11}^2 & \rho_1\sigma_{11}\sigma_{02} & \rho_1\sigma_{11}\sigma_{12} \\ \rho_2\sigma_{01}\sigma_{02} & \rho_1\sigma_{11}\sigma_{02} & \sigma_{02}^2 & \rho_1\sigma_{02}\sigma_{12} \\ \rho_1\sigma_{01}\sigma_{12} & \rho_1\sigma_{11}\sigma_{12} & \rho_1\sigma_{02}\sigma_{12} & \sigma_{12}^2 \end{pmatrix}\right).$$

Time-to-event outcomes follow Cox Proportional Hazard (PH) Model with Weibull baseline function:

$$h(t) = abt^{b-1} \exp(\alpha_1 m_1(t) + \alpha_2 m_2(t)),$$

where a is the shape parameter and b is the scale parameter of the Weibull distribution. For each subject, a survival probability s_i at the true even time T_i^* iss simulated from uniform $(0, 1)$, and the true event time T_i^* is solved from the equation:

$$s_i - \exp\left\{ - \int_0^{T_i^*} abu^{b-1} \exp(\alpha_1 m_1(t) + \alpha_2 m_2(t)) du \right\} = 0.$$

Censoring times are simulated from another uniform distribution independently. The event indicator δ_i for the ith subject is determined by comparing true event time T_i^* and censoring time C_i, $\delta_i = I(T_i^* \leq C_i)$. The observed time $T_i = \min(T_i^*, C_i)$. Four censoring percentages and three different sample sizes are considered. The true parameters used in the simulations are summarized in Table 3.1 and Table 3.2.

In this simulation, 500 Monte Carlo samples are generated. Longitudinal data are simulated for 200, 500, 1000 subjects with 10 equally spaced bivariate longitudinal observations over 5 year period. Survival data are simulated for 40%, 30%, 10% and 0% of censoring. Longitudinal observations are right censored by time-to-event data. Simulation results are presented in Tables 3.3 to 3.6. Relative bias (defined as $((\hat{\theta} - \theta_{true})/\theta_{true})$, empirical standard errors (Emp. S.E.), model based standard errors (Model S.E.), coverage probability of the 95% confidence intervals (95% CI

C.P.) and Mean Square Error (MSE) (defined as $E\left[(\hat{\theta}-\theta)^2\right]$) based on 500 Monte Carlo samples are reported.

From simulation results, it can be observed that for a fixed survival censoring percentage with a given sample size, proposed JGEE method performs better than GEE and two-stage method with smaller relative bias. For a fixed survival censoring percentage and with sample size gets larger, standard error, 95% coverage probability and MSE get smaller as expected. We also observe that JGEE method tends to underestimate the standard error, this is because the standard errors of β's are calculated based on modified GEE as if the weights are known, and the standard errors of α's are obtained through partial likelihood assuming that all the true covariate values are known. It is also observed that the GEE method always performs worse than naive two-stage method, this is because for longitudinal data, missing caused by survival information is not missing completely at random, regular GEE leads to more biased results on longitudinal estimates, thus worse survival estimates. This also reflects the proposed JGEE method adding weight in the GEE method does adjust for missing data.

For a given sample size, it can be observed from simulation results that, with censoring percentage gets larger, proposed JGEE approach gets smaller relative bias and larger MSE. This is because in the longitudinal part, missingness is caused by either censoring or time-to-event; missing data come from censoring are non-informative, which do not affect parameter estimation; while missing caused by event is nonignorable missing, the less censoring, the more events have happened, thus more bias would be for longitudinal estimates if the missing is ignored. On the

other hand, survival effect depends on true unobserved longitudinal measures, the more bias in longitudinal estimation, the more bias result in survival estimation.

Table 3.1: True parameter values for the two longitudinal models and the proportional hazard function.

Outcome 1		Outcome 2		Time to event Outcome	
Parameter	Value	Parameter	Value	Parameter	Value
β_{01}	0.5	β_{02}	1	a	0.005
β_{11}	0.4	β_{12}	0.2	b	1.1
β_{21}	0.2	β_{22}	0.5	α_1	1.0
				α_2	1.5

Table 3.2: True values for random errors, random intercept and correlation for the longitudinal models, and censoring percentage from survival.

Scenario	σ_{01}	σ_{11}	σ_{02}	σ_{12}	ρ_1	ρ_2	Censor
1	0.2	0.15	0.2	0.15	0.1	0.4	40%
2	0.2	0.15	0.2	0.15	0.1	0.4	30%
3	0.2	0.15	0.2	0.15	0.1	0.4	10%
4	0.2	0.15	0.2	0.15	0.1	0.4	0%

Table 3.3: Simulation results for comparing the two-stage approach with the joint generalized estimating equation (JGEE) approach (40% censoring).

		Two-Stage					GEE					JGEE				
Sample Size	True Parm	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE
$n = 200$	$\beta_{01}(0.5)$	0.068	0.051	0.048	87.1%	0.004	0.163	0.061	0.059	68.2%	0.010	0.058	0.054	0.051	90.4%	0.004
	$\beta_{11}(0.4)$	-0.052	0.024	0.023	82.1%	0.001	-0.067	0.030	0.027	75.0%	0.002	0.013	0.028	0.026	92.8%	0.001
	$\beta_{21}(0.2)$	-0.064	0.061	0.059	94.2%	0.004	-0.233	0.083	0.082	89.6%	0.009	-0.057	0.063	0.061	94.4%	0.004
	$\beta_{02}(1.0)$	0.036	0.043	0.042	85.3%	0.003	0.093	0.047	0.046	49.0%	0.011	0.038	0.046	0.044	84.5%	0.003
	$\beta_{12}(0.2)$	-0.134	0.021	0.021	73.9%	0.001	-0.246	0.022	0.022	35.0%	0.003	-0.055	0.025	0.023	88.2%	0.001
	$\beta_{22}(0.5)$	-0.023	0.053	0.051	94.4%	0.003	-0.087	0.063	0.060	86.6%	0.006	-0.022	0.054	0.052	93.8%	0.003
	$\alpha_1(1.0)$	-0.090	0.273	0.279	94.2%	0.083	-0.122	0.281	0.287	93.2%	0.094	-0.055	0.307	0.297	95.0%	0.097
	$\alpha_2(1.5)$	-0.055	0.246	0.259	95.6%	0.067	-0.054	0.262	0.269	94.4%	0.075	-0.012	0.272	0.274	95.4%	0.074
$n = 500$	$\beta_{01}(0.5)$	0.067	0.030	0.031	81.7%	0.002	0.160	0.037	0.038	42.0%	0.008	0.053	0.032	0.033	87.5%	0.002
	$\beta_{11}(0.4)$	-0.054	0.015	0.015	67.9%	0.001	-0.066	0.020	0.018	63.6%	0.001	0.015	0.018	0.017	92.5%	0.000
	$\beta_{21}(0.2)$	-0.049	0.037	0.038	96.2%	0.001	-0.215	0.050	0.053	87.2%	0.004	-0.034	0.037	0.039	95.8%	0.001
	$\beta_{02}(1.0)$	0.034	0.027	0.027	75.2%	0.002	0.093	0.030	0.030	13.6%	0.009	0.033	0.029	0.028	78.2%	0.002
	$\beta_{12}(0.2)$	-0.134	0.014	0.013	49.8%	0.001	-0.248	0.015	0.014	6.8%	0.003	-0.046	0.017	0.014	85.1%	0.000
	$\beta_{22}(0.5)$	-0.019	0.033	0.033	93.1%	0.001	-0.092	0.038	0.039	77.4%	0.004	-0.019	0.034	0.033	93.8%	0.001
	$\alpha_1(1.0)$	-0.073	0.166	0.168	93.1%	0.033	-0.099	0.166	0.173	89.8%	0.037	-0.035	0.184	0.179	94.0%	0.035
	$\alpha_2(1.5)$	-0.061	0.154	0.158	90.9%	0.032	-0.065	0.160	0.163	89.8%	0.035	-0.022	0.168	0.166	94.4%	0.029
$n = 1000$	$\beta_{01}(0.5)$	0.071	0.022	0.022	59.8%	0.002	0.163	0.029	0.027	17.4%	0.008	0.059	0.024	0.023	74.4%	0.001
	$\beta_{11}(0.4)$	-0.058	0.011	0.010	40.8%	0.001	-0.070	0.013	0.013	39.4%	0.001	0.009	0.013	0.012	92.4%	0.000
	$\beta_{21}(0.2)$	-0.066	0.027	0.027	90.8%	0.001	-0.231	0.042	0.038	74.6%	0.004	-0.053	0.028	0.027	94.0%	0.001
	$\beta_{02}(1.0)$	0.035	0.019	0.019	50.6%	0.002	0.093	0.021	0.021	1.4%	0.009	0.035	0.020	0.020	58.4%	0.002
	$\beta_{12}(0.2)$	-0.136	0.009	0.009	19.6%	0.001	-0.249	0.010	0.010	0.0%	0.003	-0.048	0.011	0.010	80.4%	0.000
	$\beta_{22}(0.5)$	-0.025	0.023	0.023	91.8%	0.001	-0.095	0.028	0.027	56.8%	0.003	-0.024	0.023	0.023	91.4%	0.001
	$\alpha_1(1.0)$	-0.079	0.107	0.118	91.8%	0.018	-0.110	0.106	0.121	87.8%	0.023	-0.047	0.117	0.125	94.2%	0.016
	$\alpha_2(1.5)$	-0.064	0.109	0.111	84.8%	0.021	-0.072	0.112	0.114	85.4%	0.024	-0.027	0.118	0.116	93.0%	0.016

Table 3.4: Simulation results for comparing the two-stage approach with the weighted joint generalized estimating equation (JGEE) approach (30% censoring).

		Two-Stage					GEE					JGEE				
Sample Size	True Parm	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE
$n = 200$	$\beta_{01}(0.5)$	0.071	0.049	0.047	86.3%	0.004	0.177	0.061	0.059	63.2%	0.012	0.061	0.052	0.050	90.0%	0.004
	$\beta_{11}(0.4)$	-0.056	0.022	0.022	81.7%	0.001	-0.073	0.027	0.025	75.2%	0.002	0.011	0.026	0.024	92.6%	0.001
	$\beta_{21}(0.2)$	-0.068	0.060	0.058	93.6%	0.004	-0.259	0.086	0.083	89.4%	0.010	-0.061	0.062	0.060	94.0%	0.004
	$\beta_{02}(1.0)$	0.039	0.042	0.041	84.1%	0.003	0.102	0.047	0.046	41.0%	0.013	0.039	0.045	0.043	84.1%	0.004
	$\beta_{12}(0.2)$	-0.144	0.020	0.020	68.3 %	0.001	-0.268	0.021	0.020	25.8%	0.003	-0.062	0.024	0.021	87.1%	0.001
	$\beta_{22}(0.5)$	-0.024	0.052	0.051	93.8%	0.003	-0.099	0.064	0.060	86.0%	0.006	-0.024	0.054	0.052	93.2%	0.003
	$\alpha_1(1.0)$	-0.094	0.239	0.242	92.6%	0.066	-0.129	0.240	0.248	91.8%	0.074	-0.062	0.261	0.257	94.4%	0.072
	$\alpha_2(1.5)$	-0.053	0.229	0.233	92.0%	0.059	-0.060	0.243	0.240	91.8%	0.067	-0.014	0.252	0.245	93.4%	0.064
$n = 500$	$\beta_{01}(0.5)$	0.070	0.030	0.030	79.3%	0.002	0.178	0.037	0.038	35.0%	0.009	0.056	0.032	0.032	86.3%	0.002
	$\beta_{11}(0.4)$	-0.058	0.014	0.014	59.6%	0.002	-0.075	0.018	0.016	51.0%	0.001	0.011	0.017	0.015	92.0%	0.000
	$\beta_{21}(0.2)$	-0.051	0.036	0.037	94.4%	0.001	-0.243	0.051	0.054	83.8%	0.005	-0.036	0.037	0.038	95.4%	0.001
	$\beta_{02}(1.0)$	0.036	0.027	0.026	72.2%	0.002	0.102	0.029	0.029	07.2%	0.011	0.035	0.029	0.028	73.8%	0.002
	$\beta_{12}(0.2)$	-0.142	0.014	0.013	39.6%	0.001	-0.268	0.014	0.013	2.6%	0.003	-0.052	0.017	0.014	81.9%	0.000
	$\beta_{22}(0.5)$	-0.022	0.032	0.032	94.0%	0.001	-0.105	0.038	0.039	73.0%	0.004	-0.022	0.033	0.033	94.0%	0.001
	$\alpha_1(1.0)$	-0.071	0.145	0.146	91.8%	0.026	-0.101	0.148	0.150	88.2%	0.032	-0.034	0.161	0.155	92.4%	0.027
	$\alpha_2(1.5)$	-0.062	0.141	0.142	90.1%	0.028	-0.074	0.144	0.146	85.8%	0.033	-0.026	0.152	0.148	93.8%	0.025
$n = 1000$	$\beta_{01}(0.5)$	0.075	0.022	0.021	56.6%	0.002	0.181	0.028	0.027	10.2%	0.009	0.061	0.023	0.023	72.8%	0.001
	$\beta_{11}(0.4)$	-0.062	0.010	0.010	29.8%	0.001	-0.078	0.012	0.012	24.6%	0.001	0.007	0.012	0.011	92.2%	0.000
	$\beta_{21}(0.2)$	-0.068	0.027	0.026	92.0%	0.001	-0.261	0.042	0.038	70.6%	0.004	-0.054	0.028	0.027	94.6%	0.001
	$\beta_{02}(1.0)$	0.037	0.018	0.019	45.2%	0.002	0.103	0.021	0.021	0.2%	0.011	0.036	0.020	0.020	52.0%	0.001
	$\beta_{12}(0.2)$	-0.146	0.009	0.009	8.8%	0.001	-0.271	0.010	0.009	0.0%	0.003	-0.056	0.011	0.010	74.6%	0.000
	$\beta_{22}(0.5)$	-0.026	0.023	0.023	92.0%	0.001	-0.108	0.029	0.027	50.0%	0.004	-0.026	0.023	0.023	91.8%	0.001
	$\alpha_1(1.0)$	-0.078	0.095	0.103	89.6%	0.015	-0.111	0.095	0.105	82.8%	0.021	-0.046	0.104	0.109	94.6%	0.013
	$\alpha_2(1.5)$	-0.066	0.097	0.099	82.0%	0.019	-0.081	0.099	0.102	77.6%	0.024	-0.032	0.104	0.104	91.0%	0.013

Table 3.5: Simulation results for comparing the two-stage approach with the joint generalized estimating equation (JGEE) approach (10% censoring).

		Two-Stage					GEE					JGEE				
Sample Size	True Parm	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE
$n = 200$	$\beta_{01}(0.5)$	0.077	0.047	0.045	83.9%	0.004	0.209	0.060	0.057	50.4%	0.015	0.065	0.051	0.048	88.7%	0.004
	$\beta_{11}(0.4)$	-0.062	0.020	0.020	76.5%	0.001	-0.083	0.024	0.022	61.0%	0.002	0.009	0.024	0.022	92.8%	0.001
	$\beta_{21}(0.2)$	-0.071	0.058	0.057	94.6%	0.004	-0.323	0.086	0.083	87.8%	0.012	-0.065	0.061	0.059	94.0%	0.004
	$\beta_{02}(1.0)$	0.042	0.041	0.040	80.9%	0.003	0.119	0.047	0.045	27.0%	0.016	0.042	0.044	0.042	82.3%	0.004
	$\beta_{12}(0.2)$	-0.159	0.018	0.018	56.9%	0.001	-0.299	0.018	0.018	11.4%	0.004	-0.070	0.021	0.019	85.1%	0.001
	$\beta_{22}(0.5)$	-0.027	0.052	0.049	93.6%	0.003	-0.124	0.064	0.060	79.8%	0.008	-0.028	0.053	0.050	93.4%	0.003
	$\alpha_1(1.0)$	-0.085	0.202	0.195	92.0%	0.048	-0.125	0.203	0.198	89.8%	0.057	-0.054	0.220	0.206	92.8%	0.051
	$\alpha_2(1.5)$	-0.056	0.189	0.195	92.2%	0.043	-0.075	0.195	0.200	88.8%	0.057	-0.021	0.204	0.204	94.8%	0.043
$n = 500$	$\beta_{01}(0.5)$	0.075	0.028	0.029	73.9%	0.002	0.206	0.037	0.037	19.8%	0.012	0.061	0.030	0.031	83.5%	0.002
	$\beta_{11}(0.4)$	-0.064	0.013	0.012	47.9%	0.001	-0.086	0.016	0.014	35.6%	0.001	0.008	0.015	0.014	91.0%	0.000
	$\beta_{21}(0.2)$	-0.058	0.034	0.036	93.9%	0.001	-0.298	0.052	0.054	79.4%	0.006	-0.044	0.035	0.037	95.1%	0.001
	$\beta_{02}(1.0)$	0.039	0.026	0.025	65.4%	0.002	0.119	0.028	0.029	1.2%	0.015	0.038	0.028	0.027	69.0%	0.002
	$\beta_{12}(0.2)$	-0.158	0.013	0.011	23.6%	0.001	-0.299	0.013	0.011	0.6%	0.004	-0.064	0.015	0.012	75.4%	0.000
	$\beta_{22}(0.5)$	-0.023	0.032	0.031	92.9%	0.001	-0.127	0.037	0.038	63.6%	0.005	-0.025	0.032	0.032	93.9%	0.001
	$\alpha_1(1.0)$	-0.074	0.116	0.119	89.4%	0.019	-0.110	0.118	0.121	83.6%	0.026	-0.040	0.129	0.125	92.7%	0.018
	$\alpha_2(1.5)$	-0.067	0.116	0.120	86.6%	0.023	-0.088	0.118	0.122	80.4%	0.031	-0.034	0.123	0.125	91.2%	0.018
$n = 1000$	$\beta_{01}(0.5)$	0.079	0.021	0.020	49.2%	0.002	0.208	0.028	0.026	4.2%	0.012	0.066	0.022	0.022	66.5%	0.002
	$\beta_{11}(0.4)$	-0.067	0.009	0.009	15.4%	0.001	-0.088	0.011	0.010	9.0%	0.001	0.004	0.011	0.010	91.3%	0.000
	$\beta_{21}(0.2)$	-0.070	0.026	0.025	91.3%	0.001	-0.306	0.040	0.038	62.2%	0.005	-0.056	0.027	0.026	93.1%	0.001
	$\beta_{02}(1.0)$	0.041	0.018	0.018	37.8%	0.002	0.120	0.021	0.020	0.0%	0.015	0.040	0.019	0.019	42.7%	0.001
	$\beta_{12}(0.2)$	-0.160	0.008	0.008	1.4%	0.001	-0.301	0.009	0.008	0.0%	0.004	-0.067	0.010	0.009	61.6%	0.000
	$\beta_{22}(0.5)$	-0.028	0.022	0.022	90.7%	0.001	-0.132	0.029	0.027	33.2%	0.005	-0.030	0.023	0.023	89.8%	0.001
	$\alpha_1(1.0)$	-0.083	0.078	0.083	81.9%	0.013	-0.122	0.078	0.085	68.0%	0.021	-0.054	0.085	0.088	91.7%	0.001
	$\alpha_2(1.5)$	-0.072	0.077	0.084	75.8%	0.018	-0.095	0.079	0.084	63.4%	0.027	-0.041	0.083	0.087	88.6%	0.011

Table 3.6: Simulation results for comparing the two-stage approach with the joint generalized estimating equation (JGEE) approach (0% censoring).

		Two-Stage					GEE					JGEE				
Sample Size	True Parm	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE	Relative Bias	Emp. S.E.	Model S.E.	95%CI C.P.	MSE
$n = 200$	$\beta_{01}(0.5)$	0.078	0.046	0.044	84.3%	0.004	0.221	0.060	0.057	47.2%	0.016	0.066	0.050	0.048	86.9%	0.004
	$\beta_{11}(0.4)$	-0.064	0.019	0.019	72.9%	0.001	-0.088	0.023	0.021	56.4%	0.002	0.008	0.023	0.021	91.8%	0.001
	$\beta_{21}(0.2)$	-0.069	0.057	0.056	94.8%	0.003	-0.339	0.084	0.083	86.6%	0.012	-0.061	0.059	0.058	94.2%	0.004
	$\beta_{02}(1.0)$	0.043	0.041	0.039	79.7%	0.004	0.125	0.047	0.044	22.6%	0.018	0.043	0.044	0.041	81.1%	0.004
	$\beta_{12}(0.2)$	-0.165	0.018	0.017	51.6%	0.001	-0.309	0.017	0.017	7.8%	0.018	-0.075	0.021	0.018	81.5%	0.001
	$\beta_{22}(0.5)$	-0.027	0.051	0.049	93.0%	0.003	-0.132	0.063	0.060	77.6%	0.008	-0.029	0.052	0.050	93.4%	0.003
	$\alpha_1(1.0)$	-0.081	0.186	0.179	92.0%	0.041	-0.122	0.188	0.182	88.2%	0.050	-0.049	0.204	0.190	92.6%	0.044
	$\alpha_2(1.5)$	-0.060	0.184	0.183	89.6%	0.042	-0.083	0.186	0.186	87.0%	0.050	-0.028	0.196	0.190	92.8%	0.040
$n = 500$	$\beta_{01}(0.5)$	0.078	0.027	0.028	72.1%	0.002	0.220	0.036	0.036	16.4%	0.013	0.063	0.030	0.030	83.8%	0.002
	$\beta_{11}(0.4)$	-0.067	0.012	0.012	40.5%	0.002	-0.091	0.015	0.014	27.2%	0.013	0.006	0.015	0.013	91.2%	0.000
	$\beta_{21}(0.2)$	-0.060	0.034	0.035	94.0%	0.001	-0.323	0.052	0.053	80.4%	0.007	-0.046	0.035	0.036	95.2%	0.001
	$\beta_{02}(1.0)$	0.041	0.025	0.025	63.3%	0.002	0.125	0.028	0.028	0.4%	0.007	0.040	0.028	0.026	66.3%	0.002
	$\beta_{12}(0.2)$	-0.164	0.012	0.011	19.0%	0.002	-0.310	0.012	0.011	0.2%	0.004	-0.069	0.015	0.012	68.9%	0.000
	$\beta_{22}(0.5)$	-0.023	0.031	0.031	93.4%	0.001	-0.136	0.036	0.038	57.2%	0.006	-0.026	0.032	0.032	93.6%	0.001
	$\alpha_1(1.0)$	-0.075	0.108	0.110	88.8%	0.017	-0.114	0.110	0.111	81.6%	0.025	-0.042	0.119	0.116	91.8%	0.016
	$\alpha_2(1.5)$	-0.070	0.107	0.112	83.2%	0.023	-0.094	0.107	0.114	74.8%	0.031	-0.038	0.113	0.117	91.8%	0.016
$n = 1000$	$\beta_{01}(0.5)$	0.081	0.020	0.020	45.3%	0.002	0.219	0.027	0.026	2.2%	0.013	0.067	0.022	0.021	63.6%	0.002
	$\beta_{11}(0.4)$	-0.069	0.009	0.008	11.5%	0.001	-0.091	0.010	0.010	6.4%	0.001	0.003	0.011	0.009	90.5%	0.002
	$\beta_{21}(0.2)$	-0.070	0.026	0.025	90.9%	0.001	-0.327	0.040	0.038	59.6%	0.006	-0.056	0.026	0.026	92.9%	0.002
	$\beta_{02}(1.0)$	0.042	0.018	0.018	31.7%	0.002	0.126	0.021	0.020	0.0%	0.016	0.041	0.019	0.019	39.2%	0.002
	$\beta_{12}(0.2)$	-0.165	0.008	0.008	0.8%	0.001	-0.311	0.008	0.008	0.0%	0.016	-0.070	0.009	0.008	58.2%	0.000
	$\beta_{22}(0.5)$	-0.029	0.022	0.022	89.5%	0.001	-0.141	0.029	0.027	28.0%	0.006	-0.031	0.023	0.022	88.9%	0.001
	$\alpha_1(1.0)$	-0.088	0.074	0.077	79.2%	0.013	-0.128	0.073	0.078	61.4%	0.022	-0.058	0.080	0.081	89.5%	0.010
	$\alpha_2(1.5)$	-0.075	0.073	0.079	70.1%	0.013	-0.099	0.074	0.080	54.6%	0.022	-0.044	0.078	0.082	87.7%	0.010

3.5 Data Application

In this section, the proposed method is applied to data from a longitudinal cohort to examine the association of longitudinally collected cognitive memory score and cognitive executive score and their association with the risk of coronary artery disease (CAD). The cognitive scores are counting number of items the participant answered correctly in the questionnaire instrument, with higher scores indicating better cognitive function.

Existing work of CAD and cognition mainly focused on examing the cognitive funciton in patients with heart failure (Cannon et al., 2015). It is well reported that CAD is associated with cognition function decline (Borowicz et al., 1996; Selnes et al., 1999; Saxton et al., 2000; Trojano et al., 2003). Zheng et al. (2012) have showed that history of coronary artery disease was associated with greater declines in global, verbal memory and executive cognition. However, no one has investigated changes in cognitive memory and executive scores over time before diagnosis of CAD and the association between longitudinal cognitive memory and cognitive executive scores and the risk of CAD.

3.5.1 The Longitudinal Data Cohort

The study population comes from the Indianapolis cohort of the Indianapolis-Ibadan Dementia Project (IIDP). The IIDP is a 20-year study of dementia in elderly African Americans living in Indianapolis, Indiana, Ibadan and Nigeria, all were age 65 or older. The project recruitment was conducted in two phases: for the first recruitment, 2212 African Americans age 65 or older who live in Indianapolis were enrolled in 1992; for

the second enrollment in 2001, additional 1893 community-dwelling African American with age 70 years and older were recruited. Detailed description of the IIDP cohort can be found at Hendrie et al. (2001) and Hall et al. (2009).

The cognitive memory and executive scores are from IIDP measured with Community Screening Instrument for Dementia (CSID). The CSID was developed as a screening tool for Dementia in populations with various cultural backgrounds and literacy levels. CSID had been widely used as a screening tool for demeina and evaluates multiple cognitive domains, including language, attention, memory, orientation, praxis, comprehension and motor response (Hall et al., 1996). Details in the CSID instrument has been published in Hall et al. (1996). The CSID has demonstrated good reliability and validity in detecting dementia in various populations (Hendrie et al., 2001; Hall et al., 2000).

The event of CAD was defined as the diagnosis of CAD using ICD-9 codes or cause of death by CAD. For patients with an event, age at CAD diagnosis is used as the event time; for subjects without CAD diagnosis, age at last evaluation time in IIDP or age of death other than CAD is used as the right censoring point.

The current work focuses on a subset of 1323 male patients, with a total number of 3230 observations. During the study follow-up, 365(27.59%) of the subjects were diagnosed with CAD by the end of study period. The overall average number of longitudinal cognitive measurements is 2.4(SD1.4). The average length of follow up is 3.65 years (range 0 to 17.25). Demographic characteristics of the study population including baseline age, year of education, baseline smoking status and diabetes condition are summarized in Table 3.7. CAD and non-CAD patients are not different in baseline age, education years, baseline smoking status, baseline diabetes percent-

age and baseline cognitive memory levels, but they are different in baseline cognitive executive levels.

Figure 3.1 and 3.2 plot longitudinal cognitive memory and executive scores over time. The blue dots are cognitive scores for Non-CAD group, and red dots are cognitive scores for CAD group. The blue and red lines represent fitted population average for Non-CAD and CAD group respectively. From figure 3.1, population mean cognitive memory score over time for patients in CAD and Non-CAD group are both decreasing. From figure 3.2, population mean cognitive executive score over time for patients in Non-CAD group stays flat, while population mean cognitive executive score decreases over time for patients in CAD group. Figure 3.1 and 3.2 indicates a potential association between the risk of CAD and longitudinal memory and executive cognition over time.

Table 3.7: Comparison of baseline demographic of male subjects between CAD and non-CAD groups ($n = 1323$)

Baseline Characters	All (n=1323)	CAD (n=365)	Non-CAD (n=958)	p-value
Age at baseline, mean(sd)	75.28(6.29)	75.08(6.22)	75.36(6.32)	0.474
Year of Education, mean(sd)	9.92(3.41)	9.86(3.40)	9.94(3.41)	0.681
Smoking, n (%)	1033(78.08%)	298(81.64%)	735(76.72%)	0.053
Diabetes, n (%)	338(25.55%)	106(29.04%)	232(24.22%)	0.072
Baseline memory, mean (sd)	25.79(4.66)	26.19(4.22)	25.64(4.81)	0.057
Baseline executive, mean (sd)	31.22(5.33)	31.87(5.07)	30.97(5.41)	0.006
Number of measures, mean(sd)	2.44(1.40)	2.19(1.26)	2.54(1.44)	<.0001
Years of follow up, median(range)	2.25(0, 17.25)	1.92(0, 16.18)	2.40(0, 17.25)	0.0002

Figure 3.1: Observed longitudinal memory cognitive scores over time for the CAD and Non-CAD group.

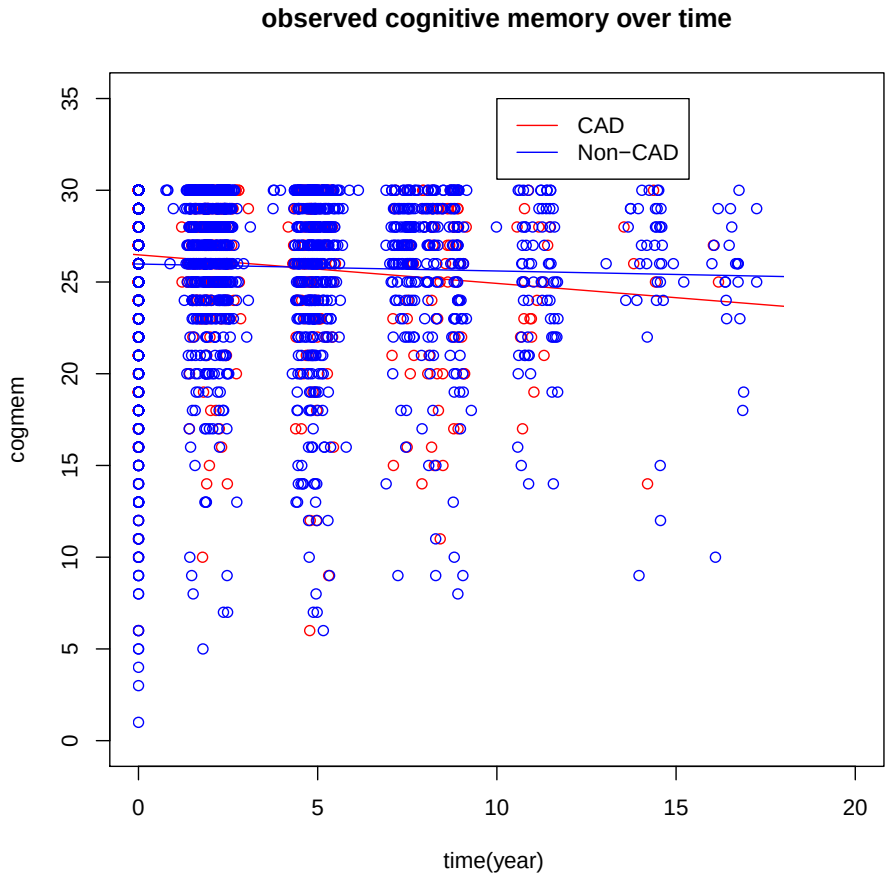
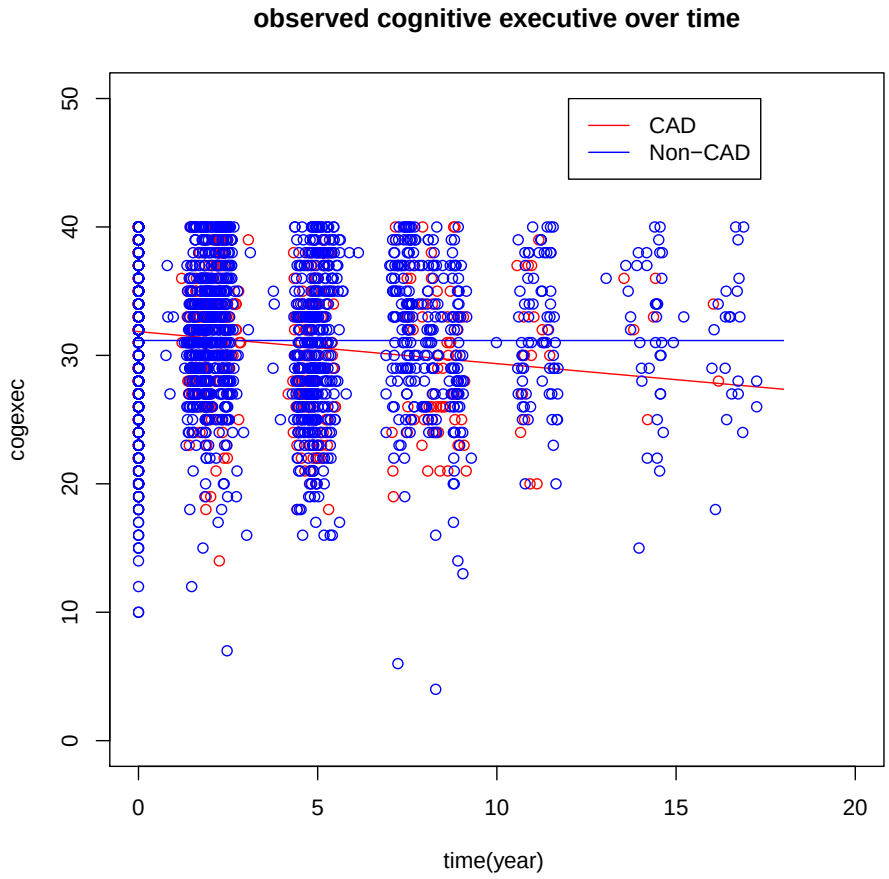


Figure 3.2: Observed longitudinal executive cognitive scores over time for the CAD and Non-CAD group.



3.5.2 Application to IIDP Data

The proposed JGEE method is applied to the longitudinal cohort with 1323 male patients. Joint modeling with two-stage approach are compared with joint modeling with proposed JGEE method.

Consider the following longitudinal model:

$$\log(E(y_l(t_{ij}))) = m_l(t_{ij}) = \beta_{0l}+\beta_{1l}t_{ij}+\beta_{2l}grade_i+\beta_{3l}baseage_i+b_{0li}+b_{1li}t_{ij}, \quad (3.12)$$

where y_l, $l = 1, 2$ denote the observed longitudinal cognitive memory and executive scores. Grade indicates year of education, baseage denotes age at baseline. $\mathbf{b}_i = (b_{01i}, b_{11i}, b_{02i}, b_{12i})^T$ is random effect with mean zero and variance covariance matrix $\mathbf{G}$. The survival model is assumed to follow the Cox PH model with form:

$$h(t) = h_0(t)\exp\{\gamma_1 baseage_i + \gamma_2 smoke_i + \gamma_3 diabetes_i + \alpha_1 m_1(t) + \alpha_2 m_2(t)\}. \quad (3.13)$$

This hazard function assumes that the logarithm of hazard of CAD depends on baseline age, smoking status, history of diabetes, and the logarithm of expected cognitive memory and executive score.

We also considered the hazard function assumes that the logarithm of hazard of CAD depends on baseline age, smoking status, history of diabetes, and the logarithm of observed cognitive memory and executive score:

$$\begin{aligned} h(t) = h_0(t)\exp\{&\gamma_1 baseage_i + \gamma_2 smoke_i + \gamma_3 diabetes_i \\ &+ \alpha_1 \log cogmem + \alpha_2 \log cogexec\}, \end{aligned} \quad (3.14)$$

where cogmem represents the observed cognitive memory score and cogexec denotes the observed cognitive executive score. Results for Cox model (3.14) with observed memory and executive scores at baseline as covariates and Cox model (3.14) with observed memory and executive scores as time dependent covariates are presented in Table 3.8 and 3.9 respectively. Note that model (3.14) has different assumption with the joint models (defined by (3.12) and (3.13)).

Estimated parameters applying the two-stage and proposed JGEE method for joint models are summarized in Tables 3.10 and 3.11. It can be observed that on population average, the logarithm of expected memory and executive cognitive score are both decreasing over time. Subjects with higher education level have higher cognitive memory and executive scores. Older subjects have lower cognitive memory and executive scores. More years of educated subjects have higher log expected memory and executive cognitive scores. For time to CAD, higher age is related to higher risk of developing CAD, subjects with diabetes have higher risk of CAD. Subjects with higher logarithm of memory cognitive score have lower risk of developing CAD. Subjects with higher logarithm of expected executive cognitive score have lower risk of developing CAD. However, either memory or executive score reach statistical significance for risk of CAD.

Table 3.8: Parameter estimates, standard errors and 95% CI using Cox model with logarithm of observed memory and executive scores at baseline. α_1 and α_2 are association estimates between the risk of CAD and the logarithm of observed memory and executive scores

Parameter	Estimate	StdErr	Lower 95% CI	Upper 95% CI
Time to CAD				
Age at baseline	-0.129	0.010	-0.149	-0.109
Smoke	0.333	0.136	0.066	0.600
Diabetes	0.331	0.116	0.103	0.559
α_1	-0.245	0.291	-0.815	0.326
α_2	0.696	0.361	-0.012	1.404

Table 3.9: Parameter estimates, standard errors and 95% CI using Cox model with logarithm of observed memory and executive scores as time dependent covariates. α_1 and α_2 are association estimates between the risk of CAD and the logarithm of observed memory and executive scores

Parameter	Estimate	StdErr	Lower 95% CI	Upper 95% CI
Time to CAD				
Age at baseline	0.128	0.013	0.102	0.153
Smoke	0.179	0.134	-0.084	0.442
Diabetes	0.362	0.118	0.118	0.593
α_1	0.090	0.256	-0.413	0.593
α_2	-0.287	0.343	-0.960	0.386

Table 3.10: Parameter estimates, standard errors and 95% CI using joint models with the two-stage approach. α_1 and α_2 are association estimates between the risk of CAD and the current value at the event time point

Parameter	Estimate	StdErr	Lower 95% CI	Upper 95% CI
Longitudinal memory cognitive score				
Intercept	3.521	0.048	3.427	3.615
Time	-0.008	0.001	-0.009	-0.006
Year of Education	0.017	0.001	0.015	0.019
Age at baseline	-0.006	0.001	-0.007	-0.005
Longitudinal executive cognitive score				
Intercept	3.525	0.050	3.426	3.623
Time	-0.007	0.001	-0.009	-0.005
Year of Education	0.018	0.001	0.016	0.020
Age at baseline	-0.003	0.001	-0.005	-0.002
Time to CAD				
Age at baseline	0.123	0.014	0.096	0.150
Smoke	0.152	0.136	-0.113	0.418
Diabetes	0.371	0.118	0.140	0.602
α_1	-0.353	0.496	-1.326	0.619
α_2	-0.713	0.539	-1.770	0.343

Table 3.11: Parameter estimates, standard errors and 95% CI with joint generalized estimating equation (JGEE) approach. α_1 and α_2 are association estimates between the risk of CAD and the current value at the event time point

Parameter	Estimate	StdErr	Lower 95% CI	Upper 95% CI
Longitudinal memory cognitive score				
Intercept	3.734	0.195	3.351	4.117
Time	-0.011	0.004	-0.019	-0.002
Year of Education	0.013	0.005	0.003	0.022
Age at baseline	-0.008	0.002	-0.013	-0.004
Longitudinal executive cognitive score				
Intercept	3.518	0.193	3.140	3.897
Time	-0.008	0.004	-0.016	-0.001
Year of Education	0.012	0.004	0.004	0.020
Age at baseline	-0.003	0.002	-0.008	0.002
Time to CAD				
Age at baseline	0.127	0.013	0.101	0.153
Smoke	0.180	0.136	-0.086	0.445
Diabetes	0.365	0.118	0.135	0.596
α_1	-0.039	0.810	-1.626	1.548
α_2	-0.149	1.164	-2.431	2.132

3.6 Discussion

In this work, we have extended the joint generalized estimating equation approach (JGEE) to parameter estimation in joint modeling of multiple longitudinal count measures and a time-to-event outcome. The performance of the proposed approach has been evaluated through extensive simulation studies. We have illustrated the proposed method using data from a longitudinal cohort study to examine the association of longitudinally collected cognitive memory score and cognitive executive score and their association with the risk of coronary artery disease (CAD).

The proposed JGEE method enjoys several advantages. Firstly, JGEE method is computationally efficient to carry out as it avoids the direct maximization of the joint likelihood. Secondly, JGEE method corrects biases from the naive two-stage method by adjusting for informative truncation from the time-to-event outcome with inverse intensity weighting. Thirdly, the JGEE method estimates the longitudinal and survival parameters simultaneously by updating inverse intensity weight and solving the estimating equations iteratively.

The current work also has some limitations. Through simulation studies, we observed that although the JGEE method reduces biased from the naive two-stage method, but the proposed estimators can still be biased. We know in the case of log linear model for Poisson data, the marginal mean and the conditional mean are shifted by an offset. Thus the first estimating equation in the JGEE will need to be adjusted to correctly reflect this shift in order to produce parameter estimates that correspond to those specified in the longitudinal model. In addition, estimated standard errors are underestimating the true variance, because the standard errors

of fixed effects in longitudinal models are calculated based on modified GEE as if the weights are known, and the standard errors of the risk coefficient estimators are obtained through partial likelihood assuming that all the true covariate values are known.

In this work, we have focused on joint models with longitudinal Poisson outcomes, it is straightforward to extend to joint models with various distributed longitudinal outcomes, including binary and a mixture of different outcomes. A unified algorithm can be developed to estimate those different types of longitudinal outcomes by specifying different link functions for different distributions. Furthermore, JGEE method does not inherit consistency and asymptotic property as the likelihood method does. Thus, theoretical justification on the consistency and asymptotic of the parameters are necessary.

Chapter 4

Joint Models for Multiple Longitudinal Outcomes from the Exponential Family and Time-to-Event Data

In Chapter 2, we introduced joint models for analyzing the association between multiple longitudinal continuous measures and a time-to-event outcome. In Chapter 3, we considered joint models with multiple longitudinal count responses and one time-to-event outcome. In this chapter, we accommodate different types of longitudinal responses in a unified framework, where the longitudinal outcomes can be members of the exponential family. This chapter is organized as follows. Section 4.1 specifies the formulation and parameterizations of the joint models in full generality. Section 4.2 presents the details of the proposed joint generalized estimation equation (JGEE) procedure for parameter estimation. The remaining sections illustrate the consistency and asymptotic properties of the estimators.

4.1 Model Formulation

4.1.1 Longitudinal Models

Let $\mathbf{Y}_i = (\mathbf{y}_{i1}^T, \cdots, \mathbf{y}_{iL}^T)^T$ denotes L-variate response vector for the i-th subject, $\mathbf{y}_{il} = (y_{il}(t_{i1}), \cdots, y_{il}(t_{in_i}))^T$ is a $n_i \times 1$ vector of longitudinal response at different time points, where $i = 1, \cdots, N$ and $l = 1, \cdots, L$. $y_{il}(t_{ij})$ is the l-th response collected on subject i at time point t_{ij}, where $j = 1, \cdots, n_i$. Let $\mathbf{t}_i = (t_{i1}, \cdots, t_{in_i})^T$ denotes time points for subject i. Assuming the l-th longitudinal outcome has the set of covariates

$\mathbf{X}_{il} = (\mathbf{X}_{il}^T(t_{i1}), \cdots, \mathbf{X}_{il}^T(t_{in_i}))^T$, where $\mathbf{X}_{il}^T(t_{ij}) = (1, x_{il1}(t_{ij}), \cdots, x_{ilp_l}(t_{ij}))^T$ is a $(p_l + 1) \times 1$ vector of covariates for the l-th response of subject i at time t_{ij}.

The l-th response for i-th subject is modeled by a multivariate generalized linear mixed effects model:

$$g_l \{E(y_{il}(t)|\mathbf{b}_{il})\} = m_{il}(t) = \mathbf{X}_{il}^T(t)\boldsymbol{\beta}_l + \mathbf{Z}_{il}^T(t)\mathbf{b}_{il}, \tag{4.1}$$

where $g_l(\cdot)$ denotes a known one-to-one monotonic link function of l-th longitudinal response; for example, if longitudinal response is a proportion, $g_l(\cdot)$ can be the logit link function; if the longitudinal response variable is a count, $g_l(\cdot)$ can be a log link function; for continuous outcomes, $g_l(\cdot)$ is the identity function; $\mathbf{X}_{il}^T(t)$ and $\mathbf{Z}_{il}^T(t)$ denote row vectors of the design matrix for fixed and random effects corresponding to the l-th longitudinal outcome for the i-th subject at time point t; $\boldsymbol{\beta}_l$ is the corresponding parameter vector of the fix effects; $\mathbf{b}_{il}$ is the according vector of random effects. Let $\mathbf{b}_i = (\mathbf{b}_{i1}^T, \cdots, \mathbf{b}_{il}^T)^T$ denote random effects for the i-th subject with mean $\mathbf{0}$ and variance covarianc $\mathbf{G}$. Random effects for different outcomes are correlated in the same subject, but independent for different subjects. The unknown function $m_{il}(t)$ describes the true profile for the l-th outcome.

For the longitudinal model, we have the conditional mean

$$E(y_{il}|\mathbf{b}_{il}) = g_l^{-1}(\mathbf{X}_{il}^T\boldsymbol{\beta}_l + \mathbf{Z}_{il}^T\mathbf{b}_{il}),$$

and the conditional variance

$$Var(y_{il}|\mathbf{b}_{il}) = \phi h(E(y_{il}|\mathbf{b}_{il})),$$

where $h(\cdot)$ represents a function of $E\left(y_{il}|\mathbf{b}_{il}\right)$. For example, for binary responses, $g(\cdot)$ is the logit link function, then $h(E\left(y_{il}|\mathbf{b}_{il}\right)) = E(y_{il}|\mathbf{b}_{il})(1-E(y_{il}|\mathbf{b}_{il}))$ and $\phi = 1$; for Poisson count responses, $g(\cdot)$ is the log link function, then $h(E\left(y_{il}|\mathbf{b}_{il}\right)) = E\left(y_{il}|\mathbf{b}_{il}\right)$ and $\phi = 1$.

4.1.2 The Survival Model

For a time to event outcome, let T_i^* be the true event time for subject i. Some subjects may not experience the event at the end of study, or they may drop out during the study, so their event times are right censored. Let C_i denote the censoring time for the i-th subject. The observed event time $T_i = \min(T_i^*, C_i)$. Define the event indicator as $\delta_i = I(T_i^* \leq C_i)$, here $I(\cdot)$ represents indicator function. Assume that censoring times and event times are independent. The effects of longitudinal outcomes on survival times are captured by a hazard model:

$$h_i(t) = h_0(t) \exp\left\{ \mathbf{u}_i^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l\left(m_{il}(t)\right) \right\}, \tag{4.2}$$

where $h_0(t)$ is the baseline risk function; $\mathbf{u}_i$ denotes a vector of time independent covariates with corresponding regression coefficient vector $\boldsymbol{\gamma}$; function $f_l(\cdot)$ is assumed to describe the true longitudinal profile for the l-th outcome; and α_l denotes a parameter measuring the effect of the l-th longitudinal process to the time to event outcome. $\exp\{\gamma_j\}$ denotes the hazard ratio for one unit change in u_{ij} at any time t. $\exp\{\alpha_l\}$ denotes the relative increase in the risk for an event at time t that results from one unit increase in $f_l(m_{il}(t))$ at the same time point. Parameters α_l's $(l = 1, 2, \cdots, L)$ are of primary interest in this work.

Function $f_l(\cdot)(l = 1, 2, \cdots, L)$ are assumed to be known and reflect the association between the longitudinal outcomes and the event process. Identity function quantifies the effect of the true current underlying longitudinal outcomes on the hazard for an event; a first derivative with respect to time t implies that the risk for an event at time t depends on the slope of the true longitudinal trajectory at time t; an integration function over time indicates that the event risk of depends on the cumulative history of the longitudinal biomarkers over time.

4.2 Estimation Method

Different approaches can be used for parameter estimation for the joint models defined above, the two-stage method can result in biased results; the maximum likelihood based approach and Bayesian method utilizes the full likelihood based on the joint models can be computationally intensive as the likelihood function or the posterior distribution function involves integration over multiple random effects. In this section, we present the joint generalized estimating equation methodology for parameter estimation of joint models for multiple longitudinal counts biomarkers and time-to-event outcome.

Recall from the longitudinal model that $g_l\left\{E\left(y_{il}|\mathbf{b}_{il}\right)\right\} = m_{il} = \mathbf{X}_{il}^T\boldsymbol{\beta}_l + \mathbf{Z}_{il}^T\mathbf{b}_{il}$. Then

$$E(y_{il}|\mathbf{b}_{il}) = g_l^{-1}(\mathbf{X}_{il}^T\boldsymbol{\beta}_l + \mathbf{Z}_{il}^T\mathbf{b}_{il}) = g_l^{-1}(m_{il}).$$

For Gaussian distributed $\mathbf{b}_{il}$ with mean $\mathbf{0}$ and covariance $\mathbf{G}_l$. The marginal mean of y_{il} is

$$E(y_{il}) = E_{\mathbf{b}_{il}}(E(y_{il}|\mathbf{b}_{il})) = E\left(g_l^{-1}\left(\mathbf{X}_{il}^T\boldsymbol{\beta}_l + \mathbf{Z}_{il}^T\mathbf{b}_{il}\right)\right).$$

Thus the expression for the marginal mean depends on the link functions. For the identity link ($g_l(u) = u$), we have $E(y_{il}) = \mathbf{X}_{il}^T \boldsymbol{\beta}_l$. For the log link ($g_l(u) = \log(u)$), we have $E(y_{il}) = e^{\mathbf{X}_{il}^T \boldsymbol{\beta}_l} E(e^{\mathbf{Z}_{il}^T \mathbf{b}_{il}}) = e^{\mathbf{X}_{il}^T \boldsymbol{\beta}_l + \mathbf{Z}_{il}^T \mathbf{G}_l \mathbf{Z}_{il}^T/2}$. That is, the random effect leads to an offset in the marginal mean. However, in general, there is no exact closed form for the marginal mean.

The variance of y_{il} does not have simple form except the linear link.

$$\begin{aligned} Var(y_{il}) &= Var(E(y_{il}|\mathbf{b}_{il})) + E(Var(y_{il}|\mathbf{b}_{il})) \\ &= Var(g_l^{-1}(\mathbf{X}_{il}^T \boldsymbol{\beta}_l + \mathbf{Z}_{il}^T \mathbf{b}_{il})) + E(\phi h(E(y_{il}|\mathbf{b}_{il}))). \end{aligned}$$

For the covariance, with $l \neq k$,

$$Cov(y_{il}, y_{ik}) = Cov\left[E(y_{il}|\mathbf{b}_{il}), E(y_{ik}|\mathbf{b}_{ik})\right] + E\left[Cov(y_{il}|\mathbf{b}_{il}, y_{ik}|\mathbf{b}_{ik})\right].$$

That is the correlation among different responses are introduced by correlation among random effects $\mathbf{b}_i$.

In Chapter 2, for the continuous longitudinal outcomes, we can estimate the fixed effect by a straightforward weighted GEE; in Chapter 3, with longitudinal count outcomes, the weighted GEE can be applied with an offset variable defined by the random effects; but for other distributions when there is no exact closed form for the marginal mean, an approximate GEE can be used with a linearized response. When using GEE approach, only an approximation for the variance $\mathbf{V}_i$ is necessary to obtain consistent and nearly efficient inferences for $\boldsymbol{\beta}$, when the number of subjects, N, is large relative to the number of observations per subject, n_i, and random effect variance $\mathbf{G}$ is given.

4.2.1 Linearization of the Longitudinal Model

Recall from previous introduction

$$E(y_{il}|\mathbf{b}_{il}) = g_l^{-1}(\mathbf{X}_{il}^T\boldsymbol{\beta}_l + \mathbf{Z}_{il}^T\mathbf{b}_{il}) = g_l^{-1}(m_{il})$$

and

$$E(y_{il}) = E_{\mathbf{b}_{il}}(E(y_{il}|\mathbf{b}_{il})) = E\left(g_l^{-1}(m_{il})\right).$$

Thus the marginal model can be considered as

$$y_{il} = g_l^{-1}(\mathbf{X}_{il}^T\boldsymbol{\beta}_l + \mathbf{Z}_{il}^T\mathbf{b}_{il}) + \epsilon_i.$$

Following Wolfinger and O'connell (1993), take first order Taylor series of $g_l^{-1}(m_{il})$ about $\tilde{\boldsymbol{\beta}}_l$ and $\tilde{\mathbf{b}}_{il}$ yields

$$y_{il} \approx g_l^{-1}(\tilde{m}_{il}) - \tilde{\Delta}_l(\mathbf{X}_{il}^T\tilde{\boldsymbol{\beta}}_l + \mathbf{Z}_{il}^T\tilde{\mathbf{b}}_{il}) + \tilde{\Delta}_l(\mathbf{X}_{il}^T\boldsymbol{\beta}_l + \mathbf{Z}_{il}^T\mathbf{b}_{il}),$$

where

$$\tilde{\Delta}_l = \left(\frac{\partial g_l^{-1}(m_{il})}{\partial m_{il}}\right)_{\tilde{\boldsymbol{\beta}}_l, \tilde{\mathbf{b}}_{il}}$$

is a diagonal matrix of derivatives of the conditional mean evaluated at $\tilde{\boldsymbol{\beta}}_l$, $\tilde{\mathbf{b}}_{il}$. Rearranging terms yields the expression

$$\tilde{\Delta}_l^{-1}(y_{il} - g_l^{-1}(\tilde{m}_{il})) + \mathbf{X}_{il}^T\tilde{\boldsymbol{\beta}}_l + \mathbf{Z}_{il}^T\tilde{\mathbf{b}}_{il} \approx \mathbf{X}_{il}^T\boldsymbol{\beta}_l + \mathbf{Z}_{il}^T\mathbf{b}_{il}.$$

Let

$$y_{il}^* \equiv \tilde{\Delta}^{-1}(g_l^{-1}(m_{il}) - g_l^{-1}(\tilde{m}_{il})) + \mathbf{X}_{il}^T\tilde{\boldsymbol{\beta}}_l + \mathbf{Z}_{il}^T\tilde{\mathbf{b}}_{il}.$$

Thus we can consider the model

$$y_{il}^* = \mathbf{X}_{il}^T\boldsymbol{\beta}_l + \mathbf{Z}_{il}^T\mathbf{b}_{il} + \epsilon_{il},$$

which is a linear mixed model with response y_{il}^*, with fixed effects $\boldsymbol{\beta}_l$, random effects $\mathbf{b}_{il}$, $Var(\epsilon_{il}) = Var(y_{il}^*|\mathbf{b}_{il}) = \tilde{\Delta}_l^{-1}Var(y_{il})\tilde{\Delta}_l^{-1}$, $V_{il}^* = Var(y_{il}^*) = \mathbf{Z}_{il}^TVar(\mathbf{b}_{il})\mathbf{Z}_{il} + Var(\epsilon_{il})$, and $\mu_{il} = E(y_{il}^*) = \mathbf{X}_{il}^T\boldsymbol{\beta}_l$. And y_{il}^* is a linear transformation of y_{il}:

$$y_{il}^* \equiv \tilde{\Delta}^{-1}(y_{il} - g_l^{-1}(\tilde{m}_{il})) + \mathbf{X}_{il}^T\tilde{\boldsymbol{\beta}}_l + \mathbf{Z}_{il}^T\tilde{\mathbf{b}}_{il}.$$

Let $\mathbf{Y}_i^* = (\mathbf{y}_{i1}^{*T}, \cdots, \mathbf{y}_{iL}^{*T})^T$ denotes corresponding linearized L-variate response vector for the i-th subject, $\mathbf{y}_{il}^* = (y_{il}^*(t_{i1}), \cdots, y_{il}^*(t_{in_i}))^T$. The longitudinal model with approximated response will be

$$\mathbf{Y}_i^* = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i.$$

Then $E(\mathbf{Y}_i^*) = \mathbf{X}_i\boldsymbol{\beta}$ and the variance covariance of $\mathbf{Y}_i^*$ is $\mathbf{V}_i^* = Cov(\mathbf{Y}_i^*) = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^T + \boldsymbol{\Sigma}_i$, where $\mathbf{X}_i = \begin{pmatrix} \mathbf{X}_{i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{i2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{X}_{iL} \end{pmatrix}_{(\sum_{j=1}^L(p_j+1))\times(\sum_{j=1}^L(p_j+1))}$,

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_L \end{pmatrix}_{(\sum_{j=1}^{L}(p_j+1))\times 1}, \mathbf{Z}_i = \begin{pmatrix} \mathbf{Z}_{i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{i2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{Z}_{iL} \end{pmatrix},$$

$$\mathbf{b}_i = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_L \end{pmatrix} \text{ with covariane matrix } \mathbf{G} = \begin{pmatrix} \mathbf{G}_1 & \mathbf{G}_{12} & \cdots & \mathbf{G}_{1L} \\ \mathbf{G}_{21} & \mathbf{G}_2 & \cdots & \mathbf{G}_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}_{L1} & \cdots & \mathbf{G}_{L(L-1)} & \mathbf{G}_L \end{pmatrix},$$

$$\boldsymbol{\epsilon}_i = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_L \end{pmatrix} \text{ represent independent random errors, with covariance matrix}$$

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma_1^2\mathbf{I}_{n_i} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sigma_2^2\mathbf{I}_{n_i} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \sigma_L^2\mathbf{I}_{n_i} \end{pmatrix}, \text{ where } \mathbf{I}_{n_i} \text{ is } n_i \times n_i \text{ identity matrix.}$$

4.2.2 The Joint Generalized Estimating Equation Method

Let $\mathbf{W}_i$ be the diagonal weight matrix, $\mathbf{U}(\boldsymbol{\beta})$, $\mathbf{D}(\mathbf{b})$, $L(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ are objective functions for solving $\boldsymbol{\beta}$, $\mathbf{b}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ respectively. We propose a joint generalized estimating

equation (JGEE) approach for estimation, which is to solve the following estimating equations:

Estimate $\boldsymbol{\beta}$ from:

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i^{*-1} \mathbf{W}_i (\mathbf{Y}_i^* - \mathbf{X}_i \boldsymbol{\beta}) = \mathbf{0}. \tag{4.3}$$

Predict random effects $\mathbf{b}$ from:

$$\mathbf{b} = \underset{\mathbf{b}}{\operatorname{argmin}} \sum_{i=1} (\mathbf{Y}_i^* - \mathbf{X}_i \boldsymbol{\beta} - \mathbf{Z}_i \mathbf{b}_i)^T \mathbf{V}_i^{*-1} (\mathbf{Y}_i^* - \mathbf{X}_i \boldsymbol{\beta} - \mathbf{Z}_i \mathbf{b}_i). \tag{4.4}$$

Estimate risk coefficients $\boldsymbol{\alpha}$ from:

$$\boldsymbol{\alpha} = \underset{\boldsymbol{\alpha},\boldsymbol{\gamma}}{\operatorname{argmax}} \prod_{i=1}^{N} \left[\frac{e^{\mathbf{u}_i^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l(m_{il}(T_i))}}{\sum_{j=1}^{N} I(T_j \geq T_i) e^{\mathbf{u}_j^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l(m_{jl}(T_j))}} \right]^{\delta_i}. \tag{4.5}$$

Where equation (4.3) is a weighted GEE for longitudinal outcomes. In equation (4.3), $\mathbf{W}_i$ is an $(L \times n_i) \times (L \times n_i)$ diagonal weight matrix with the j-th diagonal element as

$$W(t_{ij}; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = \frac{c(t_{ij}; \mathbf{u}_i, \boldsymbol{\gamma}_0)}{h_0(t_{ij}) \exp\left\{ \mathbf{u}_i^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l(m_{il}(t_{ij}-)) \right\}},$$

This weight is called inverse visit intensity weight motivated by Lin et al. (2004) and Buzkova and Lumley (2007), and it is used to account for missing caused by time-to-event. $c(\cdot)$ is a pre-specified function of the time-independent covriates $\mathbf{u}$ and has the potential to stabilize the influence of small values in the denominator. In addition, if we choose $c(\cdot)$ as $h_0(t_{ij}) \exp\left\{ \mathbf{u}_i^T \boldsymbol{\gamma}_0 \right\}$, when $y_i(t)$'s have no influence on the risk of

event at time t, the weight becomes identity matrix reducing to the regular GEE for equation (4.3).

Equation (4.4) minimizes the squared distance between approximate individual's response to the population average. $\mathbf{b}$ can be predicted as minimizing the squared length of residual vector:

$$\mathbf{b} = \underset{\mathbf{b}}{\operatorname{argmin}}(\mathbf{Y}^* - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})^T \mathbf{V}_i^{*-1} (\mathbf{Y}^* - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}).$$

Thus $\mathbf{b}$ is the generalized least squares estimates with explicit form

$$\hat{\mathbf{b}} = \left(\mathbf{Z}^T \mathbf{V}_i^{*-1} \mathbf{Z}\right)^{-1} \mathbf{Z}^T \mathbf{V}_i^{*-1} \left(\mathbf{Y}^* - \mathbf{X}\boldsymbol{\beta}\right).$$

With $\boldsymbol{\beta}$ is replaced by the consistent estimate $\hat{\boldsymbol{\beta}}$.

Equation (4.5) is the partial likelihood function for Cox proportional hazard model with unobserved longitudinal measures replaced by estimated longitudinal measures from equations (4.3) and (4.4)

$$\hat{m}_{il}(t_{ij}) = \mathbf{X}_i^T(t_{ij})\hat{\boldsymbol{\beta}}_l + \mathbf{Z}_i^T(t_{ij})\hat{b}_{il}^*,$$

where $\hat{\mathbf{b}}_{il}^*$ is an adjusted random effect based on the estimate from (4.5). The corresponding log partial likelihood function is :

$$l(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \sum_{i=1}^{N} \delta_i \left[\mathbf{u}_i^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f(\hat{m}_{il}(T_i)) \right.$$
$$\left. - \log \sum_{j=1}^{N} I(T_j \geq T_i) \exp\{\mathbf{u}_j^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f(\hat{m}_{jl}(T_j))\} \right].$$

Coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ can be obtained by maximizing this log partial likelihood.

Note that the weight $\mathbf{W}_i$'s in equation (4.3), and $m_{il}(T_i)$'s in equation (4.5) depend on the true model parameters, which are unknown. Equations (4.3), (4.4) and (4.5) can be solved iteratively with $\mathbf{W}_i$, $m_{il}(T_i)$ replaced by their estimators. Initial value of $\mathbf{W}_i$ can be obtained from a naive two-stage or GEE method; equation (4.3) can provide an estimator of fixed effects $\hat{\boldsymbol{\beta}}$; given $\hat{\boldsymbol{\beta}}$, equation (4.4) can provide a prediction of random effects $\hat{\mathbf{b}}$; from equation (4.3) and (4.4), an estimator of m can be obtained and $f_l(\hat{m}_{il}(t)) = f_l(\mathbf{X}_{il}^T(t)\hat{\boldsymbol{\beta}}_l + \mathbf{Z}_{il}^T(t)\hat{\mathbf{b}}_{il})$ can be plugged into equation (4.5). The three equations can be solved iteratively.

The variance of $\hat{\boldsymbol{\beta}}$ will be calculated as:

$$\begin{aligned} &\left(\mathbf{X}^T\mathbf{V}^{*-1}\mathbf{W}\mathbf{X}\right)^{-1} \\ &\left(\mathbf{X}^T\mathbf{V}^{*-1}\mathbf{W}(\mathbf{Y}^* - \mathbf{X}\boldsymbol{\beta})(\mathbf{Y}^* - \mathbf{X}\boldsymbol{\beta})^T\mathbf{W}\mathbf{V}^{*-1}\mathbf{X}\right) \\ &\left(\mathbf{X}^T\mathbf{V}^{*-1}\mathbf{W}\mathbf{X}\right)^{-1}, \end{aligned} \tag{4.6}$$

with $\mathbf{V}^*$, $\boldsymbol{\beta}$ and $\mathbf{W}$ are replaced by estimates.

The variance estimate of $\hat{\boldsymbol{\alpha}}$ will be obtained by inverse of the information matrix of the log partial likelihood:

$$V(\hat{\boldsymbol{\alpha}}) = -\left[\frac{\partial^2 l(\hat{\boldsymbol{\alpha}})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}\right]^{-1} \tag{4.7}$$

The proposed approach uses an iterative computational approach and avoids high dimensional integration required in likelihood or Bayesian methods. The proposed method can be implemented in standard statistical software. Because the standard errors of fixed effects in the longitudinal model are calculated based on modified

GEE with linearized longitudinal responses as if the weights are known, and the standard errors of the risk coefficient estimators are obtained through partial likelihood assuming that all the true covariate values are known, estimated standard errors are likely to be underestimating the true variance. In the following section, we discuss the large sample properties of the proposed estimators.

4.3 Consistency of Estimators

4.3.1 Zero mean of the estimating function for $\boldsymbol{\beta}$

The regression parameters $\boldsymbol{\beta}$ are identified under estimating equation (4.3) and survival assumption (4.2) and the fact that

$$E\left[\mathbf{U}(\boldsymbol{\beta})\right] = E\left[\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i^{*-1} \mathbf{W}_i (\mathbf{Y}_i^* - \mathbf{X}_i \boldsymbol{\beta})\right] = \mathbf{0}. \tag{4.8}$$

Let $N_i(t) = \sum_{i=1}^{n_i} I[T_i^* \leq t]$ be the number of observations of the ith subject by time t. $dN_i(t)$ denote the indicator for the ith individual being observed to have an event in $[t, t+dt)$. Let $\xi_i(t) = I(T_i^* \geq t)$ indicate whether subject i is still at risk at time t. From survival assumption (4.2):

$$P\left\{dN_i(t)\right\} = \xi_i(t) h(t) dt.$$

Suppose the hazard rate at time t is not related to the timing of the visits prior to t. Assume $\mathbf{V}_i^{*-1}$ is given, $\mathbf{W}_i$ is known, and the approximated $\mathbf{Y}_i^*$ are true

and known.

$$
\begin{aligned}
E\left[\mathbf{U}(\boldsymbol{\beta})\right] &= E_{\mathbf{Y}^*,\mathbf{T}}\left[\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{V}_i^{*-1}\int_0^\tau \frac{c(t)}{h(t)}dN_i(t)(\mathbf{Y}_i^*-\mathbf{X}_i\boldsymbol{\beta})\right] \\
&= E_{\mathbf{Y}^*}\left\{E_{\mathbf{T}|\mathbf{Y}^*}\left[\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{V}_i^{*-1}\int_0^\tau \frac{c(t)}{h(t)}dN_i(t)(\mathbf{Y}_i^*-\mathbf{X}_i\boldsymbol{\beta})\middle|\ \mathbf{Y}^*\right]\right\} \\
&= E_{\mathbf{Y}^*}\left\{\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{V}_i^{*-1}E_{\mathbf{T}|\mathbf{Y}^*}\left[\int_0^\tau \frac{c(t)}{h(t)}dN_i(t)\middle|\ \mathbf{Y}^*\right](\mathbf{Y}_i^*-\mathbf{X}_i\boldsymbol{\beta})\right\} \\
&= E_{\mathbf{Y}^*}\left\{\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{V}_i^{*-1}E_m\left[E_{\mathbf{T}|\mathbf{Y}^*,m}\left[\int_0^\tau \frac{c(t)}{h(t)}dN_i(t)\middle|\ m_i(t^-),\mathbf{Y}^*\right]\right](\mathbf{Y}_i^*-\mathbf{X}_i\boldsymbol{\beta})\right\} \\
&= E_{\mathbf{Y}^*}\left\{\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{V}_i^{*-1}E_m\left[\int_0^\tau \frac{c(t)}{h(t)}E_{\mathbf{T}|\mathbf{Y}^*,m}\left[dN_i(t)\middle|\ m_i(t^-),\mathbf{Y}^*\right]\right](\mathbf{Y}_i^*-\mathbf{X}_i\boldsymbol{\beta})\right\} \\
&= E_{\mathbf{Y}^*}\left\{\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{V}_i^{*-1}\int_0^\tau \frac{c(t)}{h(t)}\xi_i(t)h(t)dt(\mathbf{Y}_i^*-\mathbf{X}_i\boldsymbol{\beta})\right\} \\
&= E_{\mathbf{Y}^*}\left\{\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{V}_i^{*-1}\int_0^\tau c(t)\xi_i(t)dt(\mathbf{Y}_i^*-\mathbf{X}_i\boldsymbol{\beta})\right\} \\
&= \sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{V}_i^{*-1}\int_0^\tau c(t)\xi_i(t)dtE_{\mathbf{Y}^*}\left[\mathbf{Y}_i^*-\mathbf{X}_i\boldsymbol{\beta}\right] \\
&= \mathbf{0}.
\end{aligned}
$$

Consistent estimates of $\boldsymbol{\beta}$ can be obtained as the solution of equation (4.3) with the weight $\mathbf{W}_i$ be substituted by suitable consistent estimators.

4.3.2 Zero mean of the estimating function for $\boldsymbol{\alpha}$

Equation (4.5) corresponds to maximizing the logarithm of the partial likelihood:

$$
\begin{aligned}
l(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \delta_i \Bigg[& \mathbf{u}_i^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l(m_{il}(T_i)) \\
& - \log \sum_{j=1}^{N} I(T_j \geq T_i) \exp \left\{ \mathbf{u}_j^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l(m_{jl}(T_j)) \right\} \Bigg] .
\end{aligned}
\tag{4.9}
$$

Setting the derivative with respect to the parameters provides the score estimating equations $S(\boldsymbol{\alpha})$, the score equation corresponding to parameter α_l is:

$$
\begin{aligned}
S(\alpha_l) = \frac{\partial l(\boldsymbol{\alpha})}{\partial \alpha_l} &= \sum_{i=1}^{N} \delta_i \Bigg[f_l(m_{il}(T_i)) \\
&\quad - \frac{\sum_j I(T_j \geq T_i) \exp\{\mathbf{u}_j^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l(m_{jl}(T_i))\} f_l(m_{jl}(T_i))}{\sum_j I(T_j \geq T_i) \exp\{\mathbf{u}_j^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l(m_{jl}(T_i))\}} \Bigg] . \\
&= \sum_{i=1}^{N} \delta_i \left[f_l(m_{il}(T_i)) - \sum_j f_l(m_{jl}(T_i)) p_j \right] ,
\end{aligned}
$$

where

$$
p_j = \frac{I(T_j \geq T_i) \exp\{\mathbf{u}_j^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l(m_{jl}(T_i))\}}{\sum_j I(T_j \geq T_i) \exp\{\mathbf{u}_j^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l(m_{jl}(T_i))\}}
$$

could be considered as proportional to the hazard of the individual failing.

The regression parameters $\boldsymbol{\alpha}$ are identified under estimating equation (4.3) and survival assumption (4.2) and the fact that $E\left[S(\boldsymbol{\alpha})\right] = \mathbf{0}$. Condition on true longitudinal process m_{jl}'s are known, suppose function f is known, thus covariates $f_l(m_{il})$ are known.

$$
\begin{aligned}
E\left[S(\alpha_l)\right] &= E\left[\sum_{i=1}^{N}\delta_i\left(f_l(m_{il}(T_i)) - \sum_j f_l(m_{jl}(T_i))p_j\right)\right] \\
&= \sum_{i=1}^{N} E\left[\delta_i\left(f_l(m_{il}(T_i)) - \sum_j f_l(m_{jl}(T_i))p_j\right)\right] \\
&= \sum_{i=1}^{N} E\left[E\left[\delta_i\left(f_l(m_{il}(T_i)) - \sum_j f_l(m_{jl}(T_i))p_j\right)|f, m_i(t^-), \mathbf{u}_i, \boldsymbol{\delta}_i\right]\right] \\
&= \sum_{i=1}^{N} E\left[\delta_i\left(E\left[f_l(m_{il}(T_i))|f, m_i(t^-), \mathbf{u}_i\right] - \sum_j f_l(m_{jl}(T_i))p_j\right)\right] \\
&= \sum_{i=1}^{N} E\left[\delta_i\left(\sum_j f_l(m_{jl}(T_i))p_j - \sum_j f_l(m_{jl}(T_i))p_j\right)\right] = 0
\end{aligned}
$$

Because m_{il}'s are unknown, to obtain the estimator of $\boldsymbol{\alpha}$, m_{il}'s are substituted by their estimates $\hat{m}_{il}(t) = \mathbf{X}_{il}^T(t)\hat{\boldsymbol{\beta}}_l + \mathbf{Z}_{il}^T(t)\hat{\mathbf{b}}_{il}$, expectation of $S(\boldsymbol{\alpha}, \hat{\mathbf{m}})$ is no longer zero at $\boldsymbol{\alpha}$. Uncertainty of estimates $\hat{\mathbf{b}}_{il}$ relative to true values of $\mathbf{b}_{il}$ can be considered as subject to measurement error. Assume $\mathbf{b}$ is normally distributed, suppose $\hat{\mathbf{b}}$ are subject to additive and normal measurement error $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Lambda})$, $\hat{\mathbf{b}}_{il} = \mathbf{b}_{il} + \boldsymbol{\epsilon}_{il}$. Then $\hat{m}_{il}(t) = \mathbf{X}_{il}^T(t)\hat{\boldsymbol{\beta}}_l + \mathbf{Z}_{il}^T(t)\mathbf{b}_{il} + \mathbf{Z}_{il}^T(t)\epsilon_{il}$, consider the measurement error of $\hat{m}_{il}(t)$ relative to $m_{il}(t)$ is $\mathbf{Z}_{il}^T(t)\epsilon_{il}$. Nakamura (1990) have proposed a corrected score function method and Nakamura (1992) applied it to the proportional hazards model that allows a simple correction to the ordinary partial likelihood.

Suppose f is identity function (the hazard for an event depends on the effect of the true underlying longitudinal outcomes), the corrected score function following

Nakamura's method is

$$\tilde{S}(\boldsymbol{\alpha}, \hat{\mathbf{m}}) = S(\boldsymbol{\alpha}, \hat{\mathbf{m}}) + \mathbf{Z}^T \boldsymbol{\Lambda} \mathbf{Z} \boldsymbol{\alpha}.$$

Nakamura's corrected maximum partial likelihood estimates(MPLE) is $\hat{\boldsymbol{\alpha}}$ that satisfies

$$\tilde{S}(\boldsymbol{\alpha}, \hat{\mathbf{m}}) = \mathbf{0}.$$

In the special case of no measurement error, Nakamura's corrected MPLE becomes a regular MPLE. Theoretical justification can be found in the paper by Kong and Gu (1999). The authors have shown that the corrected MPLE is consistent and asymptotically normally distributed.

4.4 Asymptotic Properties of Estimators

In this section, we derive the large sample behavior of the JGEE estimators. The property of $\hat{\boldsymbol{\beta}}$ is under the assumption that the number of independent subjects goes to infinity and the number of observations are finite with an upper bound, and the weights are known in JGEE. The property of $\hat{\boldsymbol{\alpha}}$ is under the assumption that the number of independent subjects goes to infinity and the longitudinal covariates are known without random error.

4.4.1 Asymptotic properties of $\hat{\boldsymbol{\beta}}$

Assuming the weights are known

For general estimating equations, Liang and Zeger (1986) have proposed a sandwich variance covariance matrix estimator for $\hat{\boldsymbol{\beta}}$, the variance covariance estimator is consistent even if the working covariance matrix is not true covariance matrix of $\mathbf{Y}_i$. If the covariance matrix of $\mathbf{Y}_i$ is correctly specified, a consistent estimator for covariance matrix of $\hat{\boldsymbol{\beta}}$ is model based or naive covariance estimator. Following the sandwich variance covariance of GEEs, the robust sandwich variance covariance matrix of $\hat{\boldsymbol{\beta}}$ from (4.3) will be similar when the weight is considered as known.

The large sample distribution of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and variance covariance matrix follows from the asymptotic normality of $\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i^{*-1} \mathbf{W}_i (\mathbf{Y}_i^* - \mathbf{X}_i \boldsymbol{\beta})$ in (4.3) and a series of Taylor expansions. Take the first two terms of the Taylor expansion for $\mathbf{U}(\boldsymbol{\beta})$ at estimator $\hat{\boldsymbol{\beta}}$:

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{U}(\hat{\boldsymbol{\beta}}) + \frac{\partial \mathbf{U}^T}{\partial \boldsymbol{\beta}} |_{\boldsymbol{\beta}^*} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}),$$

where $\boldsymbol{\beta}^*$ is on the line segment between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. Then

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(-\frac{1}{N} \frac{\partial \mathbf{U}^T}{\partial \boldsymbol{\beta}} |_{\boldsymbol{\beta}^*} \right)^{-1} \frac{1}{\sqrt{N}} \mathbf{U}(\boldsymbol{\beta}).$$

By the strong law of large numbers, $-\frac{1}{N}\frac{\partial \mathbf{U}^T}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}^*}$ converges to fisher information matrix. That is

$$-\frac{1}{N}\frac{\partial \mathbf{U}^T}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}^*} = -E\left(\frac{\partial \mathbf{U}^T}{\partial \boldsymbol{\beta}}\right) = \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i^{*-1} \mathbf{W}_i \mathbf{X}_i \overset{\text{defined as}}{=} A(\boldsymbol{\beta}).$$

And

$$\begin{aligned} E\left[\mathbf{U}(\boldsymbol{\beta})\right] &= E\left[\mathbf{X}_i^T \mathbf{V}_i^{*-1} \mathbf{W}_i (\mathbf{Y}_i^* - \mathbf{X}_i \boldsymbol{\beta})\right] \\ &= \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i^{*-1} \mathbf{W}_i \left(E(\mathbf{Y}_i^*) - \mathbf{X}_i \boldsymbol{\beta}\right) = \mathbf{0}. \end{aligned}$$

In addition,

$$\begin{aligned} Var\left(\frac{1}{\sqrt{N}}\mathbf{U}(\boldsymbol{\beta})\right) &= \frac{1}{N}\sum_{i=1}^{N} Var\left(\mathbf{U}_i(\boldsymbol{\beta})\right) = \frac{1}{n}\sum_{i=1}^{n} E\left[\mathbf{U}_i(\boldsymbol{\beta})\mathbf{U}_i^T(\boldsymbol{\beta})\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} E\left[\mathbf{X}_i^T \mathbf{V}_i^{*-1} \mathbf{W}_i (\mathbf{Y}_i^* - \mathbf{X}_i\boldsymbol{\beta})(\mathbf{Y}_i^* - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{W}_i \mathbf{V}_i^{*-1} \mathbf{X}_i\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i^{*-1} \mathbf{W}_i (\mathbf{Y}_i^* - \mathbf{X}_i\hat{\boldsymbol{\beta}})(\mathbf{Y}_i^* - \mathbf{X}_i\hat{\boldsymbol{\beta}})^T \mathbf{W}_i \mathbf{V}_i^{*-1} \mathbf{X}_i \\ &\overset{\text{defined as}}{=} B(\boldsymbol{\beta}), \end{aligned}$$

where $(\mathbf{Y}_i^* - \mathbf{X}_i\hat{\boldsymbol{\beta}})(\mathbf{Y}_i^* - \mathbf{X}_i\hat{\boldsymbol{\beta}})^T$ is the empirical estimation of $Cov(\mathbf{Y}_i^* - \mathbf{X}_i\boldsymbol{\beta})$. By central limit theorem:

$$\frac{1}{\sqrt{N}}\mathbf{U}(\boldsymbol{\beta}) \sim N(\mathbf{0}, B(\boldsymbol{\beta})).$$

Therefore, condition on the weights $\mathbf{W}$ are known, as $N \longrightarrow \infty$,

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim N(\mathbf{0}, A(\boldsymbol{\beta})^{-1} B(\boldsymbol{\beta}) A(\boldsymbol{\beta})^{-1}),$$

where $A(\boldsymbol{\beta})$ and $B(\boldsymbol{\beta})$ are defined above.

4.4.2 Asymptotic properties of $\hat{\boldsymbol{\alpha}}$

Assuming the longitudinal covariates are known

Risk coefficients are estimated by maximizing the following log partial likelihood function:

$$l(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \sum_{i=1}^{N} \delta_i \left[\mathbf{u}_i^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l(m_{il}(T_i)) \right. \\ \left. - \log \sum_j I(T_j \geq T_i) \exp\{\mathbf{u}_j^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l(m_{jl}(T_i))\} \right]. \tag{4.10}$$

Conditional on m_{il}'s are known, and function $f_l(\cdot)$ are known, then $f_l(m_{il})$ can be consider as known covariates. Regression parameter estimators $\hat{\boldsymbol{\alpha}}$ obtained from solving the score estimating equations (4.10) will yield consistent estimators of true $\boldsymbol{\alpha}$. The statistic $\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$ converge in distribution to normal random variables with mean $\mathbf{0}$ and variance covariance matrix $\boldsymbol{\Sigma}$. The asymptotic properties of maximum partial likelihood estimator holds following studies such as Tsiatis (1981), Andersen and Gill (1982), Næs (1982), Bailey (1983) and Lin and Wei (1989).

Setting derivative of the logarithm of the partial likelihood (4.10) with respect to the parameters to zero yields the score equations, the parameter estimator $\hat{\alpha}_l$ is obtained by solving:

$$
\begin{aligned}
S(\alpha_l) = \frac{\partial l(\boldsymbol{\alpha})}{\partial \alpha_l} &= \sum_{i=1}^{N} \delta_i \left[f_l(m_{il}(T_i)) \right. \\
&\left. - \frac{\sum_j I(T_j \geq T_i) \exp\{\mathbf{u}_j^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l(m_{jl}(T_i))\} f_l(m_{jl}(T_i))}{\sum_j I(T_j \geq T_i) \exp\{\mathbf{u}_j^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l(m_{jl}(T_i))\}} \right]. \qquad (4.11) \\
&= \sum_{i=1}^{N} \delta_i \left[f_l(m_{il}(T_i)) - \frac{p^{(1)}(\alpha_l, T_i)}{p^{(0)}(\alpha_l, T_i)} \right] = 0,
\end{aligned}
$$

where

$$
p^{(r)}(\alpha_l, T) = \sum_j I(T_j \geq T) \exp \left\{ \mathbf{u}_j^T \boldsymbol{\gamma} + \sum_{l=1}^{L} \alpha_l f_l(m_{jl}(T)) \right\} f_l(m_{jl}(T))^r.
$$

Take second derivatives:

$$
A_{ll}(\boldsymbol{\alpha}) = \frac{\partial^2 l}{\partial \alpha_l \partial \alpha_l} = \sum_{i=1}^{N} \delta_i \left[\frac{p^{(2)}(\alpha_l, T_i)}{p^{(0)}(\alpha_l, T_i)} - \left\{ \frac{p^{(1)}(\alpha_l, T_i)}{p^{(0)}(\alpha_l, T_i)} \right\}^{\otimes 2} \right].
$$

$$
A_{lh}(\boldsymbol{\alpha}) = \frac{\partial^2 l}{\partial \alpha_l \partial \alpha_h} = \sum_{i=1}^{N} \delta_i \left[\frac{p^{(1)}(\alpha_l, T_i) f_h(m_{jh}(T_i))}{p^{(0)}(\alpha_l, T_i)} - \frac{p^{(1)}(\alpha_l, T_i)}{p^{(0)}(\alpha_l, T_i)} \frac{p^{(1)}(\alpha_h, T_i)}{p^{(0)}(\alpha_l, T_i)} \right].
$$

Denote the second derivatives matrix as $A(\boldsymbol{\alpha})$, with diagonal elements $A_{ll}(\boldsymbol{\alpha})$ and non-diagonal elements $A_{lh}(\boldsymbol{\alpha})$ defined above. The Fisher information matrix can be written as:

$$
\mathbf{I}(\boldsymbol{\alpha}) = -E(A(\boldsymbol{\alpha})).
$$

The variance covariance matrix of $\hat{\boldsymbol{\alpha}}$ with the score estimating equation $S(\hat{\boldsymbol{\alpha}}) = \mathbf{0}$ is asymptotically $\mathbf{I}(\hat{\boldsymbol{\alpha}})^{-1}$, which is inverse negative the expectation of the second derivative of Cox's partial likelihood. The standard errors calculated based on partial

likelihood with estimated longitudinal outcomes are likely to be biased and tend to be smaller than the true variance of these risk coefficent estimators. Because this method assumes all the longitudinal covariates as known, it does not take into account the uncertainty of those random effects carried by the covariates.

Chapter 5

Conclusion

In this thesis, we studied joint models for multiple longitudinal responses and their relation to the survival outcome. We started from an approach to handle multiple longitudinal continuous responses and a time-to-event outcome in Chapter 2. In Chapter 3, we extended the method to handle multiple longitudinal Poisson responses and a time-to-event outcome. In Chapter 4, we offered a general formulation to handle various longitudinal response and a time-to-event outcome. The proposed methodology is applicable to a number of medical research areas.

Joint models for multiple continuous longitudinal biomarkers and a time-to-event outcome were examined in Chapter 2. We developed a Joint Generalized Estimating Equation (JGEE) approach for parameter estimation. We investigated the finite sample behavior of the estimators with a series of simulations. The simulation compared the naive two-stage method, the GEE approach (same estimating equations as JGEE but with identity weight) to the JGEE method. The results demonstrated that the proposed JGEE method can provide more accurate parameter estimates than the two-stage or the GEE method. Our approach was also illustrated with a data from a longitudinal cohort to examine the association of longitudinally collected low density lipoprotein (LDL), high density lipoprotein (HDL) measures and their association with the risk of coronary artery disease (CAD).

In Chapter 3, we extended the proposed JGEE method to multiple longitudinal count data and a time-to-event outcome. The performance of the propsed JGEE method was assessed through extensive simulation studies. We have shown that the proposed JGEE method yields adequate risk effect estimates. The proposed method was applied to data from a longitudinal cohort to examine the association between longitudinal collected cognitive memory score and cognitive executive score and their association with the risk of coronary artery disease (CAD).

In Chapter 4, we introduced a general formulation of joint models to accommodate different types of longitudinal responses and a time-to-event outcome. The JGEE method is further extended to handle the general joint models. And the large sample properties of the parameter estimators are investigated.

The propose JGEE method has several advantages. First, the JGEE method can easily handle mixed types of longitudinal processes, which comes naturally from the feature of generalized estimating equations. Second, the JGEE method is computationally efficient to carry out as it avoids the direct maximization of the joint likelihood, thus it can be implemented in standard software. In comparison, maximum likelihood with EM algorithm and Bayesian approach can be extremely computationally intensive because of the integration over multiple random effects. Third, JGEE method reduces biases from the naive two-stage method by adjusting for informative truncation from the time-to-event outcome with inverse intensity weighting.

The current work also has some limitations and further development is necessary. Through simulation studies, we observed that although the JGEE method reduces biased from the naive two-stage method, but the proposed estimators can still be biased, especially in non-normal situations. We theorize that the bias of the

JGEE estimator arises from the non-identical functional form for the marginal mean from the conditional mean in the non-normal setting. We know in the case of log linear model for Poisson data, the marginal mean and the conditional mean are shifted by an offset. Thus the first estimating equation in the JGEE will need to be adjusted to correctly reflect this shift in order to produce parameter estimates that correspond to those specified in the longitudinal model.

In addition, additional work is also needed to refine the prediction of longitudinal components to be used in the estimating equation based on the partial likelihood. Moreover, because the standard errors of fixed effects in longitudinal model are calculated based on modified GEE as if the weights are known, and the standard errors of the risk coefficient estimators are obtained through partial likelihood assuming that all the true covariate values are known, estimated standard errors are underestimating the true variance. Further theoretical work needs to be conducted for improved variance estimations based on the joint modeling framework.

There are also many potential extensions to our joint modeling framework. For survival sub-models, the current work focuses on one time-to-event with right censoring. More complicated scenarios such as multiple events, recurrent events, competing risks, or interval censoring can be considered. For longitudinal sub-models, one can consider capturing the longitudinal trajectories more flexibly with a smoothing spline models. Another topic that this work did not address is procedures for model selection, especially in settings where many potential covariates need to be considered in the joint models.

In summary, joint models are powerful analytical tools for longitudinal and time-to-event data and widely applicable to medical research. Our proposed method

offers a computationally efficient method for parameter estimation from a number of longitudinal processes and can increase the utility of joint models in solving real world problems.

BIBLIOGRAPHY

Albert, P. S. and J. H. Shih (2010). On estimating the relationship between longitudinal measurements and time-to-event data using a simple two-stage procedure. *Biometrics 66*(3), 983–987.

Andersen, P. K. and R. D. Gill (1982). Cox's regression model for counting processes: a large sample study. *The annals of statistics*, 1100–1120.

Asar, O. and O. Ilk (2013). mmm: an r package for analyzing multivariate longitudinal data with multivariate marginal models. *Computer Methods and Programs in Biomedicine 112*(3), 649–654.

Augustin, T. and R. Schwarz (2002). Cox's proportional hazards model under covariate measurement error. In *Total Least Squares and Errors-in-Variables Modeling*, pp. 179–188. Springer.

Bailey, K. R. (1983). The asymptotic joint distribution of regression and survival parameter estimates in the cox regression model. *The Annals of Statistics*, 39–48.

Borowicz, L. M., M. A. Goldsborough, O. A. Selnes, and G. M. McKhann (1996). Neuropsychologic change after cardiac surgery: a critical review. *Journal of cardiothoracic and vascular anesthesia 10*(1), 105–112.

Brown, E. R. and J. G. Ibrahim (2003). Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. *Biometrics 59*(3), 686–693.

Brown, E. R., J. G. Ibrahim, and V. DeGruttola (2005). A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics 61*(1), 64–73.

Buzkova, P. and T. Lumley (2007). Longitudinal data analysis for generalized linear models with follow-up dependent on outcome-related variables. *Canadian Journal of Statistics 35*(4), 485–500.

Bycott, P. and J. Taylor (1998). A comparison of smoothing techniques for cd4 data measured with error in a time-dependent cox proportional hazards model. *Statistics in medicine 17*(18), 2061–2077.

Cannon, J. A., J. J. McMurray, and T. J. Quinn (2015). 'hearts and minds': association, causation and implication of cognitive impairment in heart failure. *Alzheimer's research & therapy 7*(1), 22.

Carroll, R., D. Ruppert, and L. Stefanski (1995). Nonlinear measurement error models. *Monographs on Statistics and Applied Probability.(Chapman and Hall, New York) Volume 63*.

Chi, Y.-Y. and J. G. Ibrahim (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics 62*(2), 432–445.

Choi, J., J. Cai, D. Zeng, and A. F. Olshan (2015). Joint analysis of survival time and longitudinal categorical outcomes. *Statistics in biosciences 7*(1), 19–47.

Dafni, U. G. and A. A. Tsiatis (1998). Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, 1445–1462.

De Gruttola, V. and X. M. Tu (1994). Modelling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics*, 1003–1014.

Dunson, D. B. and A. H. Herring (2005). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics 6*(1), 11–25.

Faucett, C. L. and D. C. Thomas (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a gibbs sampling approach. *Statistics in medicine 15*(15), 1663–1685.

Fieuws, S., G. Verbeke, B. Maes, and Y. Vanrenterghem (2008). Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics 9*(3), 419–431.

Ghisletta, P., J. J. McArdle, and U. Lindenberger (2006). Longitudinal cognition-survival relations in old and very old age: 13-year data from the berlin aging study. *European Psychologist 11*(3), 204–223.

Gorfine, M., L. Hsu, and R. L. Prentice (2004). Nonparametric correction for covariate measurement error in a stratified cox model. *Biostatistics 5*(1), 75–87.

Gueorguieva, R. and G. Sanacora (2006). Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Statistics in medicine 25*(8), 1307–1322.

Hall, K. S., S. Gao, O. Baiyewu, K. A. Lane, O. Gureje, J. Shen, A. Ogunniyi, J. R. Murrell, F. W. Unverzagt, J. Dickens, et al. (2009). Prevalence rates for dementia and alzheimer's disease in african americans: 1992 versus 2001. *Alzheimer's & Dementia 5*(3), 227–233.

Hall, K. S., S. Gao, C. L. Emsley, A. O. Ogunniyi, O. Morgan, and H. C. Hendrie (2000). Community screening interview for dementia (csi 'd'); performance in five disparate study sites. *International journal of geriatric psychiatry 15*(6), 521–531.

Hall, K. S., A. O. Ogunniyi, H. C. Hendrie, and B. O. Osuntokun (1996). Indianapolis, usa, and ibadan, nigeria. *International Journal of Methods in Psychiatric Research 6*, 129–142.

Hall, K. S., A. O. Ogunniyi, H. C. Hendrie, B. O. Osuntokun, S. L. Hui, B. S. Musick, C. A. Rodenberg, F. W. Unverzagt, O. Guerje, and O. Baiyewu (1996). A cross-cultural community based study of dementias. *International Journal of Methods in Psychiatric Research 6*(3), 129–142.

Hatfield, L. A., M. E. Boye, M. D. Hackshaw, and B. P. Carlin (2012). Multilevel bayesian models for survival times and longitudinal patient-reported outcomes with many zeros. *Journal of the American Statistical Association 107*(499), 875–885.

He, B. and S. Luo (2016). Joint modeling of multivariate longitudinal measurements and survival data with applications to parkinson's disease. *Statistical methods in medical research 25*(4), 1346–1358.

Henderson, C. R. (1963). Selection index and expected genetic advance. *Statistical genetics and plant breeding 982*, 141–163.

Hendrie, H. C., A. Ogunniyi, K. S. Hall, O. Baiyewu, F. W. Unverzagt, O. Gureje, S. Gao, R. M. Evans, A. Ogunseyinde, A. Adeyinka, et al. (2001). Incidence of dementia and alzheimer disease in 2 communities: Yoruba residing in ibadan,

nigeria, and african americans residing in indianapolis, indiana. *Jama 285*(6), 739–747.

Hickey, G. L., P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona (2016). Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Medical Research Methodology 16*(1), 117.

Hogan, J. W. and N. M. Laird (1998). Increasing efficiency from censored survival data by using random effects to model longitudinal covariates. *Statistical Methods in Medical Research 7*(1), 28–48.

Hu, C. and D. Lin (2002). Cox regression with covariate measurement error. *Scandinavian Journal of Statistics 29*(4), 637–655.

Huang, W., S. L. Zeger, J. C. Anthony, and E. Garrett (2001). Latent variable model for joint analysis of multiple repeated measures and bivariate event times. *Journal of the American Statistical Association 96*(455), 906–914.

Huang, Y., G. Dagne, and L. Wu (2011). Bayesian inference on joint models of hiv dynamics for time-to-event and longitudinal data with skewness and covariate measurement errors. *Statistics in medicine 30*(24), 2930–2946.

Huang, Y. and C. Wang (2000). Cox regression with accurate covariates unascertainable: A nonparametric-correction approach. *Journal of the American Statistical Association 95*(452), 1209–1219.

Hughes, M. D. (1993). Regression dilution in the proportional hazards model. *Biometrics*, 1056–1066.

Ibrahim, J. G., M.-H. Chen, and D. Sinha (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. *Statistica Sinica*, 863–883.

Ibrahim, J. G., H. Chu, and L. M. Chen (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology 28*(16), 2796–2801.

Kong, F. H. and M. Gu (1999). Consistent estimation in cox proportional hazards model with covariate measurement errors. *Statistica Sinica 9*(4), 953–970.

Kulich, M. and D. Lin (2000). Additive hazards regression with covariate measurement error. *Journal of the American Statistical Association 95*(449), 238–248.

Law, N. J., J. M. Taylor, and H. Sandler (2002). The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics 3*(4), 547–563.

Lawrence Gould, A., M. E. Boye, M. J. Crowther, J. G. Ibrahim, G. Quartey, S. Micallef, and F. Y. Bois (2015). Joint modeling of survival and longitudinal non-survival data: current methods and issues. report of the dia bayesian joint modeling working group. *Statistics in Medicine 34*(14), 2181–2195.

Li, Y. and L. Ryan (2006). Inference on survival data with covariate measurement error–an imputation-based approach. *Scandinavian journal of statistics 33*(2), 169–190.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Liang, K.-Y., S. L. Zeger, and B. Qaqish (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 3–40.

Lin, D. Y. and L.-J. Wei (1989). The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association 84*(408), 1074–1078.

Lin, H., C. E. McCulloch, and S. T. Mayne (2002). Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Statistics in Medicine 21*(16), 2369–2382.

Lin, H., D. O. Scharfstein, and R. A. Rosenheck (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(3), 791–813.

Lin, H., B. W. Turnbull, C. E. McCulloch, and E. H. Slate (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association 97*(457), 53–65.

Lipsitz, S. R., G. M. Fitzmaurice, J. G. Ibrahim, D. Sinha, M. Parzen, and S. Lipshultz (2009). Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: an application to acquired immune deficiency syndrome data. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 172*(1), 3–20.

Luo, S. (2014). A bayesian approach to joint analysis of multivariate longitudinal data and parametric accelerated failure time. *Statistics in medicine 33*(4), 580–594.

McCulloch, C. (2008). Joint modelling of mixed outcome types using latent variables. *Statistical Methods in Medical Research 17*(1), 53–73.

McCulloch, C. E. and J. M. Neuhaus (2001). *Generalized linear mixed models*. Wiley Online Library.

Mertens, A. and P. Holvoet (2001). Oxidized ldl and hdl: antagonists in atherothrombosis. *The FASEB journal 15*(12), 2073–2084.

Murawska, M., D. Rizopoulos, and E. Lesaffre (2012). A two-stage joint model for nonlinear longitudinal response and a time-to-event with application in transplantation studies. *Journal of Probability and Statistics 2012*.

Næs, T. (1982). The asymptotic distribution of the estimator for the regression parameter in cox's regression model. *Scandinavian Journal of Statistics*, 107–115.

Nakamura, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika 77*(1), 127–137.

Nakamura, T. (1992). Proportional hazards model with covariates subject to measurement error. *Biometrics*, 829–838.

Nelder, J. A. and R. J. Baker (1972). Generalized linear models. *Encyclopedia of statistical sciences*.

Pepe, M. S., S. G. Self, and R. L. Prentice (1989). Further results on covariate measurement errors in cohort studies with time to response data. *Statistics in Medicine 8*(9), 1167–1178.

Prentice, R. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika 69*(2), 331–342.

Proust-Lima, C., P. Joly, J.-F. Dartigues, and H. Jacqmin-Gadda (2009). Joint modelling of multivariate longitudinal outcomes and a time-to-event: a nonlinear latent class approach. *Computational statistics & data analysis 53*(4), 1142–1154.

R Brown, E. and J. G Ibrahim (2003). A bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics 59*(2), 221–228.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R.* CRC Press.

Rizopoulos, D. and P. Ghosh (2011). A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in medicine 30*(12), 1366–1380.

Rizopoulos, D., G. Verbeke, and E. Lesaffre (2009). Fully exponential laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(3), 637–654.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association 90*(429), 106–121.

Robinson, G. K. (1991). That blup is a good thing: the estimation of random effects. *Statistical science*, 15–32.

Rochon, J. (1996). Analyzing bivariate repeated measures for discrete and continuous outcome variables. *Biometrics*, 740–750.

Saxton, J., G. Ratcliff, A. Newman, S. Belle, L. Fried, J. Yee, and L. Kuller (2000). Cognitive test performance and presence of subclinical cardiovascular disease in the cardiovascular health study. *Neuroepidemiology 19*(6), 312–319.

Schatz, I. J., K. Masaki, K. Yano, R. Chen, B. L. Rodriguez, and J. D. Curb (2001). Cholesterol and all-cause mortality in elderly people from the honolulu heart program: a cohort study. *The lancet 358*(9279), 351–355.

Seaman, S. R. and I. R. White (2013). Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research 22*(3), 278–295.

Self, S. and Y. Pawitan (1992). Modeling a marker of disease progression and onset of disease. In *AIDS Epidemiology*, pp. 231–255. Springer.

Selnes, O. A., M. A. Goldsborough, L. M. Borowicz, C. Enger, S. A. Quaskey, and G. M. McKhann (1999). Determinants of cognitive change after coronary artery bypass surgery: a multifactorial problem. *The Annals of Thoracic Surgery 67*(6), 1669–1676.

Shelton, B. J., G. H. Gilbert, B. Liu, and M. Fisher (2004). A sas macro for the analysis of multivariate longitudinal binary outcomes. *Computer Methods and Programs in Biomedicine 76*(2), 163–175.

Song, X., M. Davidian, and A. A. Tsiatis (2002). An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. *Biostatistics 3*(4), 511–528.

Spiegelman, D., R. J. Carroll, and V. Kipnis (2001). Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Statistics in medicine 20*(1), 139–160.

Sweeting, M. J. and S. G. Thompson (2011). Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal 53*(5), 750–763.

Trojano, L., R. A. Incalzi, D. Acanfora, C. Picone, P. Mecocci, F. Rengo, C. I. S. Investigators, et al. (2003). Cognitive impairment: a key feature of congestive heart failure in the elderly. *Journal of neurology 250*(12), 1456–1463.

Tseng, Y.-K., F. Hsieh, and J.-L. Wang (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika 92*(3), 587–603.

Tsiatis, A., V. Degruttola, and M. Wulfsohn (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association 90*(429), 27–37.

Tsiatis, A. A. (1981). A large sample study of cox's regression model. *The Annals of Statistics*, 93–108.

Tsiatis, A. A. and M. Davidian (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 809–834.

Wang, C., J. Douglas, and S. Anderson (2002). Item response models for joint analysis of quality of life and survival. *Statistics in medicine 21*(1), 129–142.

Wang, C., L. Hsu, Z. Feng, and R. L. Prentice (1997). Regression calibration in failure time regression. *Biometrics*, 131–145.

Wang, Y. and J. M. G. Taylor (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association 96*(455), 895–905.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika 61*(3), 439–447.

Weverling-Rijnsburger, A. W., G. J. Blauw, A. M. Lagaay, D. L. Knock, A. E. Meinders, and R. G. Westendorp (1997). Total cholesterol and risk of mortality in the oldest old. *The Lancet 350*(9085), 1119–1123.

Weverling-Rijnsburger, A. W., I. J. Jonkers, E. van Exel, J. Gussekloo, and R. G. Westendorp (2003). High-density vs low-density lipoprotein cholesterol as the risk factor for coronary artery disease and stroke in old age. *Archives of internal medicine 163*(13), 1549–1554.

Wilson, P. W. (1990). High-density lipoprotein, low-density lipoprotein and coronary artery disease. *The American journal of cardiology 66*(6), A7–A10.

Wolfinger, R. and M. O'connell (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of statistical Computation and Simulation 48*(3-4), 233–243.

Wu, L., W. Liu, and X. Hu (2010). Joint inference on hiv viral dynamics and immune suppression in presence of measurement errors. *Biometrics 66*(2), 327–335.

Wu, L., W. Liu, G. Y. Yi, and Y. Huang (2011). Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. *Journal of Probability and Statistics 2012*.

Wulfsohn, M. S. and A. A. Tsiatis (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 330–339.

Xie, S. X., C. Wang, and R. L. Prentice (2001). A risk set calibration method for failure time regression by using a covariate reliability sample. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(4), 855–870.

Xu, J. and S. L. Zeger (2001a). The evaluation of multiple surrogate endpoints. *Biometrics 57*(1), 81–87.

Xu, J. and S. L. Zeger (2001b). Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 50*(3), 375–387.

Yang, L. (2013). *Joint models for longitudinal and survival data*. Ph. D. thesis, faculty of the University Graduate School in partial fulfillment of the requirements for the degree Doctor of Philosophy in the Department of Biostatistics, Indiana University.

Yang, L., M. Yu, and S. Gao (2016). Joint models for multiple longitudinal processes and time-to-event outcome. *Journal of Statistical Computation and Simulation*, 1–19.

Ye, W., X. Lin, and J. M. Taylor (2008a). A penalized likelihood approach to joint modeling of longitudinal measurements and time-to-event data. *Statistics and its Interface 1*, 33–45.

Ye, W., X. Lin, and J. M. Taylor (2008b). Semiparametric modeling of longitudinal measurements and time-to-event data–a two-stage regression calibration approach. *Biometrics 64*(4), 1238–1246.

Zeger, S. L., K.-Y. Liang, and P. S. Albert (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 1049–1060.

Zheng, L., W. J. Mack, H. C. Chui, L. Heflin, D. Mungas, B. Reed, C. DeCarli, M. W. Weiner, and J. H. Kramer (2012). Coronary artery disease is associated with cognitive decline independent of changes on magnetic resonance imaging in cognitively normal elderly adults. *Journal of the American Geriatrics Society 60*(3), 499–504.

Zorn, C. J. (2001). Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science*, 470–490.

Zucker, D. M. (2005). A pseudo–partial likelihood method for semiparametric survival regression with covariate errors. *Journal of the American Statistical Association 100*(472), 1264–1277.

Zucker, D. M. and D. Spiegelman (2008). Corrected score estimation in the proportional hazards model with misclassified discrete covariates. *Statistics in medicine 27*(11), 1911–1933.

CURRICULUM VITAE

Mengjie Zheng

EDUCATION

- PhD, Biostatistics August 2013 - July 2018

  Department of Biostatistics, Indiana University, Richard M. Fairbanks School of Public Health.

- MS, Computational Mathematics August 2011 - May 2013

  Department of Mathematics and Statistics, University of Minnesota Duluth.

- BS, Information and Computing Science August 2007 - June 2011

  Department of Mathematics, Shenyang University of Technology, Liaoning, China.

WORKING EXPERIENCE

- Biostatistician Intern, Puma Biotechnology, South San Francisco, August 2017 - May 2018
- Research Assistant, Indiana University School of Medicine, July 2014 - August 2017
- Summer Intern, Roche Diagnostics Corporation, May 2016 - July 2016
- Teaching Assistant, Indiana University Purdue University Indianapolis School of Science, August 2013 - June 2014
- Teaching Assistant, University of Minnesota Duluth, August 2011 - May 2013

RESEARCH EXPERIENCE

- PhD Dissertation Research, Indiana University Purdue University Indianapolis

  Joint Modeling of Longitudinal and Survival Data Using Estimation Equation Approach.

  - **Invited talk**: Lifetime Data Science (LIDA) 2017, Storrs, CT.
  - **Contributed papers presentation**: Eastern North American Region (ENAR) 2017, Washington DC.
  - **Invited seminar talk**: Cincinnati Children's Hospital Medical Center Division of Biostatistics, 2017, Cincinnati OH.
  - **Poster**: International Chinese Statistical Association (ICSA) Midwest conference 2016, Chicago, IL.

- Graduate Thesis Project, University of Minnesota Duluth

  Penalized Maximum Likelihood Estimation of Two-parameter Exponential distributions.

SELECT PUBLICATIONS

- Hendrie, H.C., **Zheng, M.**, Lane, K., Ambuehl, R., Unverzagt, F. W., Callahan, C. M., Gao, S. et al. (2018). Changes of glucose levels precede dementia in African American patients with diabetes but not in Caucasians. Alzheimer's & Dementia.
- Li, S., **Zheng, M.**, Gao, S. (2017). Joint modeling of longitudinal cholesterol measurements and time to onset of dementia in an elderly African American cohort. Biostatistics & Epidemiology.

- Hendrie, H.C., **Zheng, M.**, Lane, K., Ambuehl, R., Unverzagt, F. W., Callahan, C. M., Gao, S. et al. (2016). Glucose level decline precedes dementia in elderly African Americans with diabetes. Alzheimer's & Dementia.
- Li, S., **Zheng, M.**, Gao, S. et al. Low HDL cholesterol but not high LDL cholesterol is associated with the incidence of coronary artery disease in older African Americans. *(submitted to Age and Ageing)*