2018

Estimation Procedures for Complex Survival Models and Their Applications in Epidemiology Studies

Jie Zhou
University of South Carolina

Estimation Procedures for Complex Survival Models and Their Applications in Epidemiology Studies

by

Jie Zhou

Bachelor of Science
East China Normal University, 2012

Master of Science in Public Health
University of South Carolina, 2014

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Biostatistics

The Norman J. Arnold School of Public Health

University of South Carolina

2018

Accepted by:

Jiajia Zhang, Major Professor

Alexander C. McLain, Committee Member

James W. Hardin, Committee Member

Xuemei Sui, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

Acknowledgments

Many thanks to my dissertation advisor, Dr. Jiajia Zhang, for her continual support and advice throughout all my Ph.D study. Without her, this project could not have been completed. She has been a tremendous mentor for me. I would like to thank her for encouraging my research and for allowing me to grow as a research scientist. Her advice on both research as well as on my career have been priceless.

I would also like to thank Dr. Alex McLain, Dr. James Hardin and Dr. Xuemei Sui for serving as my committee members even at hardship. I would also like to thank you for letting my defense be an enjoyable moment, and for your brilliant comments and suggestions, appreciations for you.

My thanks also go to my parients, without their support and love, I will not have courage to complete my study in United States. I also thank my husband, Weinan Xu, for supporting me all the time.

Abstract

In this dissertation, we aim to address three important questions in practice, which can be solved through complex survival models. The first project focuses on studying the longitudinal fitness effect on cardiovascular disease (CVD) mortality. In the second project, we study the disease-death relation between CVD and all-cause mortality and evaluate important covariate effects on the disease or death transitions. In the third project, we compare antiretroviral treatment (ART) for HIV patients and consider both treatment effect and side effect of the drugs. The first two projects are motivated by the Aerobics Center Longitudinal Study (ACLS) datasets and the third project is based on the Health Sciences South Carolina (HSSC) HIV datasets.

The ACLS is a prospective study and involves patients in the Cooper Clinic in Dallas, TX. Participants had repeated measures of cardiorespiratory fitness (fitness), which is an objective measure of physical activity, during the study. Fatal outcomes, such as the CVD or all-cause mortality information, are available by the end of study. In the first project, we develop a novel joint model that allows the estimation of a time-varying exposure on a survival outcome with a varying coefficient model. Specifically, the flexible generalized odds rate models are applied to CVD mortality with an age-dependent coefficient to account for nonlinear age varying effect of fitness.

For the second project, we consider the interval censored disease incidence time, which is caused by the intermittent observations, and apply the Markov illness-death regression models to study the transition intensities among three states: disease-free, CVD and death, and estimate the covariate effects, such as age, fitness, smoking etc.,

on these transitions. We adopt the Expectation-Maximization (EM) algorithm to estimate the proposed models in the first two projects, and the covariance matrix of the estimated parameters is approximated numerically based on the profile likelihood.

HSSC is a biomedical research collaborative consisting of four of the state's largest health systems. We are interested in comparing the antiretroviral treatment (ART) for HIV patients in the HSSC. The HIV datasets in HSSC include both the time to treatment or virologic failures and side effects after drug administration. In the last project, we propose to model time to treatment or virologic failure and time to severe side effects of ART under the competing risks model framework. A restricted optimal treatment regime is defined based on cumulative incidence functions, where we minimize the risk of treatment or virologic failures while controlling the risk of serious drug-induced side effects. The estimation approach is derived using a penalized value search method.

The proposed models and their estimation algorithms are validated through extensive simulation studies and applied to either the ACLS datasets or the HSSC HIV datasets to achieve the purposes of the study.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and Motivations

Time to event data are commonly occurred in practice, such as medical and epidemiology studies. For example, in the Aerobics Center Longitudinal Study (ACLS) database, we are interested in the time to CVD or all-cause mortality, and in the Health Sciences South Carolina (HSSC) HIV dataset, after drug administration, the treatment or virologic failure and incidence of serious side-effects are two competing risks, and time to both events are of our interest.

In this part, we introduce the ACLS and HSSC datasets in Chapter 1.1 and Chapter 1.2, respectively. Specifically, we focus on the data structures and the aims of our projects for each database. Some preliminary data analysis results and motivations are also given in Section 1.1 for Project 1 and Project 2 and Section 1.2 for Project 3. Finally, the outline of the dissertation is illustrated in Chapter 1.3.

1.1 ACLS Database

The proposed research is based on the Aerobics Center Longitudinal Study (ACLS) database, which involves patients in the Cooper Clinic in Dallas, TX. The patients went to the clinic for periodic preventive medical examinations and for counseling regarding health and lifestyle behaviors. At the time of their examination, ACLS was described to the patients and the written informed consent for enrollment for the follow-up study was obtained. Participants were mostly Caucasian (>95%) and well-educated.

A prospective study design is used to analyze the ACLS database. The main expo-

sure variable is the cardiorespiratory fitness, which is quantified as the total duration of a symptom limited maximal treadmill exercise test (Balke and Ware, 1959), is used as the measure of physical activity (Physical Activity of Sports Medicine, 2013). All tests were supervised by a physician and conducted in accord with standardized exercise testing procedures. It is a more reliable measure of recent activity levels than self-reported values. Other potential confounders we consider in the model, including age, gender, BMI, smoking status and family history of CVD, were recorded during the initial visit.

During the study, participants had the cardiovascular disease (CVD) either reported or diagnosed in each clinical visit. Fatal outcomes, including the CVD and all-cause mortality, were from mortality surveillance, principally through the National Death Index (NDI), which covers all deaths in the United States after 2004.

We have two different aims motivated by the ACLS database, and they are studied separately in the first two projects. In Project 1, we are interested in studying the longitudinal effect of cardiorespiratory fitness on time to CVD mortality, and adjust other baseline covariates. In Project 2, the transitions among three states, including disease-free, CVD and all-cause mortality, are studied through the illness death modeling structure, and the covariates' effects on the transitions are estimated.

Preliminary for Project 1

For the first project, we aim to evaluate the longitudinal fitness effect on CVD mortality under the joint modeling framework. Patients in the ACLS had periodic preventive medical examinations, including longitudinal measurements of cardiorespiratory fitness ("fitness") where subjects completed a standard exercise test (Balke and Ware, 1959), an objective measure for physical activity (Physical Activity of Sports Medicine, 2013).

We include 3,980 participants, who were enrolled in the ACLS during $1970 \sim 1980$ and had at least three follow-up visits by the end of year 2004. Among them about 145 (3.64%) participants died because of CVD. 437 patients are females and 3,543 of them are males. The number of follow-up visits for each participant ranges from 3 to 30 with median equals to 5.

Based on the ACLS, Blair et al. (1996) discovered an inverse association between the baseline fitness and CVD mortality. Similarly, we first look at the baseline data for the preliminary analysis. A Cox proportional hazards (PH) model (David et al., 1972) is fitted for CVD mortality, where we include the baseline fitness and adjust BMI, family history of CVD, smoking status, gender and age as potential confounders.

Table 1.1 Estimated PH Models for ACLS Baseline Data

	Without Interaction			With Interaction		
Variable	Estimate	StDev	P value	Estimate	StDev	P value
BMI	0.103	0.030	0.001	0.110	0.030	0.000
FamilyCVD	0.124	0.168	0.462	0.133	0.168	0.427
Smoke	0.218	0.233	0.349	0.230	0.233	0.323
Female	-0.342	0.371	0.356	-0.348	0.371	0.347
AGE	0.117	0.011	0.000	0.181	0.035	0.000
Fitness	-0.013	0.023	0.572	0.192	0.107	0.072
AGE×Fitness	-	-	-	-0.004	0.002	0.050

The estimated results are summarized in the left part of Table 1.1. The adjusted baseline fitness is found to have a protective effect on CVD mortality (coefficient= -0.013), but the effect is not significant (p value= 0.572). However, if we consider age as an effect modifier of fitness and include the "age×fitness" in the model, as summarized in the right part of Table 1.1, we find the interaction term is marginally significant (p value= 0.050). This indicates that the effect of fitness on CVD mortality changes over age.

Previous analysis based on the baseline fitness does not account for the whole

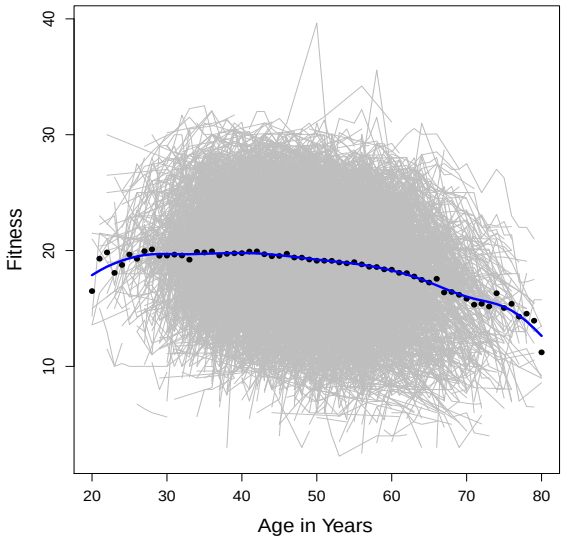

Figure 1.1 Profile Plots of Longitudinal Fitness.

pattern of fitness during a person's life span. To study the longitudinal effect of fitness, we need to use the repeated measures of fitness for each subject. Moreover, it is well known that there are changes in the overall level of fitness with age. For example, Figure 1.1 displays the longitudinal fitness profiles for all participants in the ACLS over age. It can be seen that the mean fitness is around 20 and gradually decreases with age.

Further, while the standard exercise test is an objective measure of physical activity which is superior to self-report, the values appear to be subjected to measurement error. This measurement error could be due to true measurement error in the equipment, or small biological fluctuations in the subjects fitness level on the day of the measurement (e.g., a bad night of sleep). Also considering the effect of fitness on CVD mortality is modified by age, we seek to model the association between a time-varying covariate that is subject to measurement error and a survival outcome with a varying-coefficient joint model.

Preliminary for Project 2

In the second project, we aim to study the CVD incidence and all-cause mortality in the ACLS data under the illness-death modeling framework. Specifically, 5236 CVD-free participants, who were enrolled in the ACLS during 1970 $\sim$ 1980, are included in the analysis and being followed until the end of year 2004. During the study, each participant had a sequence of follow-up visits, say $0 = v_0 < v_1 < \cdots < v_K < \infty$, and had the cardiovascular disease (CVD) either reported or diagnosed during each visit. We also have both the death information and its major cause (CVD or other cause) for each participant from mortality surveillance, principally through the National Death Index (NDI).

As a result, each subject has the risk of developing CVD, or dies directly without CVD. We have intermittent CVD diagnosis information for each subject, and the true incidence time of CVD is either between two consecutive visits, say (v_{k-1}, v_k), or right censored. The exact death information is obtainable through NDI, therefore, we assume the death time is only subject to right censoring, and the only case that it is right censored is because the patient is still alive at the end of study.

Figure 1.2 shows the distribution of the participants with regard to their disease and death status at the end of the study. Among all the participants, 353 (6.74%) have CVD diagnosed during the study and 274 out of them died eventually. There are 479 (9.15%) subjects died without CVD and 4404 (84.11%) were still alive and were CVD-free at their last follow-up visits.

We are interested in studying the following three problems based on the ACLS data: (1) estimate the transition intensities between the states including disease-free, CVD and death; (2) compare the survival experience for subjects with and without CVD; and (3) explore the covariate effects in each transition process. To achieve

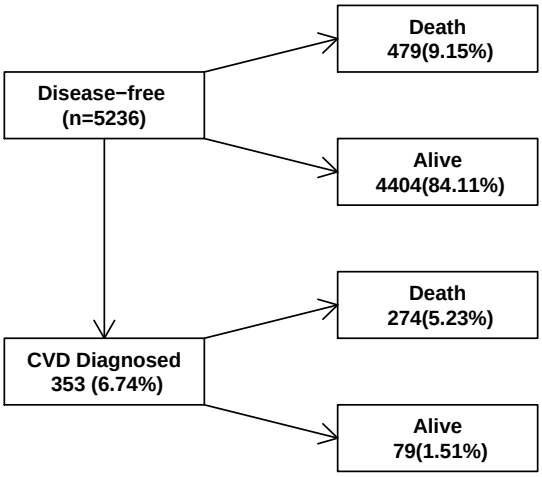

Figure 1.2 Illness-Death Process in ACLS Data.

these goals, we consider the Markov multi-state regression models and estimate the transition intensities and covariate coefficients based on the intermittent observations of CVD incidence.

1.2 HSSC HIV Dataset

Health Sciences South Carolina (HSSC) (`https://www.healthsciencessc.org/`) is a biomedical research collaborative consisting of four of the state's largest health systems namely University Medical Center, Spartanburg Regional Healthcare System, McLeod Health, AnMed Health, and Self Regional Healthcare. The HSSC database includes several datasets that can be linked based on the subject and visit ID numbers. The datasets we used include the patient's demographic information, visit information, diagnosis, medication order history and laboratory test results.

We are interested in comparing the antiretroviral treatment (ART) among the population with HIV diagnosis in HSSC. There are three most commonly used ART class-

es: nucleoside reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs) and protease inhibitors (PIs) (Günthard et al., 2016). The drug class information can be searched on the website: `http://bioportal.bioontology.org/ontologies/RXNORM`, and linked with the medication order history dataset based on the RxNORM code.

In general, drugs in the same class share common properties, whereas drugs in different class have different treatment effects. For example, NNRTIs is associated with faster virologic suppression and PIs recover more CD4 cells (Organization, 2016; Günthard et al., 2016). Modern ART consists a combination of at least three agents from two classes (Günthard et al., 2016). Common combinations such as "NNRTIs+NRTIs" and "PIs+NRTIs" have also been compared in literature regarding to their treatment effects measured via the level of virologic suppression or CD4 recovery. (Staszewski et al., 1999; Haubrich et al., 2009; Smith et al., 2009; Borges et al., 2016).

The evaluation for ART requires considerations of both treatment effects and side effects among different populations. The information about side effects of the drugs is based on patients' diagnosis records during their visits, where the ICD-9 or ICD-10 code are used to find the symptoms that related to the drug-induced side effects. In Project 3, we aim to obtain the optimal treatment regime for different populations that can minimize the risk of treatment or virologic failure while controlling the risk of long-term side effects under a tolerable limit based on the HSSC HIV dataset.

Preliminary for Project 3

In the third project, we aim to find the optimal ART treatment for HIV patients based on the HSSC HIV dataSET. Jiang et al. (2017) considered the optimal regime of "NNRTIs+NRTIs" and "PIs+NRTIs" to maximize the longest initial treatment duration based on a data set from HIV/AIDS clinical observational study. Similarly, we

compare “NNRTIs+NRTIs" and “PIs+NRTIs" with respect to both their treatment or virologic failures and the serious drug-induced side effects after the drug administration.

In HSSC data set, there are 426 patients who took drug combinations “NNRTIs+NRTIs" or “PIs+NRTIs" and had complete laboratory measures. We define “risk 1" as treatment or virologic failure, which is monitored by either CD4 counts ($\leq$ 500 cells/mm^3) or HIV viral load ($\geq$200 copies/mL), and “risk 2" as the drug-induced long-term side effects. Days to either risk, whichever came first after drug administration, were recorded.

We compared the cumulative incidence functions of the two HIV treatments among patients below and above 50-year-old for risk 1 in Figure 1.3 and for risk 2 in Figure 1.4, separately. Based on these curves, we found that “NNRTIs+NRTIs" has generally lower risk of treatment or virologic failure but higher risk of having serious side effects than “PIs+NRTIs" among younger patients. In contrast, among senior patients, “NNRTIs+NRTIs" has lower risk of treatment or virologic failure before 1000 days after drug administration, but similar performance after 1000 days compared with “PIs+NRTIs". Moreover, the risks of side effects with these two types of drugs reversed compared with younger patients.

Therefore, the criteria to assign “NNRTIs+NRTIs" or “PIs+NRTIs" to each individual are not consistent. In Project 3, we discuss an optimal treatment regime, which can balance between the treatment efficacy and side effects, under the competing risks framework. Specifically, we define a restricted optimal treatment regime that minimizes the t-year cumulative incidence function of the main risk while controlling the t-year cumulative incidence of the other risk under a predetermined level.

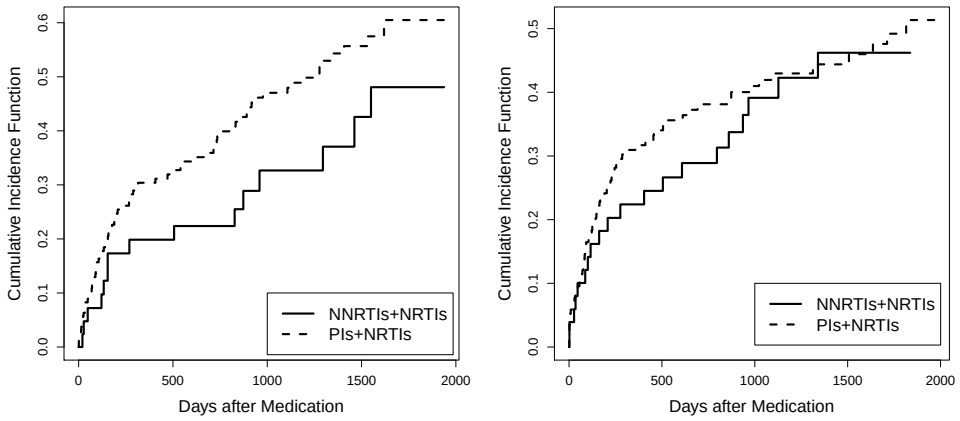

Figure 1.3 Risk 1 Cumulative Incidence Functions (left: age < 50, right: age ≥ 50)

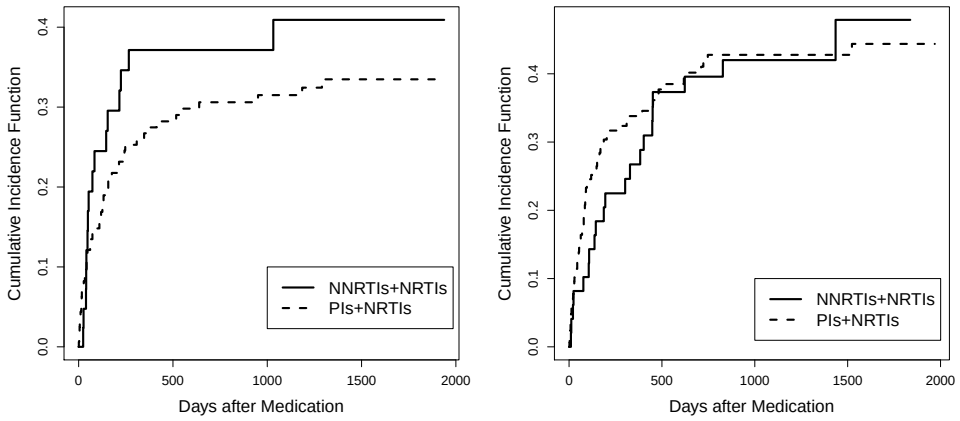

Figure 1.4 Risk 2 Cumulative Incidence Functions (left: age < 50, right: age ≥ 50)

1.3 Outline of Dissertation

In the rest of the dissertation, we present the proposed three projects separately in Chapters 2, 3 and 4. In each part, we discuss the proposed models and present the details of the estimation methods for each project.

In Chapter 2, we develop a novel joint model that allows the estimation of a time-varying exposure on a survival outcome with a varying coefficient model. The flexible

generalized odds rate models are applied to CVD mortality with an age-dependent coefficient to account for nonlinear age varying effect of fitness. Our model uses a mixed effects model for the longitudinal process, and cubic B-splines to estimate the varying coefficient function. The proposed joint model is estimated based on the Expectation-Maximization (EM) algorithm, where the numerical integrals with respect to the random effects are approximated by a modified pseudo adaptive Gaussian-Hermite quadrature in the E-step. The covariance matrix is approximated numerically based on profile likelihood. All of the estimation details are presented in Section 2.2. The performance of the proposed algorithm is validated through the extensive simulation studies in Section 2.3. Finally, it is applied to a cohort in the ACLS in Section 2.4.

Chapter 3 investigates the transitions to disease or death under the semi-competing risks model. We propose to study the process of developing cardiovascular disease (CVD) and all-cause mortality in the ACLS data, and focus on the covariate effects, such as age, fitness, smoking etc., in the transitions to the CVD or death states. Due to the intermittent observations of the CVD incidence time, we have both interval censored disease time and right censored death time. The details of the estimation procedures are discussed in Section 3.2, Where we propose to use the Markov illness-death regression models and apply the expectation-maximization (EM) algorithm to derive a self-consistent estimator for the model. The variance of the estimates are approximated based on the profile likelihood function. The proposed method is evaluated through extensive simulation studies in Section 3.3, and illustrated by the application to the ACLS data in Section 3.4.

It is well accepted that individualized treatment regimes may have potential benefit to improve the clinical outcome of interest. However, the positive treatment effects often accompany with certain side effects. That is, when choosing the optimal treatment regime for a patient we need to consider both efficacy and safety issues. In Chapter 4,

we propose to model time to a primary event of interest and time to severe side effects of treatment by a competing risks model and define a restricted optimal treatment regime based on cumulative incidence functions. The estimation approach is presented in Section 4.2 and investigated through numerical studies. Specifically, a penalized value search method is derived and evaluated through extensive simulations in Section 4.3. The proposed method is applied to an HSSC HIV dataset in Section 4.4, where we minimize the risk of treatment or virologic failures while controlling the risk of serious drug-induced side effects.

Discussions and conclusions are made for each project at the end of each part. Finally, some summaries and future works are discussed in Chapter 5.

Chapter 2

A Generalized Odds Rate Joint Model for Varying Coefficients with Time-varying Exposures

Promoting a physically active lifestyle is a major national public health priority. Physical inactivity, mainly due to a sedentary lifestyle, has been shown to have a positive association with cardiovascular disease (CVD) mortality (Blair et al., 1996; Kohl 3rd, 2001; Mora et al., 2007; Nocon et al., 2008). The Aerobic Center Longitudinal Study (ACLS) enrolled 3,980 participants from the Cooper Clinic in Dallas, TX from $1970 \sim 1980$ with follow-up till 2004. Patients in the ACLS had periodic preventive medical examinations, including longitudinal measurements of cardiorespiratory fitness ("fitness") where subjects completed a standard exercise test (Balke and Ware, 1959), an objective measure for physical activity (Physical Activity of Sports Medicine, 2013).

It is well known that physical fitness has an impact on cardiovascular disease (CVD) mortality. For example, an inverse association between the baseline fitness and CVD mortality was discovered based on the ACLS by Blair et al. (1996). It is not known, however, how the effect of fitness on CVD mortality varyies with age. Based on the preliminary data analysis in Section 1.1, the effect of fitness on CVD mortality is modified by age. Moreover, we would like to utilize the repeated measures of fitness in the ACLS, other than the baseline data only, to represent participants' fitness trajectory and study the longitudinal effect of fitness on CVD mortality.

Therefore, we seek to model the association between a time-varying covariate on a survival outcome with a varying-coefficient model. In practice it is challenging to capture the association between a time-varying covariate and a survival outcome with a varying-coefficient model. Previous studies focused on either estimation of varying coefficients for time-independent variables (Cai and Sun, 2003; Tian et al., 2005), or fixed coefficients for time-dependent variables (Fisher and Lin, 1999; Zeng and Lin, 2006). To the best of our knowledge, there's no literature on survival models that consider both a time-varying covariate and its varying-effect over another variable. What complicates our situation more is that the exposure of interest is a endogeneous covariate that is subject to measurement error, where the previous methods do not apply.

The most popular tools in modeling the association between a survival outcome and an endogeneous covariate with measurement error are joint models. Specifically, a mixed effects model with normal random effects is assumed for the longitudinal observations and standard survival models are used for the survival outcome. There have been plenty of work on joint models which combines the linear mixed model with Cox proportional hazards (PH) model (Wulfsohn and Tsiatis, 1997; Bycott and Taylor, 1998; Zeng et al., 2005; Zeng and Cai, 2005). Further, the proportional odds (PO) joint model has also been studied in the literature when the PH assumption is violated (Andrinopoulou et al., 2014). Various of extensions of joint models have been made to account for complex data structures in practice, with considerations of multiple longitudinal outcomes (Song et al., 2002; Brown et al., 2005; Rizopoulos and Ghosh, 2011; Moreno-Betancur et al., 2017), competing risks (Elashoff et al., 2008; Huang et al., 2011), and cure rate models (Yu et al., 2004; Brown and Ibrahim, 2003). More overviews and extensions can be found in Tsiatis and Davidian (2004) and Rizopoulos (2012).

The existing joint models do not allow varying-coefficients, so they cannot be used to estimate the age-related association between fitness and CVD mortality. Therefore, we develop a novel joint model framework considering the following three features: (1) longitudinal process of fitness, (2) survival process of CVD mortality, and (3) the age-related fitness effects. For the longitudinal fitness process, we assume a flexible pre-specified time function with random coefficients to accommodate for subject-specific longitudinal trajectories over time. For the survival process, we propose to incorporate the generalized odds rate (GOR) model (Dabrowska and Doksum, 1988; Scharfstein et al., 1998; Zhou et al., 2017), including the PH model and the PO model (Bennett, 1983) as special cases. To investigate the age-related fitness effect on CVD mortality, we include a novel age-dependent varying coefficient for longitudinal fitness in the survival model. A clear pattern of the effect of fitness on CVD mortality with age can be described by the estimated nonlinear varying coefficient. In addition, based on the estimated point-wise confidence intervals for the varying coefficient, an age period for fitness being a significant protective effect for CVD mortality can be detected as well. The proposed model can improve understanding of how age-related changes in fitness effect CVD mortality, which can provide direct guidance in behavior consultation.

The rest of the part is organized as follows. The notations and model definitions are first introduced in Section 2.1. Specifically, the details of the estimation procedures are presented in Section 2.2, which includes the derivation of the complete likelihood function, calculation of the conditional expectations and maximization steps. The complete EM algorithm and the corresponding variance estimation are presented at the end of the section. The extensive simulation studies are performed in Section 2.3. To study the nonlinear age-dependent effect of fitness on the CVD mortality, we apply the proposed model and method to the ACLS data in Section 2.4. The final discussions and conclusions are summarized in Section 2.5.

2.1 GOR Joint Models

Let T_i denote the failure time for subject i, $i = 1, \cdots, n$. The distribution of T_i depends on a vector of baseline covariates Z_i, age $A_i(\cdot)$ and a time-varying predictor $W_i(\cdot)$. The filtration of $W_i(\cdot)$ is denoted by $\mathcal{W}_i(t) = \{W_i(s) : s \leq t\}$, which include the history of $W_i(\cdot)$ up to time t. Let $\Lambda_i(\cdot)$ denote the cumulative hazard function of T_i. Under the generalized odds rate (GOR) model, we have

$$\Lambda_i(\cdot) = \Lambda(t|Z_i, A_i, \mathcal{W}_i(t)) = G_r\left\{\int_0^t \lambda_0(s)\exp[Z_i\boldsymbol{\beta} + \psi(A_i(s))W_i(s)]ds\right\},$$

where $G_r\{\cdot\}$ is a pre-specified increasing transformation function, which is indexed by a non-negative argument r. $\lambda_0(t)$ is the baseline hazard function and will be estimated non-parametrically. $\boldsymbol{\beta}$ is the vector of coefficients for Z_i and $\psi(A_i(s))$ is the age-dependent varying coefficient for $W_i(s)$, where $A_i(s)$ is the age at time s and $\psi(s)$ is a smoothing function. For example, a possible transformation is $G_r(x) = \frac{1}{r}\log(1 + rx)$ when $r > 0$ and $G_r(x) = x$ when $r = 0$, which reduces to the PH model when $r = 0$ and the PO model when $r = 1$. We approximate the smoothing function using cubic B-splines, with $\psi(s) = \sum_{l=1}^{L}\gamma_l B_l(s)$ where $B_l(\cdot)$ $l = 1, \cdots, L$, are the B-spline basis functions.

The whole history of the longitudinal marker $\mathcal{W}_i(t)$ is not obtainable in reality. Instead, we can only observe $Y_i(\cdot)$, which is a contaminated version of $W_i(\cdot)$, at a sequence of intermittent follow-up visit times denoted by $0 = t_{i,0} < t_{i,1} < \cdots < t_{i,m_i}$. We assume the following random effects model for $\boldsymbol{Y}_i = (Y_{i,1}, \cdots, Y_{i,m_i})$, where $Y_{i,j} = Y_i(t_{i,j})$ denotes the observation for subject i at $t_{i,j}$, $j = 1, \cdots, m_i$,

$$Y_i(t|\boldsymbol{b}_i) = W_i(t|\boldsymbol{b}_i) + \epsilon(t) = g'(t)\boldsymbol{b}_i + \epsilon(t),$$

where $g(t)$ is a d-dimensional vector of known functions of t, for example, $g(t) = (1, t)'$ corresponding to a linear function of t with $d = 2$ and $\boldsymbol{b}_i = (b_{i1}, \cdots, b_{id})$ is a d-dimensional vector of random effects and is assumed to jointly follow multivariate

normal distribution $MVN(\boldsymbol{\mu}, \boldsymbol{D})$, where $\boldsymbol{\mu}$ and $\boldsymbol{D}$ are the mean vector and $d \times d$ variance-covariance matrix of $\boldsymbol{b}_i$. The error terms $\boldsymbol{\epsilon} = (\epsilon_{i,1}, \cdots, \epsilon_{i,m_i})$, where $\epsilon_{i,j} = \epsilon(t_{i,j}), j = 1, \cdots, m_i$, are assumed to follow $N(0, \sigma^2)$.

2.2 Estimation Procedures

Complete Likelihood Function

We observe $V_i = \min(T_i, C_i)$ with a censoring indicator $\delta_i = I(T_i \leq C_i)$ for $i = 1, \cdots, n$, where C_i is the right censoring time. The observed data for subject i can be denoted as $O_i = (V_i, \delta_i, A_i, Z_i, \boldsymbol{t}_i, \boldsymbol{Y}_i)$ and the parameters to be estimated include $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda_0, \boldsymbol{\mu}, \boldsymbol{D}, \sigma^2)$, where $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_L)$ is the vector of coefficients in the B-splines. Note, the notation $\Lambda(t|Z_i, A_i, \mathcal{W}_i(t))$ and $\Lambda(t|O_i, \boldsymbol{b}_i)$ are equivalent. Let $S(t|O_i, \boldsymbol{b}_i)$ denote the survival function corresponding to $\Lambda(t|O_i, \boldsymbol{b}_i)$. Under the GOR model defined in Section 2.1, $S(t|O_i, \boldsymbol{b}_i)$ can be written as the marginal survivor function of a gamma frailty model. That is, $S(t|O_i, \boldsymbol{b}_i) = \int S(t|O_i, \phi_i, \boldsymbol{b}_i) f(\phi_i) d\phi_i$ where

$$S(t|O_i, \phi_i, \boldsymbol{b}_i) = \exp\left\{-\phi_i \int_0^t \lambda_0(s) e^{Z_i\boldsymbol{\beta}} \exp[\psi(A_i(s)) W_i(s|\boldsymbol{b}_i)] ds\right\},$$

and $f(\cdot)$ is the gamma density with mean of 1 and variance r.

The complete likelihood function of $\boldsymbol{\theta}$ given the observed data $\boldsymbol{O} = (O_1, \cdots, O_n)$, the frailty terms $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_n)$ and the random effects $\boldsymbol{b} = (\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n)$ can be written as:

$$\begin{aligned}
\mathcal{L}_c(\boldsymbol{\theta}|\boldsymbol{O}, \boldsymbol{\phi}, \boldsymbol{b}) &= \prod_{i=1}^{n} p(V_i, \delta_i|\phi_i, \boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda_0) \times p(\boldsymbol{Y}_i|\boldsymbol{b}_i; \sigma^2) \times p(\boldsymbol{b}_i|\boldsymbol{\mu}, \boldsymbol{D}) \times f(\phi_i) \\
&= \prod_{i=1}^{n} \left\{\phi_i \lambda_0(V_i) e^{Z_i\boldsymbol{\beta}} \eta(V_i|A_i, \boldsymbol{b}_i; \boldsymbol{\gamma})\right\}^{\delta_i} \times \exp\left\{-\phi_i \int_0^{V_i} \lambda_0(s) e^{Z_i\boldsymbol{\beta}} \eta(s|A_i, \boldsymbol{b}_i; \boldsymbol{\gamma}) ds\right\} \\
&\quad \times (2\pi\sigma^2)^{-\frac{m_i}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{Y}_i - \boldsymbol{G}_i\boldsymbol{b}_i)'(\boldsymbol{Y}_i - \boldsymbol{G}_i\boldsymbol{b}_i)\right\} \\
&\quad \times (2\pi)^{-\frac{d}{2}} |\boldsymbol{D}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{b}_i - \boldsymbol{\mu})'\boldsymbol{D}^{-1}(\boldsymbol{b}_i - \boldsymbol{\mu})\right\} \times f(\phi_i), \qquad (2.1)
\end{aligned}$$

where $\boldsymbol{G}_i = (g(t_{i,1}), \cdots, g(t_{i,m_i}))'$ and $\eta(s|A_i, \boldsymbol{b}_i; \boldsymbol{\gamma}) = \exp\{\psi(A_i(s))W_i(s|\boldsymbol{b}_i)\}$.

The observed likelihood function $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{O})$ can be derived by integrating the frailty terms $\boldsymbol{\phi}$ and the random effects $\boldsymbol{b}$ out of (2.1). Direct maximization of the observed likelihood $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{O})$ is difficult due to the numerical integrals regarding to the random effects. Therefore, we apply the EM algorithm to estimate the proposed joint model, and assume the baseline function $\lambda_0(\cdot)$ to be nonparametric. The complex form of the likelihood function and the infinite dimension of the parameter space make this a challenging computation task.

Conditional Expectations

After dropping the terms that do not contain $\boldsymbol{\theta}$, the complete log-likelihood function can be written as the summation of three distinct parts, i.e.

$$l^c(\boldsymbol{\theta}|\boldsymbol{\phi}, \boldsymbol{b}) = l_1^c(\lambda_0, \boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\phi}, \boldsymbol{b}) + l_2^c(\sigma^2|\boldsymbol{b}) + l_3^c(\boldsymbol{\mu}, \boldsymbol{D}|\boldsymbol{b}),$$

where

$$\begin{aligned}
l_1^c(\lambda_0, \boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\phi}, \boldsymbol{b}) =& \sum_{i=1}^{n} \delta_i[\log(\lambda_0(V_i)) + Z_i\boldsymbol{\beta} + \log(\eta(V_i|A_i, \boldsymbol{b}_i; \boldsymbol{\gamma}))] \\
& - \phi_i \int_0^{V_i} \lambda_0(s) e^{Z_i\boldsymbol{\beta}} \eta(s|A_i, \boldsymbol{b}_i; \boldsymbol{\gamma})]ds, \\
l_2^c(\sigma^2|\boldsymbol{b}) =& \sum_{i=1}^{n} -\frac{m_i}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{Y}_i - \boldsymbol{G}_i\boldsymbol{b}_i)'(\boldsymbol{Y}_i - \boldsymbol{G}_i\boldsymbol{b}_i), \quad \text{and} \\
l_3^c(\boldsymbol{\mu}, \boldsymbol{D}|\boldsymbol{b}) =& \sum_{i=1}^{n} -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{D}|) - \frac{1}{2}(\boldsymbol{b}_i - \boldsymbol{\mu})'\boldsymbol{D}^{-1}(\boldsymbol{b}_i - \boldsymbol{\mu}).
\end{aligned}$$

Let $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ denote the conditional expectation of the complete log-likelihood function $l^c(\boldsymbol{\theta}|\boldsymbol{\phi}, \boldsymbol{b})$ given observed data $\boldsymbol{O} = (O_1, \cdots, O_n)$ and current estimates $\boldsymbol{\theta}^{(k)}$. Similar to previous arguments, $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ can be written as the summation of three

distinct parts,

$$\begin{aligned} Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}) =& E_{\boldsymbol{b}}\left\{E_{\boldsymbol{\phi}}\left[l^c(\boldsymbol{\theta}|\boldsymbol{\phi},\boldsymbol{b})|\boldsymbol{O},\boldsymbol{b}\right]|\boldsymbol{O},\boldsymbol{\theta}^{(k)}\right\} \\ =& E_{\boldsymbol{b}}\left\{E_{\boldsymbol{\phi}}\left[l_1^c(\lambda_0,\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{\phi},\boldsymbol{b})|\boldsymbol{b},\boldsymbol{O},\boldsymbol{\theta}^{(k)}\right]|\boldsymbol{O},\boldsymbol{\theta}^{(k)}\right\} \\ &+ E_{\boldsymbol{b}}\left\{l_2^c(\sigma^2|\boldsymbol{b})|\boldsymbol{O},\boldsymbol{\theta}^{(k)}\right\} + E_{\boldsymbol{b}}\left\{l_3^c(\boldsymbol{\mu},\boldsymbol{D}|\boldsymbol{b})|\boldsymbol{O},\boldsymbol{\theta}^{(k)}\right\} \\ =& Q_1(\lambda_0,\boldsymbol{\beta},\boldsymbol{\gamma};\boldsymbol{\theta}^{(k)}) + Q_2(\sigma^2;\boldsymbol{\theta}^{(k)}) + Q_3(\boldsymbol{\mu},\boldsymbol{D};\boldsymbol{\theta}^{(k)}). \end{aligned} \tag{2.2}$$

To evaluate the conditional expectation $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$, we need to calculate both $E(\phi_i|\boldsymbol{b}_i,O_i,\boldsymbol{\theta}^{(k)})$ and the conditional expectations of functions of $\boldsymbol{b}_i$ given O_i and current estimate $\boldsymbol{\theta}^{(k)}$. The conditional distribution of ϕ_i given $\boldsymbol{b}_i$, O_i and $\boldsymbol{\theta}^{(k)}$ is

$$p(\phi_i|\boldsymbol{b}_i,O_i) \propto \phi_i^{\delta_i} \times \exp\left\{-\phi_i\int_0^{V_i}\lambda_0(s)e^{Z_i\boldsymbol{\beta}}\eta(s|A_i,\boldsymbol{b}_i;\boldsymbol{\gamma})ds\right\} \times f(\phi_i).$$

Plugging in the density for Gamma$(r,1/r)$ and doing some algebra, it can be shown that the resulted conditional distribution is a gamma distribution with shape parameter $\delta_i + 1/r$ and scale parameter $[1/r + \int_0^{V_i}\lambda_0(s)e^{Z_i\boldsymbol{\beta}}\eta(s|A_i,\boldsymbol{b}_i;\boldsymbol{\gamma})ds]^{-1}$.

Expectations with respect to the conditional distribution of $\boldsymbol{b}_i$ given O_i and $\boldsymbol{\theta}^{(k)}$ can be approximated using a modified version of the adaptive Gaussian-Hermite (GH) quadrature. To achieve that, we first need to find the kernel of the conditional distribution of $\boldsymbol{b}_i$. From the joint distribution in Equation (2.1), we have

$$\begin{aligned} f(\boldsymbol{b}_i|O_i,\boldsymbol{\theta}^{(k)}) \propto& \left[\frac{1}{r} + \int_0^{V_i}\lambda_0(s)e^{Z_i\boldsymbol{\beta}}\eta(s|A_i,\boldsymbol{b}_i;\boldsymbol{\gamma})ds\right]^{-(\delta_i+\frac{1}{r})} \times [\eta(V_i|A_i,\boldsymbol{b}_i;\boldsymbol{\gamma})]^{\delta_i} \\ &\times \exp\left\{-\frac{1}{2\sigma^2}[\boldsymbol{b}_i'\boldsymbol{G}_i'\boldsymbol{G}_i\boldsymbol{b}_i - 2\boldsymbol{Y}_i'\boldsymbol{G}_i\boldsymbol{b}_i]\right\}\exp\left\{-\frac{1}{2}(\boldsymbol{b}_i'\boldsymbol{D}^{-1}\boldsymbol{b}_i - 2\boldsymbol{\mu}'\boldsymbol{D}^{-1}\boldsymbol{b}_i)\right\} \\ \propto& \quad S_i(\boldsymbol{b}_i) \times \exp\left\{-\frac{1}{2}(\boldsymbol{b}_i-\tilde{\mu}_i)'\tilde{\Sigma}_i^{-1}(\boldsymbol{b}_i-\tilde{\mu}_i)\right\}, \end{aligned} \tag{2.3}$$

where

$$S_i(\boldsymbol{b}_i) = \left[\frac{1}{r} + \int_0^{V_i}\lambda_0(s)e^{Z_i\boldsymbol{\beta}}\eta(s|A_i,\boldsymbol{b}_i;\boldsymbol{\gamma})ds\right]^{-(\delta_i+\frac{1}{r})}$$

is the survival part that depends on $\boldsymbol{b}_i$; and the rest is proportional to a Normal kernel with mean $\tilde{\mu_i}' = \left[\delta_i \psi(A_i + V_i) g'(V_i) + \frac{1}{\sigma^2} \boldsymbol{Y}_i' \boldsymbol{G}_i + \boldsymbol{\mu}' \boldsymbol{D}^{-1}\right] \tilde{\Sigma}_i$ and covariance matrix $\tilde{\Sigma}_i = [\boldsymbol{D}^{-1} + \boldsymbol{G}_i' \boldsymbol{G}_i / \sigma^2]^{-1}$.

Let $\mathscr{K}_i$ denote the kernel of the conditional distribution in Equation (2.3). The conditional expectation of any function $h(\boldsymbol{b}_i)$ can be calculated through

$$
\begin{aligned}
E[h(\boldsymbol{b}_i)|\boldsymbol{\theta}^{(k)}, O_i] =& \int_{-\infty}^{\infty} h(\boldsymbol{b}_i) p(\boldsymbol{b}_i|\boldsymbol{\theta}^{(k)}, O_i) d\boldsymbol{b}_i \\
=& \frac{\int_{-\infty}^{\infty} h(\boldsymbol{b}_i) \mathscr{K}_i d\boldsymbol{b}_i}{\int_{-\infty}^{\infty} \mathscr{K}_i d\boldsymbol{b}_i} \\
=& \frac{\int_{-\infty}^{\infty} h(\boldsymbol{b}_i) S_i(\boldsymbol{b}_i) \times \exp\left\{-\frac{1}{2}(\boldsymbol{b}_i - \tilde{\mu_i})' \tilde{\Sigma_i}^{-1} (\boldsymbol{b}_i - \tilde{\mu_i})\right\} d\boldsymbol{b}_i}{\int_{-\infty}^{\infty} S_i(\boldsymbol{b}_i) \times \exp\left\{-\frac{1}{2}(\boldsymbol{b}_i - \tilde{\mu_i})' \tilde{\Sigma_i}^{-1} (\boldsymbol{b}_i - \tilde{\mu_i})\right\} d\boldsymbol{b}_i} \qquad (2.4) \\
=& \frac{\int_{-\infty}^{\infty} h(\tilde{\mu_i} + \sqrt{2} \tilde{\Sigma_i}^{1/2} \boldsymbol{r}_i) S_i(\tilde{\mu_i} + \sqrt{2} \tilde{\Sigma_i}^{1/2} \boldsymbol{r}_i) \times e^{-\boldsymbol{r}_i' \boldsymbol{r}_i} d\boldsymbol{r}_i}{\int_{-\infty}^{\infty} S_i(\tilde{\mu_i} + \sqrt{2} \tilde{\Sigma_i}^{1/2} \boldsymbol{r}_i) \times e^{-\boldsymbol{r}_i' \boldsymbol{r}_i} d\boldsymbol{r}_i}, \qquad (2.5)
\end{aligned}
$$

where the transformation $\boldsymbol{r}_i = \frac{1}{\sqrt{2}} \tilde{\Sigma_i}^{-1/2} (\boldsymbol{b}_i - \tilde{\mu_i})$ was made from (2.4) to (2.5). Both of the numerator and the denominator in (2.5) can be approximated by the Gaussian-Hermite (GH) quadrature $\int_{-\infty}^{\infty} f(x) e^{-x^2} dx \approx \sum_{j=1}^{K} \pi_j f(x_j)$, where x_j's and π_j are the abscissas and weights under K nodes given by the GH quadrature.

Maximization

In the maximization steps, we need to maximize the expectation of the log-likelihood functions $Q_1(\lambda_0(\cdot), \boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{\theta}^{(k)})$, $Q_2(\sigma^2; \boldsymbol{\theta}^{(k)})$ and $Q_3(\boldsymbol{\mu}, \boldsymbol{D}; \boldsymbol{\theta}^{(k)})$ described in Equation (2.2), and update the parameters $\boldsymbol{\theta}^{(k+1)}$.

Since $Q_1(\lambda_0(\cdot), \boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{\theta}^{(k)})$ involves the infinite dimensional parameter $\lambda_0(\cdot)$, we adopt a profile approach. Specifically, we first solve the partial derivative of $Q_1(\lambda_0, \boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{\theta}^{(k)})$ with respective to $\lambda_0(\cdot)$ and derive

$$
\widetilde{\lambda_0}(t; \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{\sum_{i=1}^{n} \delta_i I(V_i = t)}{\sum_{i=1}^{n} I(V_i \geq t) e^{Z_i \beta} E_{\boldsymbol{b}_i}\{E(\phi_i|\boldsymbol{b}_i, O_i, \boldsymbol{\theta}^{(k)}) \eta(t|A_i, \boldsymbol{b}_i; \boldsymbol{\gamma})|\boldsymbol{\theta}^{(k)}, O_i\}}
$$

as a function of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

Then the Newton-Raphson algorithm is applied to maximize the expectation of the profile log-likelihood function $Q_1(\widetilde{\lambda_0}(t;\boldsymbol{\beta},\boldsymbol{\gamma}),\boldsymbol{\beta},\boldsymbol{\gamma};\boldsymbol{\theta}^{(k)})$.

The gradient functions are

$$\frac{\partial}{\partial\beta_j}Q_1(\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{i=1}^{n}\delta_i\left\{\frac{\widetilde{\lambda_0}^{\beta_j}(V_i;\boldsymbol{\beta},\boldsymbol{\gamma})}{\widetilde{\lambda_0}(V_i;\boldsymbol{\beta},\boldsymbol{\gamma})}+Z_{ij}\right\} - \int_0^{V_i}\left[\widetilde{\lambda_0}^{\beta_j}(s;\boldsymbol{\beta},\boldsymbol{\gamma})+\widetilde{\lambda_0}(s;\boldsymbol{\beta},\boldsymbol{\gamma})Z_{ij}\right]\mathcal{H}_i^1(s)ds,$$

$$\begin{aligned}\frac{\partial}{\partial\gamma_l}Q_1(\boldsymbol{\beta},\boldsymbol{\gamma}) &= \sum_{i=1}^{n}\delta_i\left\{\frac{\widetilde{\lambda_0}^{\gamma_l}(V_i;\boldsymbol{\beta},\boldsymbol{\gamma})}{\widetilde{\lambda_0}(V_i;\boldsymbol{\beta},\boldsymbol{\gamma})}+B_l(A_i+V_i)g'(V_i)E(\boldsymbol{b}_i|\boldsymbol{\theta}^{(k)},O_i)\right\}\\ &\quad - \int_0^{V_i}\widetilde{\lambda_0}^{\gamma_l}(s;\boldsymbol{\beta},\boldsymbol{\gamma})\mathcal{H}_i^1(s)+\widetilde{\lambda_0}(s;\boldsymbol{\beta},\boldsymbol{\gamma})B_l(A_i+s)\mathcal{H}_i^2(s)ds,\end{aligned}$$

where

$$\begin{aligned}\mathcal{H}_i^1(t) &= e^{Z_i\boldsymbol{\beta}}E_{\boldsymbol{b}_i}\{E(\phi_i|\boldsymbol{b}_i,O_i,\boldsymbol{\theta}^{(k)})\eta(t|A_i,\boldsymbol{b}_i;\boldsymbol{\gamma})|\boldsymbol{\theta}^{(k)},O_i\},\\ \mathcal{H}_i^2(t) &= e^{Z_i\boldsymbol{\beta}}E_{\boldsymbol{b}_i}\{g'(t)\boldsymbol{b}_iE(\phi_i|\boldsymbol{b}_i,O_i,\boldsymbol{\theta}^{(k)})\eta(t|A_i,\boldsymbol{b}_i;\boldsymbol{\gamma})|\boldsymbol{\theta}^{(k)},O_i\},\end{aligned}$$

and

$$\widetilde{\lambda_0}^{\beta_j}(t;\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{\partial}{\partial\beta_j}\widetilde{\lambda_0}(t;\boldsymbol{\beta},\boldsymbol{\gamma}) = -\frac{[\sum_{i=1}^n\delta_iI(V_i=t)]\times[\sum_{i=1}^nI(V_i\geq t)Z_{ij}\mathcal{H}_i^1(t)]}{[\sum_{i=1}^nI(V_i\geq t)\mathcal{H}_i^1(t)]^2},$$

$$\widetilde{\lambda_0}^{\gamma_l}(t;\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{\partial}{\partial\gamma_l}\widetilde{\lambda_0}(t;\boldsymbol{\beta},\boldsymbol{\gamma}) = -\frac{[\sum_{i=1}^n\delta_iI(V_i=t)]\times[\sum_{i=1}^nI(V_i\geq t)B_l(A_i+t)\mathcal{H}_i^2(t)]}{[\sum_{i=1}^nI(V_i\geq t)\mathcal{H}_i^1(t)]^2}.$$

After the Newton-Raphson algorithm is converged, parameters $(\boldsymbol{\beta},\boldsymbol{\gamma})$ are updated as $(\boldsymbol{\beta}^{(k+1)},\boldsymbol{\gamma}^{(k+1)})$, and the baseline hazard function is updated as

$$\lambda_0^{(k+1)}(t) = \widetilde{\lambda_0}(t;\boldsymbol{\beta}^{(k+1)},\boldsymbol{\gamma}^{(k+1)}).$$

From $Q_2(\sigma^2;\boldsymbol{\theta}^{(k)})$ and $Q_3(\boldsymbol{\mu},\boldsymbol{D};\boldsymbol{\theta}^{(k)})$, the parameters can be solved in closed forms and we have the following updating formula:

$$\begin{aligned}\boldsymbol{\mu}^{(k+1)} &= \sum_{i=1}^{n}E[\boldsymbol{b}_i|\boldsymbol{\theta}^{(k)},O_i]/n,\\ \boldsymbol{D}^{(k+1)} &= \sum_{i=1}^{n}E[\boldsymbol{b}_i\boldsymbol{b}_i'|\boldsymbol{\theta}^{(k)},O_i]/n, \text{and}\\ (\sigma^2)^{(k+1)} &= \frac{\sum_{i=1}^n\sum_{j=1}^{m_i}E\left\{(\boldsymbol{Y}_i-\boldsymbol{G}_i\boldsymbol{b}_i)'(\boldsymbol{Y}_i-\boldsymbol{G}_i\boldsymbol{b}_i)|\boldsymbol{\theta}^{(k)},O_i\right\}}{\sum_{i=1}^n m_i}.\end{aligned}$$

EM Algorithm

We propose to implement the Expectation Maximization (EM) algorithm to derive the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$. The complete algorithm is described below.

Give initial values $\boldsymbol{\theta}^{(0)}$ based on a two-step approach as follows:

Step 1: fit a mixed effect model $Y(t|\boldsymbol{b}) = g'(t)\boldsymbol{b} + \epsilon(t)$, and use its estimated mean and covariance as $\boldsymbol{\mu}^{(0)}$ and $\boldsymbol{D}^{(0)}$, and the variance of error term as $(\sigma^2)^{(0)}$. This can be realized by using the "nlme" package in R.

Step 2(a): for a pre-specified r, fit the GOR model that only include $\boldsymbol{Z}$ as the covariates and use the estimated coefficients as $\boldsymbol{\beta}^{(0)}$. This can be done using the R package "TransModel". The initial values for $\boldsymbol{\gamma}^{(0)}$ are set to be 0.

Step 2(b): obtain the baseline survival estimate $\hat{S}^{(0)}(t)$ from the model in Step 2(a) by predicting the survival curves for $\boldsymbol{Z} = \boldsymbol{0}$, and $\lambda_0^{(0)}(t)$ is the gradient of $-\log(\hat{S}^{(0)}(t))$.

In the kth iteration,

E-step: approximate the conditional expectations described in Section 2.2 based on $\boldsymbol{O}$ and current estimate $\boldsymbol{\theta}^{(k)}$ using Gaussian-Hermite quadrature.

M-step: maximize the expectation of the log-likelihood functions $Q_1(\lambda_0(\cdot), \boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{\theta}^{(k)})$, $Q_2(\sigma^2; \boldsymbol{\theta}^{(k)})$ and $Q_3(\boldsymbol{\mu}, \boldsymbol{D}; \boldsymbol{\theta}^{(k)})$ and update the parameters $\boldsymbol{\theta}^{(k+1)}$ as described in Section 2.2.

Iterate the E-step and M-steps until $\sum(\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)})^2 < 0.001$ or $k > 100$.

Variance Estimation

After the EM algorithm converges, we have the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$. Let $\boldsymbol{\theta}^* = \boldsymbol{\theta} \setminus \lambda_0$ denote the vector of all the parameters except the baseline hazard function λ_0. Suppose the length of the vector $\boldsymbol{\theta}^*$ is m, the variance-covariance matrix

of $\hat{\boldsymbol{\theta}}^*$ is a $m \times m$ matrix, and can be estimated by inverting the observed information matrix based on the profile likelihood.

To be specific, we define $pl(\boldsymbol{\theta}^*) = \max_{\lambda_0} n^{-1} \sum_{i=1}^{n} pl_i(\boldsymbol{\theta}^*, \lambda_0)$ as the logarithm of the profile likelihood for $\boldsymbol{\theta}^*$, where $pl_i(\boldsymbol{\theta}^*, \lambda_0)$ denote the logarithm of the observed likelihood for subject i, $i = 1, \cdots, n$. Let $I(\boldsymbol{\theta}^*) = \{v_{ll'}\}$, $l, l' = 1, \cdots, m$ denote the observed information matrix for $\hat{\boldsymbol{\theta}}^*$. The element $v_{ll'}$ can be approximated by the second-order numerical difference of $pl(\boldsymbol{\theta}^*)$. Specifically,

$$v_{ll'} = \frac{(\boldsymbol{q}(\hat{\boldsymbol{\theta}}^* + h_n \boldsymbol{e}_l) - \boldsymbol{q}(\hat{\boldsymbol{\theta}}^*))'(\boldsymbol{q}(\hat{\boldsymbol{\theta}}^* + h_n \boldsymbol{e}_{l'}) - \boldsymbol{q}(\hat{\boldsymbol{\theta}}^*))}{h_n^2},$$

where $\boldsymbol{q}(\hat{\boldsymbol{\theta}}^*) = (pl_1(\hat{\boldsymbol{\theta}}^*), \cdots, pl_n(\hat{\boldsymbol{\theta}}^*))$ is the vector of profile likelihood functions being evaluated at $\hat{\boldsymbol{\theta}}^*$, $\boldsymbol{e}_l$ is the unit vector of length m that has the lth element being 1 and other elements being 0, and $h_n = O(1/\sqrt{n})$ is a pre-specified constant that is bounded by $1/\sqrt{n}$.

2.3 Simulation Study

We generate data for the proposed joint models. For the survival time, we generate from the GOR model

$$S(t|\boldsymbol{Z}_i) = \begin{cases} \exp\{-\int_0^t \lambda_0(s) \exp\left[\boldsymbol{Z}_i\boldsymbol{\beta} + \psi(A_i + s) \times W_i(s)\right] ds\}, & r = 0, \\ \left\{1 + r \int_0^t \lambda_0(s) \exp\left[\boldsymbol{Z}_i\boldsymbol{\beta} + \psi(A_i + s) \times W_i(s)\right] ds\right\}^{-1/r}, & r > 0. \end{cases}$$

The baseline distribution for $\lambda_0(\cdot)$ is assumed to be either Weibull with shape and scale parameters equals 2, or Lognormal with log mean 0 and log standard deviation 1. The varying coefficient function in the survival model is $\psi(t) = -0.2\sin(t)$. The baseline age A_i is generated from the standard normal distribution and two baseline covariates are included: Z_1 follows Uniform (0,2) and Z_2 follows Bernoulli (0.5). Coefficients for $\boldsymbol{Z} = (Z_1, Z_2)$ are set to be $\boldsymbol{\beta} = (1, -1)$. Different models with transformation parameter $r = 0, 0.5, 1$ and 2 are used.

A linear function for the fitness over time is assumed, i.e., $W_i(t) = b_{i0} + b_{i1}t$. The random effects $\boldsymbol{b}_i = (b_{i0}, b_{i1}) \sim N(\boldsymbol{\mu}, \boldsymbol{D})$, where $\boldsymbol{\mu} = (2, 1)$ and the covariance matrix $\boldsymbol{D} = \{v_{ij}\}$ is assumed to be $v_{ij} = I(i = j) + 0.5I(i \neq j)$, that is, the variances are 1 and the covariance is 0.5. Variance of the error terms is $\sigma^2 = 0.5$.

Right censoring time C is generated from the uniform distribution, $U(0, a)$, where a is adjusted to have 50% right censoring data. Subject i is assumed to have visits $0 = t_{i0} < t_{i1} < \cdots < t_{i\nu_i} < \min\{T_i, C_i\}$, and the length between two consecutive visits are set to be 0.1. Sample size of $n = 500$ is used and 1000 replications are made for each setting.

We use 5 nodes are in the gaussian hermite quadrature and $L = 3$ knots at the percentiles for the B-splines in estimating the varying coefficient function $\phi(t)$. The simulation results are summarized in Table 2.1 for Weibull baseline distribution and in Table 2.2 for Lognormal distribution, where we report the bias, empirical standard deviation (StDev), mean of the estimated standard error (StdErr) and the coverage probability (CP) of the 95% Wald confidence intervals. The bias of all the parameters are very small, the estimated standard errors based on the profile likelihood are close to the empirical estimates and the CP is close to the nominal level 0.95. The estimated baseline cumulative hazard functions are compared with the true curves in Figures 2.1 and 2.2 for Weibull baseline distribution and in Figures 2.5 and 2.6 for Lognormal baseline distribution. The varying coefficient functions $\psi(\cdot)$ are plotted in Figures 2.3 and 2.4 for Weibull baseline distribution and in Figures 2.7 and 2.8 for Lognormal baseline distribution. The solid lines in the plots are the mean of estimates, dashed lines are the true curve and the dotted lines are the 2.5 and 97.5 quantiles of the estimates. All the curves are found to be close to the truth. More settings with regard to different functions for the varying coefficient, different sample sizes and censoring proportions have been performed as well, which give similar findings.

Table 2.1 Simulation Results for Joint Models (Weibull)

Variable	Bias	StDev	StdErr	CP	Bias	StDev	StdErr	CP
	r=0				r=0.5			
β_1	0.005	0.125	0.124	0.944	-0.003	0.139	0.145	0.954
β_2	-0.013	0.138	0.141	0.948	-0.005	0.169	0.165	0.946
μ_0	-0.001	0.047	0.048	0.952	0.003	0.046	0.048	0.964
μ_1	-0.006	0.059	0.060	0.952	-0.008	0.058	0.057	0.944
σ^2	-0.004	0.010	0.010	0.934	-0.005	0.009	0.010	0.912
V_{11}	-0.007	0.069	0.075	0.968	-0.011	0.074	0.073	0.940
V_{12}	0.012	0.066	0.065	0.940	0.014	0.062	0.062	0.952
V_{22}	-0.008	0.099	0.099	0.938	-0.005	0.098	0.093	0.944
	r=1				r=2			
β_1	0.009	0.170	0.164	0.942	0.000	0.203	0.194	0.946
β_2	-0.007	0.188	0.187	0.946	0.025	0.228	0.223	0.940
μ_0	0.000	0.049	0.048	0.946	0.002	0.047	0.047	0.944
μ_1	-0.003	0.060	0.056	0.926	-0.001	0.058	0.054	0.926
σ^2	-0.006	0.009	0.009	0.886	-0.006	0.008	0.008	0.884
V_{11}	-0.008	0.071	0.072	0.942	-0.017	0.069	0.071	0.932
V_{12}	0.021	0.062	0.062	0.938	0.013	0.058	0.059	0.964
V_{22}	0.007	0.092	0.091	0.948	-0.007	0.087	0.084	0.924

2.4 Real Data Analysis

We include patients who were enrolled between 1970 and 1980, and being followed till 2004 in the Aerobics Center Longitudinal Study (ACLS) database. The main exposure variable is the cardiorespiratory fitness (fitness), which is quantified as the maximal treadmill time in minutes during a symptom limited exercise test. All tests were supervised by a physician and conducted in accord with standardized exercise testing procedures. As an objective measure of physical activity, fitness is a more reliable measure of recent activity levels than self-reported values. Other potential confounders we adjust in the model include gender, BMI, smoking and family history of CVD. Fatal outcomes (e.g. CVD mortality) were from mortality surveillance, principally through the National Death Index (NDI).

In order to assess the longitudinal effect of fitness on the CVD mortality, we apply

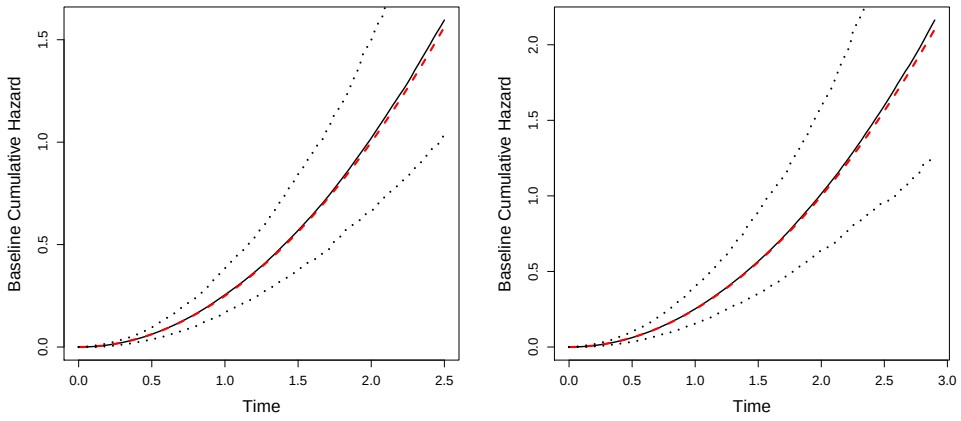

Figure 2.1 Estimated Baseline Cumulative Hazard Curves for Weibull Distribution (left: $r = 0$, right: $r = 0.5$).

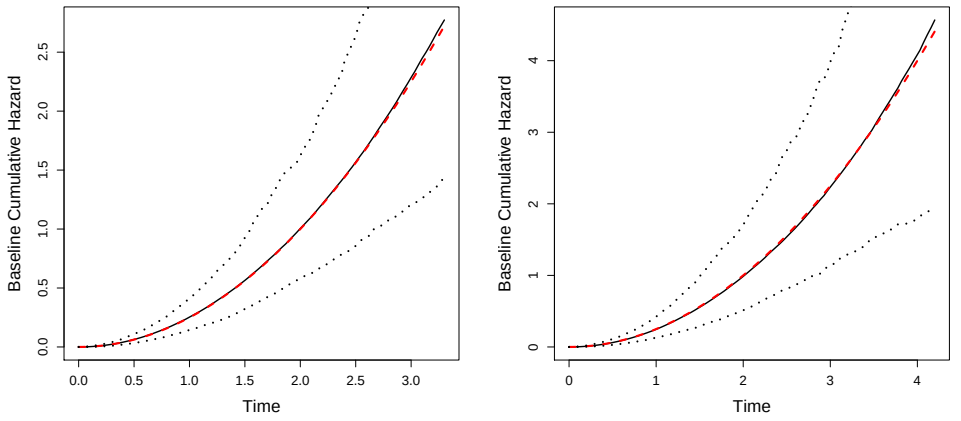

Figure 2.2 Estimated Baseline Cumulative Hazard Curves for Weibull Distribution (left: $r = 1$, right: $r = 2$).

the proposed model to the ACLS Data set. There are 3,980 patients and among them about 145 (3.64%) participants died because of CVD by year 2004. 437 patients are females and 3,543 of them are males. The number of follow-up visits for each participant ranges from 3 to 30 with median equals to 5.

We assume a linear form for the fitness trajectory over time. Similar to the simulation, we use Gaussian Hermite quadrature with 5 nodes for the approximation in the

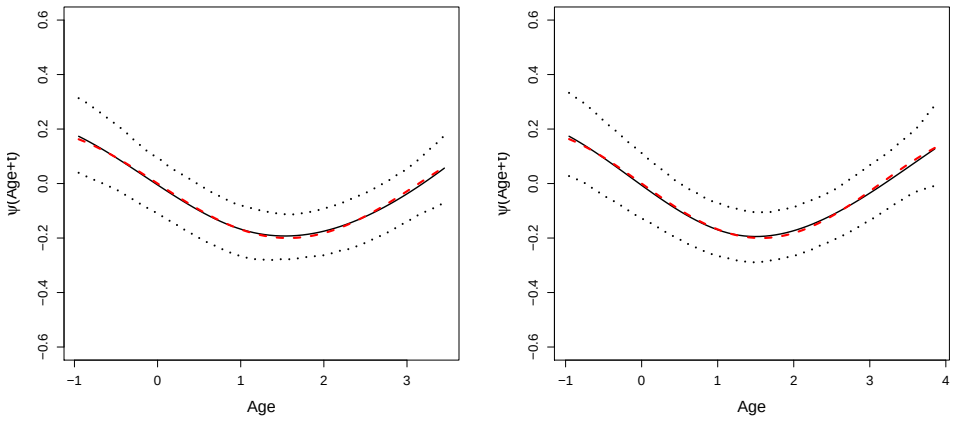

Figure 2.3 Estimated Varying Coefficient Curves $\psi(A(t))$ for Weibull Distribution (left: $r = 0$, right: $r = 0.5$).

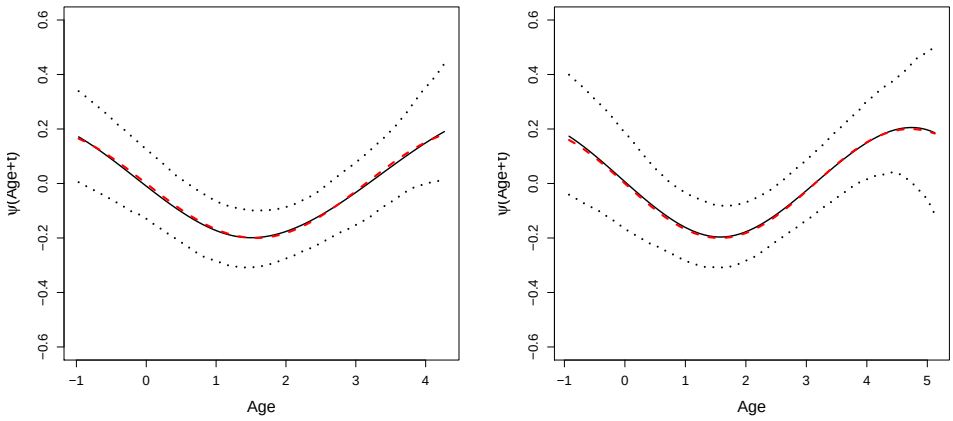

Figure 2.4 Estimated Varying Coefficient Curves $\psi(A(t))$ for Weibull Distribution (left: $r = 1$, right: $r = 2$).

E-step, which lead to similar results to using a larger number of nodes. We apply cubic B-splines with k knots being placed at percentiles to estimate the varying coefficient, where the number k can be selected based on the Akaike information criterion (AIC) in practice. For illustration, in Figure 2.9 we plot the AIC versus number of knots for three different models: a PH model ($r = 0$), a PO model ($r = 1$) and a variant of PO model ($r = 2$). Based on the curves, the PH model with 4 knots result in the smallest

Table 2.2 Simulation Results for Joint Models (Lognormal)

	r=0				r=0.5			
Variable	Bias	StDev	StdErr	CP	Bias	StDev	StdErr	CP
β_1	0.006	0.117	0.123	0.960	0.005	0.149	0.144	0.938
β_2	-0.001	0.130	0.139	0.948	-0.007	0.163	0.164	0.946
μ_0	0.000	0.050	0.050	0.938	0.004	0.050	0.049	0.954
μ_1	-0.006	0.093	0.083	0.924	-0.003	0.092	0.075	0.904
σ^2	-0.002	0.014	0.015	0.944	-0.002	0.013	0.013	0.944
V_{11}	-0.010	0.078	0.079	0.938	-0.009	0.076	0.077	0.938
V_{12}	0.007	0.079	0.090	0.972	0.014	0.078	0.080	0.944
V_{22}	0.003	0.159	0.159	0.942	0.010	0.135	0.135	0.958
	r=1				r=2			
β_1	0.008	0.165	0.160	0.932	-0.013	0.189	0.189	0.956
β_2	0.008	0.184	0.182	0.938	0.007	0.214	0.217	0.944
μ_0	0.003	0.046	0.049	0.960	0.004	0.047	0.048	0.938
μ_1	-0.001	0.082	0.068	0.894	0.001	0.075	0.061	0.896
σ^2	-0.004	0.011	0.011	0.948	-0.007	0.009	0.009	0.876
V_{11}	-0.009	0.074	0.075	0.938	-0.011	0.072	0.073	0.954
V_{12}	0.018	0.067	0.073	0.962	0.020	0.064	0.066	0.946
V_{22}	0.009	0.112	0.115	0.956	0.014	0.099	0.099	0.944

AIC.

Table 2.3 ACLS Data Analysis: Parameter Estimates in the PH Joint Model

Parameter	Estimate	StDev	P value
BMI	0.092	0.029	0.002
FamilyCVD	0.198	0.179	0.269
Smoke	0.167	0.234	0.476
Female	-0.442	0.386	0.253
μ_0	18.337	0.073	< 0.001
μ_1	0.063	0.005	< 0.001
σ^2	3.984	0.022	< 0.001
v_{11}	19.056	0.508	< 0.001
v_{12}	-0.280	0.026	< 0.001
v_{22}	0.052	0.002	< 0.001

We summarize the estimated coefficients in the PH model with 4 knots in Table 2.3. Based on the results, higher BMI will increase the risk of CVD mortality and females

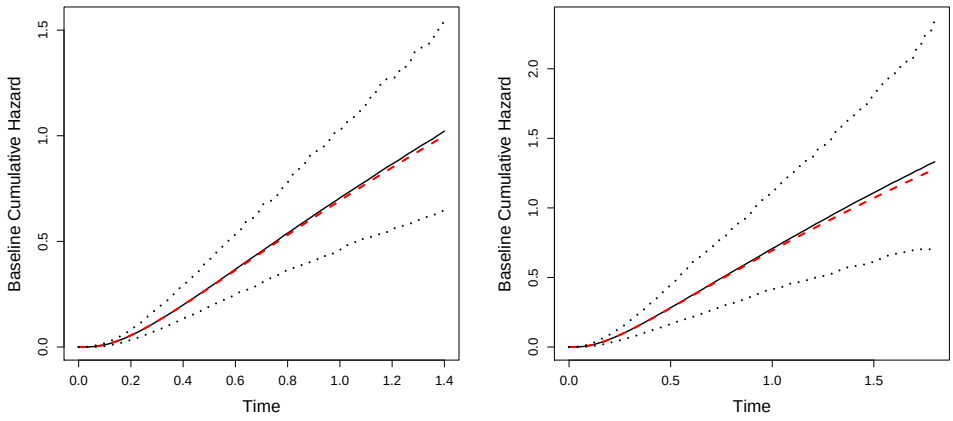

Figure 2.5 Estimated Baseline Cumulative Hazard Curves for Lognormal Distribution (left: $r = 0$, right: $r = 0.5$).

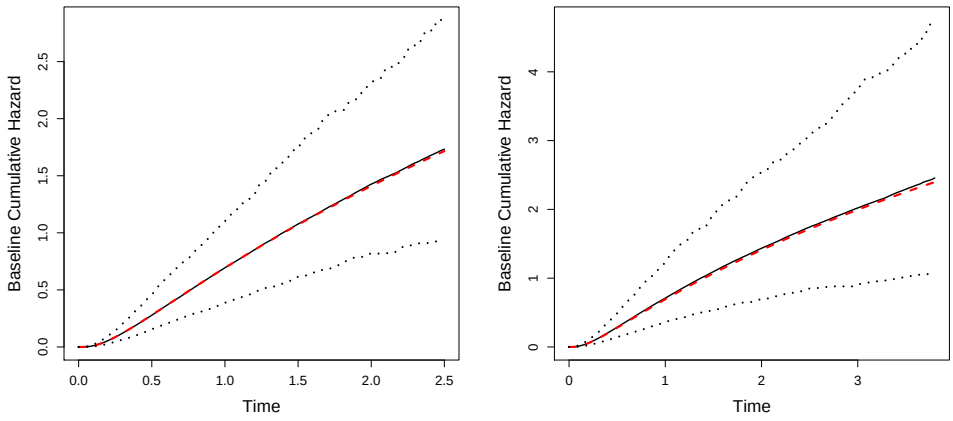

Figure 2.6 Estimated Baseline Cumulative Hazard Curves for Lognormal Distribution (left: $r = 1$, right: $r = 2$).

generally have lower risk of dying from CVD. Smoking and family history are positively associated with the risk of dying from CVD. All the terms in the longitudinal process are found to be highly significant here, indicating a significant linear trend of fitness over time. The baseline cumulative hazard curve is plotted in Figure 2.10 left panel, which is a step function with jumps at the event times.

Based on the estimated $\boldsymbol{\gamma}$ coefficients in B-splines, we can first test the hypothesis

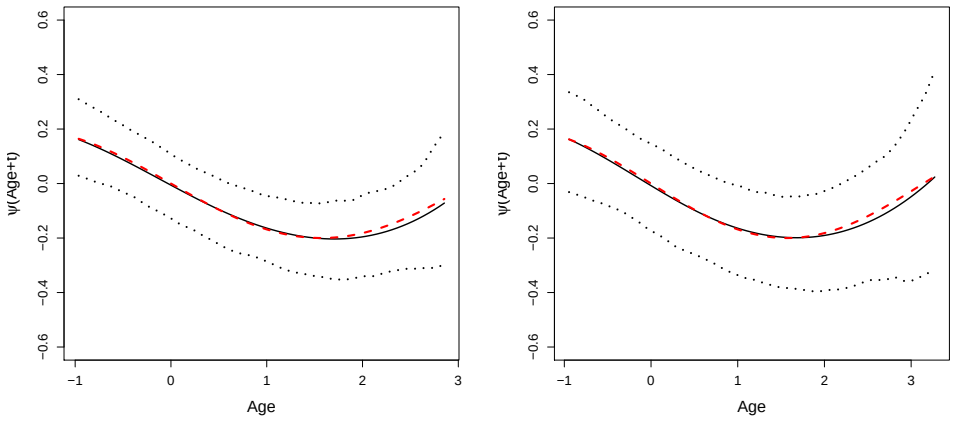

Figure 2.7 Estimated Varying Coefficient Curves $\psi(A(t))$ for Lognormal Distribution (left: $r = 0$, right: $r = 0.5$).

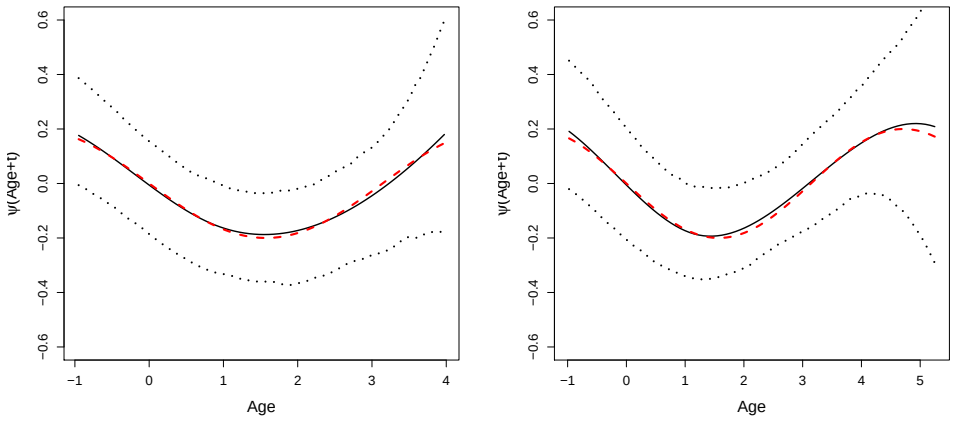

Figure 2.8 Estimated Varying Coefficient Curves $\psi(A(t))$ for Lognormal Distribution (left: $r = 1$, right: $r = 2$).

"H_0: the varying coefficient is constant over age", which is equivalent to $H_0 : M\boldsymbol{\gamma} = 0$, where

$$M = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

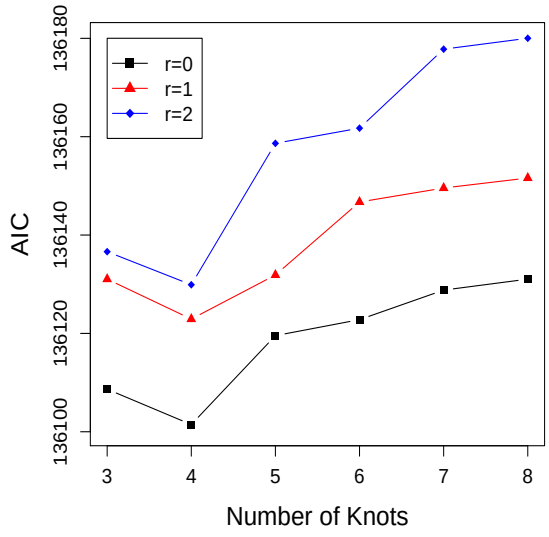

Figure 2.9 ACLS Data Analysis: Choose Knots and r Based On AIC.

when we apply 4 knots. The test statistic $(M\hat{\boldsymbol{\gamma}})(M\widehat{V_\gamma}M')^{-1}(M\hat{\boldsymbol{\gamma}})'$ follows the chi-squared distribution with 4 degree of freedom under H_0, where $\widehat{V_\gamma}$ is the estimated covariance matrix of $\hat{\boldsymbol{\gamma}}$. Specifically, the test statistic is calculated as 138.12 under the fitted model, and is greater than $\chi^2_{.95,4} = 9.49$. Therefore, we can conclude that the fitness effect on CVD mortality is significantly different over age.

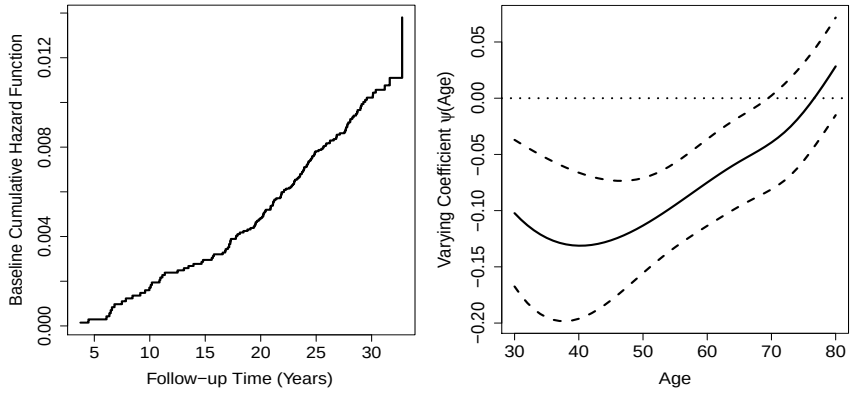

Figure 2.10 ACLS Data Analysis: Estimated Baseline Cumulative Hazard (left) and Age-dependent Varying Coefficient for Fitness (right).

The age-dependent varying coefficient curve with the 95% pointwise confidence intervals is plotted in Figure 2.10 right panel. Based on the curve, physical activity has significant protective effects on CVD mortality till age 70, no significant impact from 70 to 80. An explanation for this finding could be that after 70, age genetic factors take over as the dominate reason for CVD related mortality, and that the individuals' physical activity isn't really a factor anymore. The protective effect of physical activity is the strongest around age 40, suggesting that more exercise during middle-aged population is the most effective in reducing CVD associated mortality.

2.5 Discussions and Conclusions

We proposed a joint model with an age-dependent varying coefficient for GOR model with a longitudinal endogenous covariate measured with error. The age-related varying coefficient was flexibly modeled with cubic B-splines. The EM algorithm is applied in estimating the proposed joint model, while the variance of the estimates are approximated based on the profile likelihood function. The estimation methods are discussed and evaluated by simulation studies.

The ACLS dataset is used to illustrate the usage of the model, where we study the longitudinal effect of fitness on the CVD mortality. The effect of fitness on CVD mortality is found to change over age, and the trajectory can be clearly described by the estimated varying coefficient curve as illustrated in Section 2.4.

Chapter 3

Semiparametric Regression of the Illness-Death Model with Interval Censored Disease Incidence Time

The semi-competing model proposed by Fine et al. (2001) is the most popular framework for studying the disease-death process and their associations. The model describes a situation, where a subject can experience both a nonterminal event (e.g., disease) and a terminal event (e.g., death) during the life. The terminal event can censor the nonterminal event but not vice versa. For example, in the Aerobics Center Longitudinal Study (ACLS), participants have the risk of developing the cardiovascular disease (CVD) during their life. Whereas if a subject dies without CVD, his or her incidence time of CVD is not observable.

To study the semi-competing model, two general statistical frameworks have been proposed in the literature. The copula models (Hsieh and Huang, 2012; Li and Cheng, 2016; Yu, 2016; Zhou et al., 2017) assume the joint distribution between the nonterminal and terminal events under the condition that the nonterminal event is dependently censored by the terminal event. Such structure facilitates the estimation of the association between the two events. However, it makes the interpretation of the marginal distribution of the nonterminal event hypothetical and complicates the analysis of covariate effects (Xu et al., 2010).

Another way of viewing the problem is to consider a multi-state modeling framework

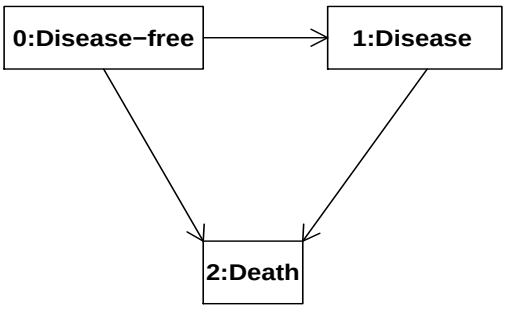

Figure 3.1 Semi-competing Diagram

with three states: 0 = disease-free, 1 = disease and 2 = death. As illustrated in Figure 3.1, there are three possible transitions between these three states and the disease process is irreversible, which means a person in state 1 can not go back to state 0 in the future. As a result, for healthy people in state 0, they have two possible survival paths from disease-free to death: die without disease ($0 \to 2$) or die with disease ($0 \to 1 \to 2$). The common interest in a multi-state model lies in the transition intensities among the three states. Markov models have been studied extensively in the literature, where the transition time is assumed to depend on the subject's current state only Siannis et al. (2007); Barrett et al. (2011). Extensions such as the semi-Markovian and non-Markovian models allow the transition from disease to death to depend both on chronological time and disease duration Datta et al. (2000); Meira-Machado et al. (2006); de Uña-Álvarez and Meira-Machado (2015). More summaries and overviews on the semi-competing models can be found in reviews by Andersen and Keiding (2002) and Meira-Machado et al. (2009).

Most previously mentioned studies assume that the nonterminal event can be either exactly observed or right censored. However, this is not true in the observational studies. For example, in the ACLS, the participants went to the Cooper Clinic in Dallas,

TX for periodic preventive medical examinations. Each participant had a sequence of follow-up visits, say $0 = v_0 < v_1 < \cdots < v_K < \infty$, and had the cardiovascular disease (CVD) either reported or diagnosed during each visit. As a result, we have intermittent CVD diagnosis information for each subject, and the true incidence time of CVD is either between two consecutive visits, say (v_{k-1}, v_k), or right censored.

In applications involving Markov processes and intermittent observations, Klein et al. (1984) and Kay (1986) applied simple parametric models with constant intensities. For the three-state disease model with periodic observations, Frydman (1992) proposed a self-consistent estimator, which extends the Turnbull's approach (Turnbull, 1976), to estimate the cumulative transition functions nonparametrically. Similar procedure was extended to the Markov illness-death model (Frydman, 1995a) and the non-Markov model (Frydman, 1995b). A common concern when studying the illness-death model with interval censored incidence time is that a subject could become diseased between two visits and thus die without being observed (Joly et al., 2002; Frydman and Szarek, 2009). This issue is ignorable when the lengths between intervals are relatively small compared with the disease development or the causes of death are available. In the ACLS data, the participants visit the clinic annually and we have both the death information and its major cause (CVD or other cause) for each participant from mortality surveillance, principally through the National Death Index (NDI). Therefore, if a patient died for reasons other than CVD, and all the previous diagnosis records showing the subject was CVD-free, then we assume the subject died without CVD.

We include 5236 CVD-free participants, who were enrolled in the ACLS during $1970 \sim 1980$, and follow them until the end of year 2004. We are interested in studying the following three problems based on the ACLS data: (1) estimate the transition intensities between the states including disease-free, CVD and death; (2) compare the

survival experience for subjects with and without CVD; and (3) explore the covariate effects in each transition process. To the best of our knowledge, there are no illness-death regression models that can study covariate effects and deal with interval censored incidence time. Therefore, we extend the self-consistent estimator proposed by Frydman (1995a) and apply the expectation-maximization (EM) algorithm to estimate the semiparametric illness-death regression model, and study the CVD-death process in the ACLS data.

The rest of the part is constructed as follows. The semi-competing risk model with interval censored data and the corresponding observed likelihood function are introduced in Section 3.1. The details of the estimation procedures are discussed in Section 3.2, where the derivation of the complete likelihood function and calculation of the conditional expectations are discussed. The complete EM algorithm and the corresponding variance estimation approach are presented at the end of the section. Extensive simulation studies based on the proposed method are performed and the results are discussed in Section 3.3. In Section 3.4, the proposed model is applied to estimate the covariate effects, such as age, fitness, smoking, etc., on the transition intensities among the three states: disease-free, CVD and death. Some final conclusions and limitations of the method are discussed in Section 3.5.

3.1 Semi-competing Risks Model

Notations

Let $\{X(t), t \geq 0\}$ denote a Markov process under the semi-competing model with a state space $\mathcal{S} = \{0, 1, 2\}$. $\mathcal{S}$ has three possible states: state 0 is disease-free, state 1 is illness and state 2 is death. We assume each subject has an initial state of 0 (i.e. $X(0) = 0$), and may or may not have disease during their life, and finally will enter

the absorb state of death. Define the transition probabilities as

$$P_{ll'}(s,t) = P(X(t) = l'|X(s) = l), \quad l \leq l' \quad \text{and} \quad l, l' \in \mathcal{S},$$

which is the probability of being in state l' at time t given that the subject is in state l at time s. The transition intensity from state l to state l' at time t is

$$\lambda_{ll'}(t) = \lim_{dt \to 0} \frac{P_{ll'}(t, t+dt)}{dt},$$

which is the instantaneous rate of moving from state l to state l'. The corresponding cumulative transition intensity function is then defined as

$$\Lambda_{ll'}(s,t) = \int_s^t \lambda_{ll'}(u)du.$$

Let T_1 and T_2 denote the time to disease and death, respectively. The distributions of T_1 and T_2 depend on $\boldsymbol{z}$, which is a vector of baseline covariates. We assume the following multiplicative proportional transition intensity model,

$$\lambda_{01}(t_1|\boldsymbol{z}) = \alpha_{01}(t_1) \times e^{\boldsymbol{\beta}'_{01}\boldsymbol{z}},$$

$$\lambda_{02}(t_2|\boldsymbol{z}) = \alpha_{02}(t_2) \times e^{\boldsymbol{\beta}'_{02}\boldsymbol{z}},$$

$$\lambda_{12}(t_2|\boldsymbol{z}, t_1) = \alpha_{12}(t_2) \times e^{\boldsymbol{\beta}'_{12}\boldsymbol{z}}, t_2 \geq t_1$$

where $\alpha_{ll'}(t)$ are the baseline intensity functions and $\boldsymbol{\beta}_{ll'}$ are the coefficients for $\boldsymbol{z}$ in the $l \to l'$ transition, $l \leq l'$ and $l, l' \in \mathcal{S}$. The first two models for T_1 and T_2 correspond to cause specific hazard functions for the competing risks part of the model in which either the disease or death happens first. The last model for $T_2|T_1$ is a Markov model, which assumes the death time for subjects who have disease depends on the observed disease time t_1, but not on the duration of the disease status.

Observed Likelihood

In observational studies, the exact disease time T_1 is unobservable and the disease information of each subject is followed through a sequence of examination visits $0 =$

$v_0 < v_1 < \cdots < v_K < \infty$. Therefore, we assume T_1 is interval censored and $[L, R)$ is the smallest observed interval that brackets T_1. If $L = 0$, T_1 is left censored; if $R = \infty$, T_1 is right censored; otherwise, T_1 is interval censored. Since the exact death information is commonly obtainable, we assume T_2 is only subject to right censoring, and the only case that T_2 is right censored is because the patient is still alive at the end of study (i.e. $T_2 > v_K$). Let $Y = \min(T_2, v_K)$ denote the last observation time. Let $\delta_1 = I(R < \infty)$ denote the indicator of observing disease by the last visit before death and $\delta_2 = I(T_2 \leq v_K)$ be an indicator of whether death is observed.

The observed data are $\mathcal{O} = \{(L_i, R_i, \delta_{1i}, Y_i, \delta_{2i}, \boldsymbol{z}_i), i = 1, \cdots, n\}$, and the parameters of interest in the model is $\boldsymbol{\theta} = (\boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{02}, \boldsymbol{\beta}_{12}, \alpha_{01}(\cdot), \alpha_{02}(\cdot), \alpha_{12}(\cdot))$. Next, we construct the likelihood function for $\boldsymbol{\theta}$ conditional on $\mathcal{O}$. We depict all possible scenarios in Figure 3.2, and describe them case-by-case.

Case 1: if subject i has $\delta_{1i} = 0$ and $\delta_{2i} = 0$, it means neither disease nor death happened at the end of study v_{i,K_i}. The observed data are $(L_i, R_i) = (v_{i,K_i}, \infty)$ and $Y_i = v_{i,K_i}$. The contribution to the likelihood is

$$P_{00}(0, L_i) = P_{00}(0, Y_i) = \exp\{-A_{01}(0, L_i)e^{\boldsymbol{\beta}_{01}'\boldsymbol{z}_i} - A_{02}(0, L_i)e^{\boldsymbol{\beta}_{02}'\boldsymbol{z}_i}\},$$

where $A_{ll'}(s, t) = \int_s^t \alpha_{ll'}(u)du$ is the baseline cumulative transition intensity function, $l \leq l'$ and $l, l' \in \mathcal{S}$.

Case 2: if subject i has $\delta_{1i} = 1$ and $\delta_{2i} = 0$, it means the disease is observed between two visits $(v_{i,k-1}, v_{i,k}]$ during the study, and the subject is still alive at the end of study v_{i,K_i}. The observed data are $(L_i, R_i) = (v_{i,k-1}, v_{i,k})$ and $Y_i = v_{i,K_i}$. The contribution to the likelihood is $P_{00}(0, L_i)P_{01}(L_i, R_i)P_{11}(R_i, Y_i)$, where

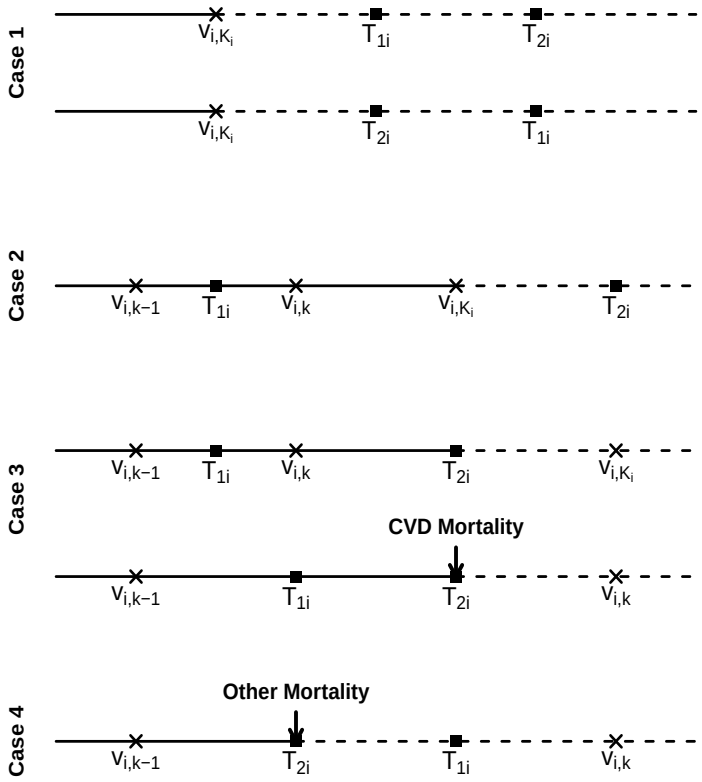

Figure 3.2 Possible follow-up cases (solid lines are observed and dashed lines are not observed)

$$P_{11}(R_i, Y_i) = \exp\{-A_{12}(R_i, Y_i)e^{\beta'_{12}\boldsymbol{z}_i}\}$$

$$P_{01}(L_i, R_i) = \int_{L_i}^{R_i} P_{00}(L_i, u)\lambda_{01}(u)P_{11}(u, R_i)du$$

$$= \int_{L_i}^{R_i} \alpha_{01}(u)e^{\beta'_{01}\boldsymbol{z}_i} \exp\left\{-A_{01}(L_i, u)e^{\beta'_{01}\boldsymbol{z}_i} - A_{02}(L_i, u)e^{\beta'_{02}\boldsymbol{z}_i} - A_{12}(u, R_i)e^{\beta'_{12}\boldsymbol{z}_i}\right\} du.$$

Case 3: if subject i has $\delta_{1i} = 1$ and $\delta_{2i} = 1$, there are two possibilities: (i) the disease was observed between two visits $(v_{i,k-1}, v_{i,k}]$, and the subject's death time T_{2i} was observed later; or (ii) The disease was not observed in the visit $v_{i,k-1}$, but the subject died before the next visit $v_{i,k}$ and the death is caused by the disease. The second scenario indicates the subject developed CVD between the visit $v_{i,k-1}$ and

death time T_{2i}. Therefore, the observed disease interval is $(L_i, R_i) = (v_{i,k-1}, v_{i,k})$ for scenario (i) and $(L_i, R_i) = (v_{i,k-1}, T_{2i})$ for scenario (ii), and the last observation time is $Y_i = T_{2i}$. The contribution to the likelihood based on either scenario is $P_{00}(0, L_i)P_{01}(L_i, R_i)P_{11}(R_i, Y_i)\lambda_{12}(Y_i)$, where $P_{11}(R_i, Y_i) = 1$ in the scenario (ii).

Case 4: if subject i has $\delta_{1i} = 0$ and $\delta_{2i} = 1$, we make the following assumption:
Assumption: If the disease is not observed till the last visit before death, i.e., $v_{i,k-1}$, and the subject dies for reasons other than the disease, we assume this subject dies without disease.

Based on the above assumption, we have observed data $(L_i, R_i) = (T_{2i}, \infty)$ and $Y_i = T_{2i}$, and the contribution to the likelihood is $P_{00}(0, L_i)\lambda_{02}(Y_i)$.

Combining the above Cases 1-4, the observed likelihood function can be written as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}|\mathcal{O}) &= \prod_{i=1}^{n} P_{00}(0, L_i) \times \left\{P_{01}(L_i, R_i)P_{11}(R_i, Y_i)\lambda_{12}^{\delta_{2i}}(Y_i)\right\}^{\delta_{1i}} \times \lambda_{02}(Y_i)^{(1-\delta_{1i})\delta_{2i}} \\ &= \prod_{i=1}^{n} P_{00}(0, L_i) \times P_{11}(R_i, Y_i)^{\delta_{1i}} \times \lambda_{12}(Y_i)^{\delta_{1i}\delta_{2i}} \times \lambda_{02}(Y_i)^{(1-\delta_{1i})\delta_{2i}} \\ &\qquad \times \left\{\int_{L_i}^{R_i} P_{00}(L_i, u)\lambda_{01}(u)P_{11}(u, R_i)du\right\}^{\delta_{1i}} \qquad (3.1)\end{aligned}$$

Direct maximization of the above observed likelihood function (3.1) is not feasible, because of the non-parametric baseline transition functions and the numerical integral. Therefore, we apply the EM algorithm and derive a self-consistent estimator, which extends the nonparametric estimator proposed by Frydman (1995a), for the proposed model with interval censored incidence time.

3.2 Estimation Procedures

Complete Likelihood Function

Let $0 = s_0 < s_1 < \cdots < s_M$ be a sequence of unique and ordered time points that contains all the observational time of disease or death, i.e., $\{(L_i, R_i < \infty, Y_i), i =$

$1, \cdots, n\}$. We introduce latent variables $N_{im} = I(T_{1i} \in (s_{m-1}, s_m])$ as the indicators of subject i having the disease in the subinterval $(s_{m-1}, s_m]$, $i = 1, \cdots, n; m = 1, \cdots, M$. Since the baseline transition intensity functions $\alpha_{01}(\cdot)$, $\alpha_{02}(\cdot)$ and $\alpha_{12}(\cdot)$ will be estimated nonparametrically, their maximum likelihood estimators are discrete functions that can only have positive values at the observed time points $s_m, m = 1, \cdots, M$. Therefore, conditional on these N_{im}'s, the integration part in the likelihood function (3.1) can be written as

$$\begin{aligned} P_{01}(L_i, R_i) &= \int_{L_i}^{R_i} P_{00}(L_i, u)\lambda_{01}(u)P_{11}(u, R_i)du \\ &= \sum_{m=1}^{M} N_{im} \times P_{00}(L_i, s_m)\lambda_{01}(s_m)P_{11}(s_m, R_i) \qquad (3.2) \\ &= \prod_{m=1}^{M} [P_{00}(L_i, s_m)\lambda_{01}(s_m)P_{11}(s_m, R_i)]^{N_{im}}, \qquad (3.3) \end{aligned}$$

where $N_{im} = 0$ if $s_m \leq L_i$ or $s_m > R_i$. The equations (3.2) and (3.3) are equivalent because only one of the N_{im}'s can be one for each subject, the rest will have to be zeros.

Replacing the integral part in the observed likelihood function (3.1) by (3.3), we derive the following complete likelihood function given the observed data $\mathcal{O}$ and latent indicators $\boldsymbol{N} = \{N_{im}, i = 1, \cdots, n; m = 1, \cdots, M\}$:

$$\begin{aligned} \mathcal{L}_c(\boldsymbol{\theta}|\boldsymbol{N}, \mathcal{O}) = &\prod_{i=1}^{n} P_{00}(0, L_i) \times P_{11}(R_i, Y_i)^{\delta_{1i}} \times \lambda_{02}(Y_i)^{(1-\delta_{1i})\delta_{2i}} \times \lambda_{12}(Y_i)^{\delta_{1i}\delta_{2i}} \\ &\times \left\{ \prod_{m=1}^{M} [P_{00}(L_i, s_m)\lambda_{01}(s_m)P_{11}(s_m, R_i)]^{N_{im}} \right\}^{\delta_{1i}}. \qquad (3.4) \end{aligned}$$

Conditional Expectations

Let $\boldsymbol{\theta}^{(d)}$ denote the updated vector of parameters in the dth iteration. Under the multiplicative proportional transition intensities model, the conditional expectation of the complete log-likelihood function given the observed data $\mathcal{O}$ and current estimate

$\boldsymbol{\theta}^{(d)}$ can be written as the summation of three separate functions:

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(d)}) =& E\{\log \mathcal{L}_c(\boldsymbol{\theta})|\mathcal{O},\boldsymbol{\theta}^{(d)}\}\\ =& \mathcal{Q}_1(\boldsymbol{\beta}_{01},\alpha_{01};\boldsymbol{\theta}^{(d)}) + \mathcal{Q}_2(\boldsymbol{\beta}_{02},\alpha_{02};\boldsymbol{\theta}^{(d)}) + \mathcal{Q}_3(\boldsymbol{\beta}_{12},\alpha_{12};\boldsymbol{\theta}^{(d)}),\end{aligned}$$

where

$$\begin{aligned}\mathcal{Q}_1(\boldsymbol{\beta}_{01},\alpha_{01};\boldsymbol{\theta}^{(d)}) =& \sum_{i=1}^{n}\sum_{m=1}^{M}\delta_{1i}W_{im}^{(d)}\{\log\alpha_{01}(s_m) + \boldsymbol{\beta}_{01}'\boldsymbol{z}_i - A_{01}(L_i,s_m)e^{\boldsymbol{\beta}_{01}'\boldsymbol{z}_i}\}\\ & - A_{01}(0,L_i)e^{\boldsymbol{\beta}_{01}'\boldsymbol{z}_i},\\ \mathcal{Q}_2(\boldsymbol{\beta}_{02},\alpha_{02};\boldsymbol{\theta}^{(d)}) =& \sum_{i=1}^{n}(1-\delta_{1i})\delta_{2i}[\log\alpha_{01}(Y_i) + \boldsymbol{\beta}_{02}'\boldsymbol{z}_i] - A_{02}(0,L_i)e^{\boldsymbol{\beta}_{02}'\boldsymbol{z}_i}\\ & - \sum_{m=1}^{M}\delta_{1i}W_{im}^{(d)}A_{02}(L_i,s_m)e^{\boldsymbol{\beta}_{02}'\boldsymbol{z}_i},\\ \mathcal{Q}_3(\boldsymbol{\beta}_{12},\alpha_{12};\boldsymbol{\theta}^{(d)}) =& \sum_{i=1}^{n}\delta_{1i}\delta_{2i}[\log\alpha_{12}(Y_i) + \boldsymbol{\beta}_{12}'\boldsymbol{z}_i] - \delta_{1i}A_{12}(R_i,Y_i)e^{\boldsymbol{\beta}_{12}'\boldsymbol{z}_i}\\ & - \sum_{m=1}^{M}\delta_{1i}W_{im}^{(d)}A_{12}(s_m,R_i)e^{\boldsymbol{\beta}_{12}'\boldsymbol{z}_i},\end{aligned}$$

where $W_{im}^{(d)} = E[N_{im}|\mathcal{O},\boldsymbol{\theta}^{(d)}]$ is the conditional expectation of N_{im}. From the complete likelihood (3.4), it can be shown that $N_{i1},\cdots,N_{iM}$ conditionally follow the multinomial distribution given the observed data $\mathcal{O}$ and current estimated parameter $\boldsymbol{\theta}^{(d)}$. Therefore, we have

$$W_{im}^{(d)} = E[N_{im}|\mathcal{O},\boldsymbol{\theta}^{(d)}] = \frac{\delta_{1i}I(s_m\in(L_i,R_i])P_{00}^{(d)}(L_i,s_m)\lambda_{01}^{(d)}(s_m)P_{11}^{(d)}(s_m,R_i)}{\sum_{s_l\in(L_i,R_i]}P_{00}^{(d)}(L_i,s_l)\lambda_{01}^{(d)}(s_l)P_{11}^{(d)}(s_l,R_i)},$$

for $i=1,\cdots,n$, $m=1,\cdots,M$, where $\lambda_{ll'}^{(d)}$ and $P_{ll'}^{(d)}$ are $\lambda_{ll'}$ and $P_{ll'}$ with $\boldsymbol{\theta}$ being replaced by $\boldsymbol{\theta}^{(d)}$, $l\le l'$ and $l,l'\in\mathcal{S}$.

EM Algorithm

We implement the following algorithm to derive the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ in the proposed model:

Step 0: give the initial values $\boldsymbol{\beta}_{01}^{(0)} = \boldsymbol{\beta}_{02}^{(0)} = \boldsymbol{\beta}_{12}^{(0)} = \mathbf{0}$ and $\alpha_{01}^{(0)}(s_m) = \alpha_{02}^{(0)}(s_m) = \alpha_{12}^{(0)}(s_m) = 1/M$.

Step 1: calculate the conditional expectations $W_{im}^{(d)}$, $i = 1, \cdots, n; m = 1, \cdots, M$ based on current parameters $\boldsymbol{\theta}^{(d)}$.

Step 2: maximize the conditional expectations of the profile log-likelihood functions to get the updated parameters $\boldsymbol{\theta}^{(d+1)}$. This can be done in three different sub-steps:

Step 2.1: maximize $\mathcal{Q}_1(\boldsymbol{\beta}_{01}, \widetilde{\alpha}_{01}(\cdot); \boldsymbol{\theta}^{(d)})$ with respect to $\boldsymbol{\beta}_{01}$, where

$$\widetilde{\alpha}_{01}(s_l; \boldsymbol{\beta}_{01}) = \frac{\sum_{i=1}^{n} \delta_{1i} W_{il}^{(d)}}{\sum_{i=1}^{n} e^{\boldsymbol{\beta}_{01}' \boldsymbol{z}_i} \left\{ I(L_i \geq s_l) + \sum_{m=1}^{M} \delta_{1i} W_{im}^{(d)} I(L_i < s_l \leq s_m) \right\}},$$

to get the updated estimate $\boldsymbol{\beta}_{01}^{(d+1)}$, and the updated estimate $\alpha_{01}^{(d)}(\cdot)$ is obtained by replacing $\boldsymbol{\beta}_{01}$ by $\boldsymbol{\beta}_{01}^{(d+1)}$ in the expression of $\widetilde{\alpha}_{01}(\cdot)$.

Step 2.2: maximize $\mathcal{Q}_2(\boldsymbol{\beta}_{02}, \widetilde{\alpha}_{02}(\cdot); \boldsymbol{\theta}^{(d)})$ with respect to $\boldsymbol{\beta}_{02}$, where

$$\widetilde{\alpha}_{02}(s_l; \boldsymbol{\beta}_{02}) = \frac{\sum_{i=1}^{n} \delta_{2i}(1 - \delta_{1i}) I(Y_i = s_l)}{\sum_{i=1}^{n} e^{\boldsymbol{\beta}_{02}' \boldsymbol{z}_i} \left\{ I(L_i \geq s_l) + \sum_{m=1}^{M} \delta_{1i} W_{im}^{(d)} I(L_i < s_l \leq s_m) \right\}},$$

to get the updated estimate $\boldsymbol{\beta}_{02}^{(d+1)}$, and the updated estimate $\alpha_{02}^{(d)}(\cdot)$ is obtained by replacing $\boldsymbol{\beta}_{02}$ by $\boldsymbol{\beta}_{02}^{(d+1)}$ in the expression of $\widetilde{\alpha}_{02}(\cdot)$.

Step 2.3: maximize $\mathcal{Q}_3(\boldsymbol{\beta}_{12}, \widetilde{\alpha}_{12}(\cdot); \boldsymbol{\theta}^{(d)})$ with respect to $\boldsymbol{\beta}_{12}$, where

$$\widetilde{\alpha}_{12}(s_l; \boldsymbol{\beta}_{12}) = \frac{\sum_{i=1}^{n} \delta_{1i} \delta_{2i} I(Y_i = s_l)}{\sum_{i=1}^{n} \delta_{1i} e^{\boldsymbol{\beta}_{12}' \boldsymbol{z}_i} \left\{ I(R_i < s_l \leq Y_i) + \sum_{m=1}^{M} W_{im}^{(d)} I(s_m < s_l \leq R_i) \right\}},$$

to get the updated estimate $\boldsymbol{\beta}_{12}^{(d+1)}$, and the updated estimate $\alpha_{12}^{(d)}(\cdot)$ is obtained by replacing $\boldsymbol{\beta}_{12}$ by $\boldsymbol{\beta}_{12}^{(d+1)}$ in the expression of $\widetilde{\alpha}_{12}(\cdot)$.

Step 3: iterate the Step 1 and Step2 until convergence. The convergence criteria is defined as $(\boldsymbol{\theta}^{(d+1)} - \boldsymbol{\theta}^{(d)})'(\boldsymbol{\theta}^{(d+1)} - \boldsymbol{\theta}^{(d)}) < .001$ or $d > 100$.

Variance Estimation

After the EM algorithm converges, we have the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$. Let $\boldsymbol{\theta}^* = (\boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{02}, \boldsymbol{\beta}_{12})$ denote the vector of all the coefficients in the regression models.

Suppose the length of the vector $\boldsymbol{\theta}^*$ is m, the variance-covariance matrix of $\hat{\boldsymbol{\theta}}^*$ is a $m \times m$ matrix, and can estimated by inverting the observed information matrix based on the profile likelihood.

To be specific, we define $pl(\boldsymbol{\theta}^*) = \sum_{i=1}^n pl_i(\boldsymbol{\theta}^*) = \max_{\lambda_0} n^{-1} \sum_{i=1}^n q_i(\boldsymbol{\theta}^*, \lambda_0)$ as the logarithm of the profile likelihood for $\boldsymbol{\theta}^*$, where $q_i(\boldsymbol{\theta}^*, \lambda_0)$ denote the logarithm of the observed likelihood for subject i, $i = 1, \cdots, n$. Let $I(\boldsymbol{\theta}^*) = \{v_{ll'}\}$, $l, l' = 1, \cdots, m$ denote the observed information matrix for $\hat{\boldsymbol{\theta}}^*$. The element $v_{ll'}$ can be approximated by the second-order numerical difference of $pl(\boldsymbol{\theta}^*)$. Specifically,

$$v_{ll'} = \frac{(\boldsymbol{p}(\hat{\boldsymbol{\theta}}^* + h_n \boldsymbol{e}_l) - \boldsymbol{p}(\hat{\boldsymbol{\theta}}^*))'(\boldsymbol{p}(\hat{\boldsymbol{\theta}}^* + h_n \boldsymbol{e}_{l'}) - \boldsymbol{p}(\hat{\boldsymbol{\theta}}^*))}{h_n^2},$$

where $\boldsymbol{p}(\hat{\boldsymbol{\theta}}^*) = (pl_1(\hat{\boldsymbol{\theta}}^*), \cdots, pl_n(\hat{\boldsymbol{\theta}}^*))$ is the vector of individual profile log-likelihood functions being evaluated at $\hat{\boldsymbol{\theta}}^*$, $\boldsymbol{e}_l$ is the unit vector of length m that has the lth element being 1 and other elements being 0, and $h_n = O(1/\sqrt{n})$ is a pre-specified constant that is bounded by $1/\sqrt{n}$.

3.3 Simulation Study

We generate data for the semi-competing regression model:

$$\lambda_{ll'}(t|\boldsymbol{z}) = \alpha_{ll'}(t) e^{\beta_{ll'}^{(1)} z_1 + \beta_{ll'}^{(2)} z_2}, \quad l \leq l' \quad \text{and} \quad l, l' \in \{0, 1, 2\}.$$

The baseline hazards functions are assumed to be $\alpha_{01}(t) = 0.5\sqrt{t}$, $\alpha_{02}(t) = 0.2t^2$ and $\alpha_{12}(t) = 0.5t^2$. Two baseline covariates $z_1 \sim \text{Ber}(0.5)$ and $z_2 \sim \text{U}(-2, 2)$ are included, with corresponding regression coefficients $\boldsymbol{\beta}_{01} = (1, -1)$, $\boldsymbol{\beta}_{02} = (0, 1)$ and $\boldsymbol{\beta}_{12} = (0, -1)$, respectively.

Similar to the data generation procedure for competing risks data, we first generate T from the distribution function

$$F_T(t) = 1 - \exp\{-A_{01}(0, t) e^{\boldsymbol{\beta}_{01}' \boldsymbol{z}} - A_{02}(0, t) e^{\boldsymbol{\beta}_{02}' \boldsymbol{z}}\},$$

and then generate an indicator $\epsilon = I(T_1 < T_2)$ based on the generated T from the Bernoulli distribution with probability

$$P(\epsilon = 1) = \frac{\alpha_{01}(T)e^{\beta'_{01}\boldsymbol{z}}}{\alpha_{01}(T)e^{\beta'_{01}\boldsymbol{z}} + \alpha_{02}(T)e^{\beta'_{02}\boldsymbol{z}}}.$$

If $\epsilon = 1$, we have $T_1 = T$ and we further generate T_2 from the distribution

$$F_{T_2|T_1}(t_2|t_1) = 1 - \exp\{-A_{12}(t_1, t_2)e^{\beta'_{12}\boldsymbol{z}}\}.$$

If $\epsilon = 0$, we have $T_2 = T$ and $T_1 = \infty$.

The visit times are generated independently as follows: the total number of visits K is generated from $1 + \text{Poisson}(\nu)$, and the lengths between two consecutive visits $\tau_k = v_k - v_{k-1}, k = 1, \cdots, K$ are generated from Exponential(κ). The parameters ν and κ are adjusted to have 10% and 40% right censoring proportions, which corresponds to the case when neither disease nor death happened before the last visit, i.e., $T_1 > v_K$ and $T_2 > v_K$.

Based on the generated T_1, T_2 and visit times $0 = v_0 < v_1 < \cdots < v_K$, the observed data $(L, R, \delta_1, Y, \delta_2)$ are derived as follows: $Y = \min(T_2, v_K)$ and $\delta_2 = I(T_2 \leq v_K)$; $L = \max\{v_k : v_k < T_1, k = 1, \cdots, K\}$, $R = I(\mathcal{R}_0 \neq \emptyset)\min\{\mathcal{R}_0\} + I(\mathcal{R}_0 = \emptyset)\infty$, where $\mathcal{R}_0 = \{v_k : T_1 \leq v_k \leq Y, k = 1 \cdots, K\}$ and $\emptyset$ is the null set, and $\delta_1 = I(R < \infty)$.

The estimated simulation results are presented in Table 3.1, where we report the bias, empirical standard deviation (StDev), mean of the estimated standard error (StdErr) and the coverage probability (CP) of the 95% Wald confidence intervals. In all settings, we find the bias of all the parameters are very small, the estimated standard errors based on the profile likelihood are close to the empirical estimates and the CP is close to the nominal level 0.95. The performance gets better when the sample size is relatively large and the right censoring proportion is small. In Figure 3.3 to Figure 3.8, the estimated baseline cumulative transition functions are plotted as solid curves, and compared with the true curves, which are presented as the dashed lines.

Table 3.1 Simulation Results for Illness-Death Models

	$n = 200$				$n = 500$			
Variable	Bias	StDev	StdErr	CP	Bias	StDev	StdErr	CP
10% censoring								
$\beta_{01}^{(1)}$	0.016	0.211	0.209	0.947	0.016	0.126	0.128	0.960
$\beta_{01}^{(2)}$	-0.017	0.112	0.114	0.957	-0.004	0.064	0.069	0.968
$\beta_{02}^{(1)}$	0.013	0.301	0.301	0.953	0.007	0.187	0.182	0.952
$\beta_{02}^{(2)}$	0.039	0.231	0.225	0.945	0.017	0.135	0.134	0.953
$\beta_{12}^{(1)}$	-0.020	0.221	0.223	0.945	-0.007	0.134	0.136	0.953
$\beta_{12}^{(2)}$	-0.030	0.158	0.163	0.959	-0.013	0.095	0.097	0.956
40% censoring								
$\beta_{01}^{(1)}$	0.028	0.240	0.241	0.956	0.011	0.142	0.147	0.960
$\beta_{01}^{(2)}$	-0.021	0.128	0.130	0.957	-0.008	0.077	0.078	0.956
$\beta_{02}^{(1)}$	0.022	0.445	0.445	0.962	0.007	0.264	0.262	0.949
$\beta_{02}^{(2)}$	0.053	0.329	0.335	0.957	0.039	0.200	0.194	0.951
$\beta_{12}^{(1)}$	-0.013	0.357	0.348	0.948	-0.001	0.210	0.208	0.953
$\beta_{12}^{(2)}$	-0.069	0.274	0.263	0.938	-0.028	0.150	0.154	0.947

The 2.5% and 97.5% quantiles of the estimates are added as the dotted lines in the plots. All the curves are found to be close to the truth. More settings with regard to different functions for the baseline distribution have been performed as well, which give similar findings. We do not present all the results here due to the space limit.

3.4 Real Data Analysis

We include 5236 participants, who were enrolled between 1970 and 1980 and without CVD at the time of enrollment, from the Aerobics Center Longitudinal Study (ACLS) database. The participants were followed till the end of year 2004. The patients went to the clinic for periodic preventive medical examinations and for counseling regarding health and lifestyle behaviors.

All participants were disease-free at the beginning. During the study, each subject

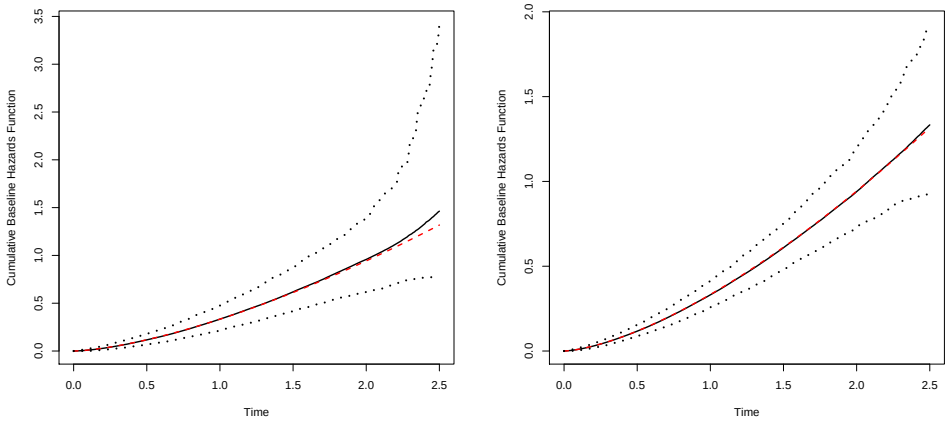

Figure 3.3 Estimated Baseline Cumulative Transition Functions α_{01} with 10% Censoring (left: n=200, right: n=500).

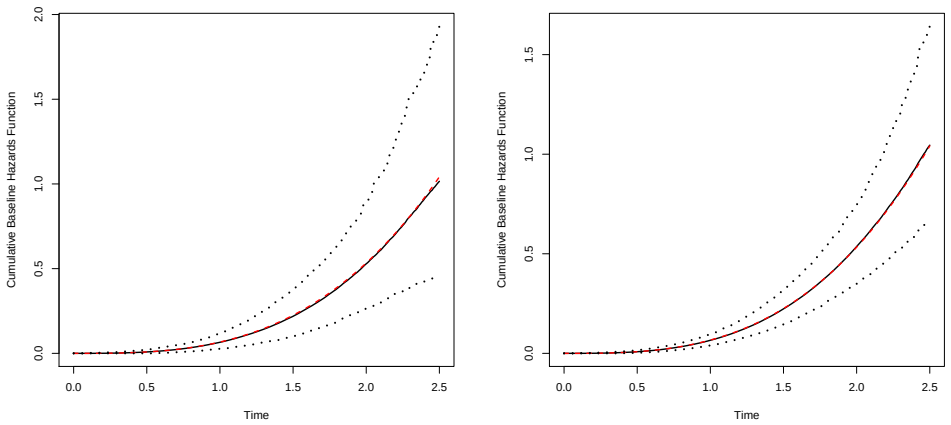

Figure 3.4 Estimated Baseline Cumulative Transition Functions α_{02} with 10% Censoring (left: n=200, right: n=500).

has the risk of developing CVD, or dies directly without CVD. We have both the death information and its major cause (CVD or other cause) for each participant from mortality surveillance, principally through the National Death Index (NDI). Based on the assumption we made in Section 3.1 case 4, if a subject died with no diagnosed CVD record and the cause of death was not CVD, we consider the subject died without CVD.

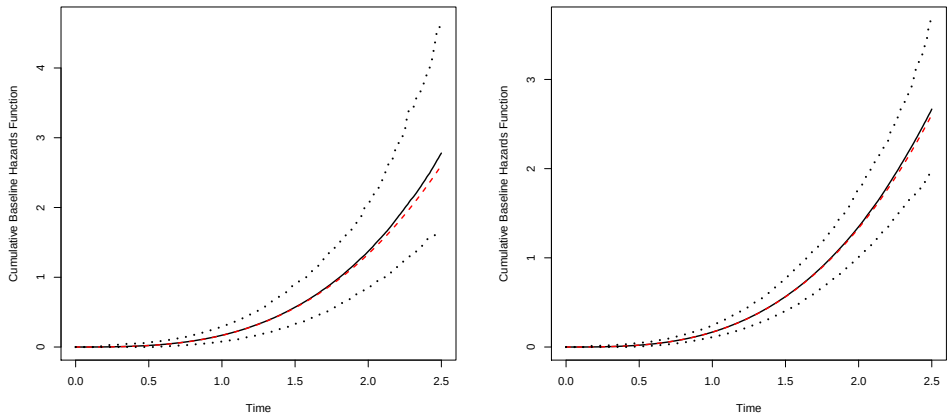

Figure 3.5 Estimated Baseline Cumulative Transition Functions α_{12} with 10% Censoring (left: n=200, right: n=500).

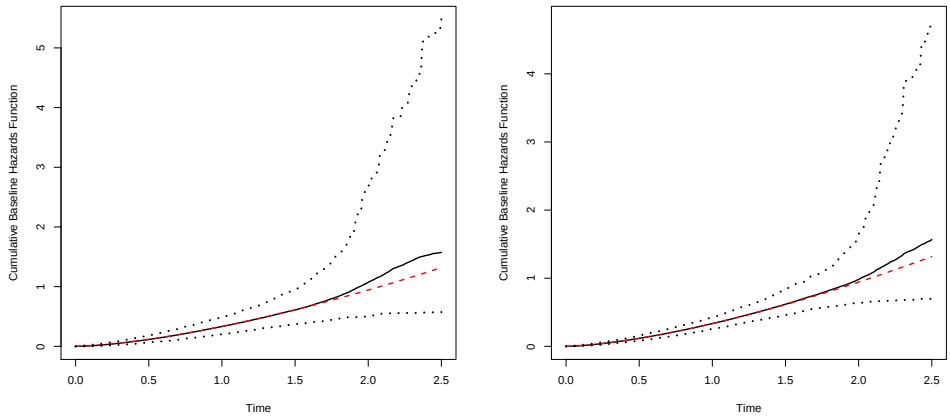

Figure 3.6 Estimated Baseline Cumulative Transition Functions α_{01} with 40% Censoring (left: n=200, right: n=500).

As a result, we have 353 (6.74%) have CVD diagnosed during the study and 274 out of them died eventually. There are 479 (9.15%) subjects died without CVD and 4404 (84.11%) were still alive and were CVD-free at their last follow-up visits. The chart for the distribution of participants was presented in Figure 1.2.

We study the covariates' effects on the transitions among the three states: disease-

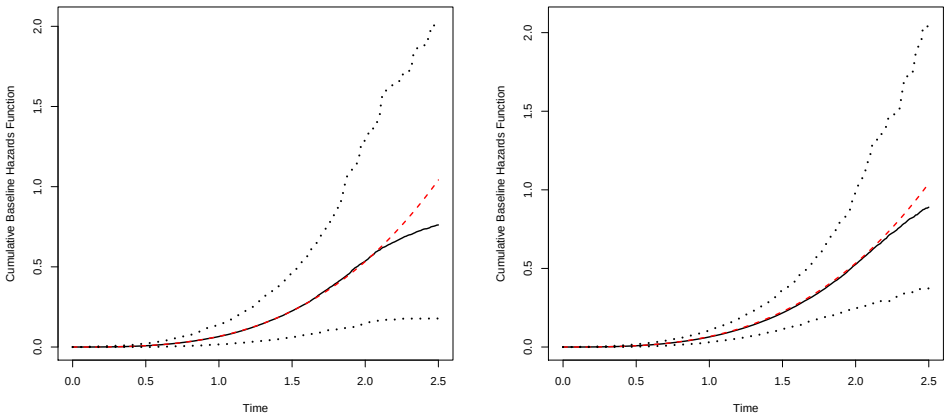

Figure 3.7 Estimated Baseline Cumulative Transition Functions α_{02} with 40% Censoring (left: n=200, right: n=500).

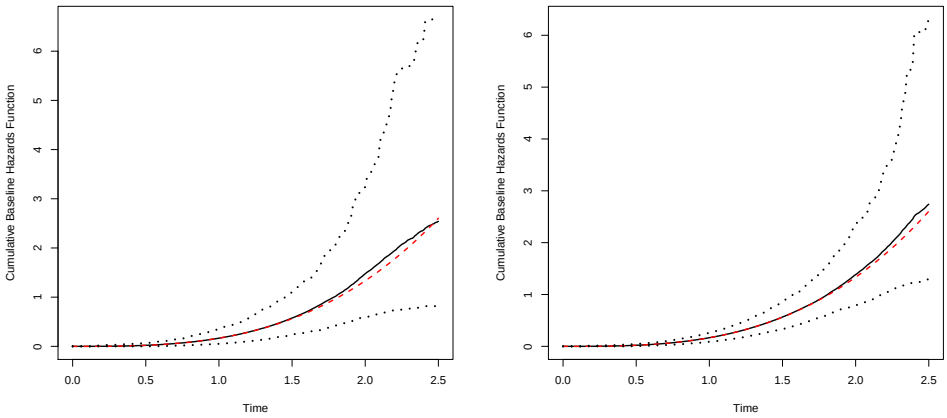

Figure 3.8 Estimated Baseline Cumulative Transition Functions α_{12} with 40% Censoring (left: n=200, right: n=500).

free, CVD and all-cause mortality (death). The covariates we are interested in include the baseline age, gender (1=female, 0=male), the average cardiorespiratory fitness (fitness), BMI and smoking status (1=smoker, 0=non-smoker). The estimated coefficients for these covariates are listed in Table 3.2. Generally, senior people have significantly greater intensity to transit to the states of CVD and death. Females have significantly

lower risk to derive CVD compared with males, but the risk to death is not significant between females to males. Fitness, which is an objective measure of the physical activity, has generally significant effect in reducing the transition intensity to both CVD and death. BMI plays a significant role among patients with CVD, and people with lower BMI tend to have longer survival. Smoking is positively related to both the development of CVD and death. Though it is not significant in the transition to CVD, it is marginally significant among both healthy subjects and subjects with CVD.

We also compare the estimated cumulative transition intensity curves in Figure 3.9. Categorical variables, gender and smoking status, are plotted separately with other continuous variables being set at the mean values. Generally, females have lower curves than males and smokers have higher curves than non-smokers among all the three transitions.

Table 3.2 ACLS Data Analysis: Estimated Coefficient in the Illness-Death Model

Transition	Variable	Coefficient	StDev	P-value
$Healthy \to CVD$	Age	0.097	0.007	0.000
	Gender	-1.297	0.286	0.000
	Fitness	-0.060	0.015	0.000
	BMI	0.032	0.021	0.126
	Smoke	0.086	0.131	0.510
$Healthy \to Death$	Age	0.087	0.006	0.000
	Gender	-0.234	0.164	0.154
	Fitness	-0.066	0.012	0.000
	BMI	0.024	0.016	0.127
	Smoke	0.181	0.109	0.097
$CVD \to Death$	Age	0.023	0.008	0.004
	Gender	-0.087	0.384	0.821
	Fitness	-0.038	0.015	0.015
	BMI	0.045	0.021	0.034
	Smoke	0.260	0.145	0.073

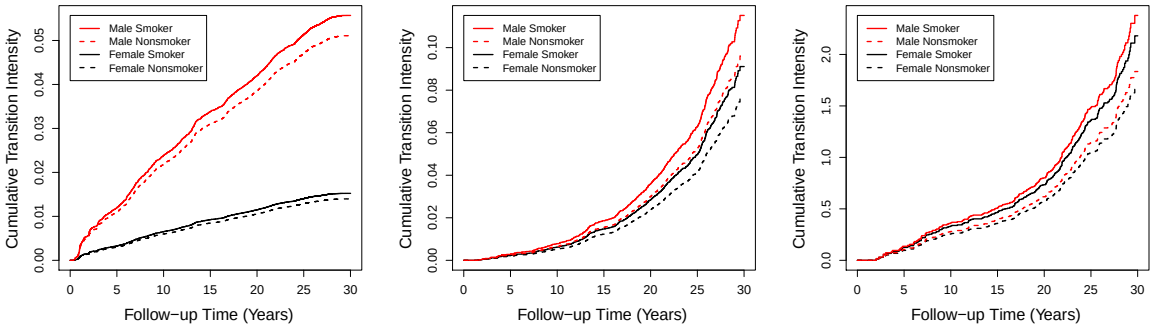

Figure 3.9 ACLS Data: Estimated Cumulative Transition Intensity Curves (from left to right: $Healthy \to CVD$, $Healthy \to Death$ and $CVD \to Death$).

3.5 Discussions and Conclusions

We extended Turnbull (1976)'s self-consistent estimator to the semi-parametric illness-death regression model, where the disease incidence time and the death time were subject to interval censoring and right censoring, respectively. The EM algorithm was applied to derive the estimates for both the coefficients and the baseline transition intensity functions. Numerical second derivative of the profile likelihood was used to approximate the observed information matrix and get the variance estimates.

Simulation studies with regard to different sample sizes and censoring proportions are performed for the proposed approach, and results from all settings suggest a good performance. The proposed method was then applied to the ACLS data to study the covariate effects, including age, gender, fitness, BMI and smoking status, in the transitions among disease-free, CVD and all-cause mortality.

Chapter 4

On Restricted Optimal Treatment Regime Estimation for Competing Risks Data

Generally, there is no uniformly best treatment for all patients because of individual heterogeneity. Personalized medicine is a paradigm that aims to tailor treatment which maximizes its effect according to patient's characteristics. The treatment effect may be determined by how well the treatment can improve the clinical outcomes of interest. For example, angiotensin converting enzyme inhibitors are evaluated regarding to how well it can control the blood pressure in hypertension studies; while in HIV studies, different antiretroviral agents are compared with respect to their ability in controlling the HIV viral load and CD4 counts. A number of works have been developed to address this question including Q-learning (Watkins and Dayan, 1992), A-learning (Murphy, 2003, 2005), direct value search methods (Zhang et al., 2012, 2013) and outcome-weighted learning (Zhao et al., 2012, 2015). The treatment can also target on survival time such as how long the HIV treatment can suppress the viral load under 200 copies/mL. When the primary endpoint to evaluate the treatment effect is survival time, Zhao et al. (2015) and Bai et al. (2017) proposed doubly robust estimators of the optimal treatment regime from a classification perspective, Jiang et al. (2017) proposed an optimal treatment regime estimation method that maximizes t-year survival probability, and Jiang et al. (2017) extended it to maximize a user-specified function of survival curve.

Our study is motivated by the HIV dataset obtained from Health Sciences South Carolina (HSSC). There are three most commonly used antiretroviral treatment (ART) classes: nucleoside reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs) and protease inhibitors (PIs) (Günthard et al., 2016). Drugs in the same class share common properties, whereas drugs in different class have different treatment effects. NNRTIs is associated with faster virologic suppression and PIs recover more CD4 cells (Organization, 2016; Günthard et al., 2016). In general, modern ART consists a combination of at least three agents from two classes (Günthard et al., 2016). Common combinations such as "NNRTIs+NRTIs" and "PIs+NRTIs" have also been compared in literature regarding to their treatment effects measured via the level of virologic suppression or CD4 recovery. (Staszewski et al., 1999; Haubrich et al., 2009; Smith et al., 2009; Borges et al., 2016). Jiang et al. (2017) considers the optimal regime of "NNRTIs+NRTIs" and "PIs+NRTIs" to maximize the longest initial treatment duration based on a data set from HIV/AIDS clinical observational study.

However, it is worthwhile pointing out that all ART drugs cause side effects. Some side effects, like headaches or occasional dizziness, may not be serious. Others such as swelling of the throat and tongue, damage to the liver and myocardial infarction, can be life-threatening (Dybul et al., 2002; Worm et al., 2010). The long-term side effects of these drugs include kidney problems, liver damage and nervous system/psychiatric effects such as insomnia, dizziness, depression and suicidal thoughts (Simpson et al., 2014). The evaluation for ART requires considerations of both treatment effects and side effects among different populations.

In HSSC data set, there are 426 patients took drug combinations "NNRTIs+NRTIs" or "PIs+NRTIs" with complete laboratory measures. We define "risk 1" as treatment or virologic failure, which is monitored by either CD4 counts ($\leq$ 500 cells/mm^3) or HIV viral load ($\geq$200 copies/mL), and "risk 2" as the drug-induced long-term side effects.

Days to either risk, whichever came first after drug administration, were recorded.

As discussed in Section 1.2, there is no consistent criteria to assign "NNRTIs+NRTIs" or "PIs+NRTIs" to each individual. How to achieve an optimal treatment regime which can balance between the treatment efficacy and side effects is a challenging problem in practice. More specifically, based on HIV patients' characteristics obtained through HSSC in SC, we aim to obtain an optimal treatment regime that can minimize the risk of treatment or virologic failure while controlling the risk of long-term side effects under a tolerable limit. Such optimal treatment regime can provide useful guidance for practitioners on antiretroviral drug consultation.

For simplicity, we only consider two risks here. The primary risk of interest is treatment failure, and the other is the risk due to adverse drug effects. Time to either risk is recorded and treated as the competing risk data. Estimation methods for competing risks data have been studied extensively in the literature (Gray, 1988; Fine and Gray, 1999; Klein and Andersen, 2005; Sun et al., 2006; Lu and Peng, 2008; Mao and Lin, 2017). However, to the best of our knowledge, there are no methods for estimating the optimal treatment regime under the competing risks framework. Therefore, there is an emerging need to develop an optimal individualized treatment regime achieving balance between risk 1 and risk 2 in practice. Toward this goal, we define a restricted optimal treatment regime that minimizes the t-year cumulative incidence function of the main risk while controlling the t-year cumulative incidence of the other risk under a predetermined level and derive its estimation procedure via a penalized value search method.

The rest of the part is organized as follows. In Section 4.1, the notation and definition of the restricted optimal treatment regime for competing risks data are introduced. An estimator for the proposed restricted optimal treatment regime is presented in Section 4.2, and the details of the implementation procedure are also discussed. Simulation

studies are conducted to examine the empirical properties of the proposed estimator in Section 4.3. An application to the HSSC HIV data is presented in Section 4.4. Finally, some conclusions and discussions are summarized in Section 4.5.

4.1 Model and Notations

Let T_1 and T_2 denote the event times of risks 1 and 2, respectively, and C denote the right-censoring time. Define $T = \min(T_1, T_2)$, $\tilde{T} = \min(T, C)$, $\delta = I(T \leq C)$ and $\epsilon = I(T = T_1) + 2I(T = T_2)$ as an indicator for competing risks. In addition, let X denote the p-dimensional vector of baseline covariates, and $\mathcal{X}$ denote the support for X. We assume two treatment options, $\mathcal{A} = \{0, 1\}$, are available. A treatment regime $d(x)$ is a mapping from $\mathcal{X}$ to $\mathcal{A}$, for example, under the linear decision rule, $d_\beta(x) = I(\beta'\tilde{x} > 0)$, where $\tilde{x} = (1, x)'$. The observed data include $O = \{(\tilde{T}_i, \delta_i, \delta_i\epsilon_i, A_i, X_i), i = 1, \cdots, n\}$.

To define the restricted optimal treatment regime, let $T_j^*(a)$ denote the potential event times for risk j, $j = 1, 2$, if treatment a was assigned to the patient, where $a \in \mathcal{A}$. Moreover, let $C^*(a)$ denote the potential censoring time. Define $T^*(a) = \min\{T_1^*(a), T_2^*(a)\}$, $\delta^*(a) = I\{T^*(a) \leq C^*(a)\}$ and $\epsilon^*(a) = I\{T^*(a) = T_1^*(a)\} + 2I\{T^*(a) = T_2^*(a)\}$. In addition, let $S^*(t; a) = P\{T^*(a) > t\}$ denote the survival function of $T^*(a)$. Then the cumulative incidence function of $T_j^*(a)$ is given by

$$F_j^*(t; a) = P\{T^*(a) \leq t, \epsilon^*(a) = j\} = \int_0^t S^*(u; a)\lambda_j^*(u; a)du, \quad j = 1, 2, \tag{4.1}$$

where $\lambda_j^*(t; a)$ is the cause-specific hazard function for $T_j^*(a)$.

Our proposed restricted optimal treatment regime is defined by

$$d^{opt} = \arg\min_{d \in \mathcal{D}} F_1^*(t_0; d),$$

where t_0 is a fixed time point of interest, and $\mathcal{D} = \{d : F_2^*(t_0; d) \leq \alpha\}$. In other words, the proposed restricted optimal treatment regime minimizes the t_0-year cumulative

incidence for risk 1 while controlling the t_0-year cumulative incidence of risk 2 under a pre-determined level α, $0 < \alpha < 1$. For simplicity, we only consider linear decision rules in this work, i.e., $d(x) = d_\beta(x) = I(\beta'\tilde{x} > 0)$. Then the restricted optimal linear decision rule is equivalent to finding β^{opt}, such that

$$\beta^{opt} = \arg\min_{\beta \in \mathcal{B}} F_1^*(t_0; \beta), \tag{4.2}$$

with the constraint $||\beta|| = 1$, where $\mathcal{B} = \{\beta : F_2^*(t_0; \beta) \leq \alpha\}$.

Let $\beta_j^* = \arg\min_\beta F_j^*(t_0; \beta)$ be the unrestricted estimators that minimize the t_0-year cumulative incidence of risk j, $j = 1, 2$. We have the following relation:

$$0 \leq F_2^*(t_0; \beta_2^*) \leq F_2^*(t_0; \beta_1^*) \leq 1.$$

If $0 \leq \alpha < F_2^*(t_0; \beta_2^*)$, $\mathcal{B}$ is a null set and there is no solution for β^{opt}; if $F_2^*(t_0; \beta_1^*) < \alpha \leq 1$, we have $\beta^{opt} = \beta_1^*$ because $\beta_1^* \in \mathcal{B}$. Otherwise, the restricted optimal treatment regime needs to be searched in the set $\mathcal{B}$.

4.2 Estimation Procedures

To estimate the restricted optimal treatment regime, we make the following three assumptions, as commonly used in the causal inference literature (Rubin, 1974). (i) The stable unit treatment value assumption (SUTVA): $T_j = AT_j^*(1) + (1-A)T_j^*(0)$, $j = 1, 2$ and $C = AC^*(1) + (1 - A)C^*(0)$. (ii) The no unmeasured confounder assumption: given covariates X, treatment assignment A is independent of potential survival times $T_j^*(a)$, $j = 1, 2$, and potential censoring times $C^*(a)$, for $a = 0, 1$. (iii) The independent censoring assumption: given covariates X and treatment assignment A, the survival times T_j, $j = 1, 2$ are independent of the censoring time C, and the censoring time C is independent of X and A.

Based on the first two assumptions, we have

$$F_j^*(t; a) = E_X\{F_j(t|A = a, X)\}, \quad j = 1, 2,$$

where $F_j(t|A=a,X) = P(T \leq t, \epsilon = j|A=a,X)$ is the conditional cumulative incidence function of risk j defined based on the observed data. The relatively restrictive censoring assumption that C is independent of X and A is needed to derive a simple, model-free estimator for the regime-specific cumulative incidence functions, which will be introduced shortly. A similar assumption was adopted in Jiang et al. (2017) for deriving estimators of the regime-specific survival function. Such an assumption can be relaxed; however, a model needs to be assumed for censoring times, and the inverse probability of the censoring-weighted technique needs to be used for constructing the estimator for the regime-specific cumulative incidence functions.

Next we propose consistent estimators for the cumulative incidence functions based on the definition in (4.1). Specifically, a consistent estimator for $F_j^*(t;\beta)$ is given by

$$\hat{F}_j(t;\beta) = \int_0^t \hat{S}(u;\beta) d\hat{\Lambda}_j(u;\beta), \quad j = 1, 2, \tag{4.3}$$

where $\hat{S}(u;\beta)$ and $\hat{\Lambda}_j(u;\beta)$ are consistent estimators for $S^*(u;\beta)$ and $\Lambda_j^*(u;\beta) = \int_0^u \lambda_j^*(s;\beta)ds$. Here $S^*(u;\beta) = S^*(u;d_\beta(\cdot))$ and $\lambda_j^*(u;\beta) = \lambda_j^*(u;d_\beta(\cdot))$.

Define the counting process of risk j as $N_{ij}(t) = I(\tilde{T}_i \leq t, \delta_i\epsilon_i = j)$ for $j = 1, 2$ and let $N_i(t) = N_{i1}(t) + N_{i2}(t)$. Moreover, let $Y_i(t) = I(\tilde{T}_i \geq t)$ denote the at-risk process for subject i. Under the assumed independent censoring assumption, Jiang et al. (2017) proposed an inverse propensity score-weighted Kaplan-Meier estimator (IPSWKME) for the survival function under regime $d_\beta(\cdot)$, i.e.,

$$\hat{S}(u;\beta) = \prod_{s \leq u} \left\{ 1 - \frac{\sum_{i=1}^n \hat{w}_i(\beta) dN_i(s)}{\sum_{i=1}^n \hat{w}_i(\beta) Y_i(s)} \right\},$$

where

$$\hat{w}_i(\beta) = \frac{A_i I(\beta'\tilde{X}_i > 0) + (1-A_i) I(\beta'\tilde{X}_i \leq 0)}{A_i\hat{\pi}(X_i) + (1-A_i)\{1-\hat{\pi}(X_i)\}},$$

and $\hat{\pi}(X_i)$ is an estimator of the propensity score $\pi(X_i) = P(A_i = 1|X_i)$. It can be shown that as long as $\hat{\pi}(X_i)$ is a consistent estimator of $\pi(X_i)$, $\hat{S}(u;\beta)$ is a consistent

estimator of the overall survival function $S^*(u;\beta)$. In practice, the propensity score is either known by design, as in randomized clinical trials, or estimated from data based on a posited parametric model, such as a logistic regression model, or estimated nonparametrically using a kernel or tree regression.

Similarly, we propose a consistent estimator of $\Lambda_j^*(u;\beta)$ as

$$\hat{\Lambda}_j(u;\beta) = \sum_{i=1}^{n} \int_0^u \frac{\hat{w}_i(\beta) I(\delta_i \epsilon_i = j)}{\sum_{k=1}^n \hat{w}_k(\beta) Y_k(s)} dN_i(s), \quad j = 1, 2.$$

Then, a consistent estimator $\hat{F}_j(t;\beta)$ of $F_j^*(t;\beta)$ can be obtained based on (4.3).

Given the consistent estimator $\hat{F}_j(t;\beta)$, $j = 1, 2$, a natural estimator of the proposed restricted optimal treatment regime can be obtained by minimizing $\hat{F}_1(t_0;\beta)$ subject to the constraint $\hat{F}_2(t_0;\beta) \leq \alpha$. However, such a restricted optimization problem may not be easy to solve. Here, we propose an approximated solution by penalization. Specifically, we define the approximated solution for β as

$$\hat{\beta}^{opt} = \arg\min_{\beta} \{\hat{F}_1(t_0;\beta) + M[\hat{F}_2(t_0;\beta) - \alpha]_+\}, \tag{4.4}$$

where M is a large number, e.g., $M = 1000$, and $[c]_+ = cI(c > 0)$. Note that when $\hat{F}_2(t_0;\beta) \leq \alpha$, no penalization is added; however, when $\hat{F}_2(t_0;\beta) > \alpha$, a large penalty is added to the target function $\hat{F}_1(t_0;\beta)$ to encourage β to satisfy the constraint. It can be seen that as $M \to \infty$, the penalized optimization problem will become the original restricted optimization problem.

The penalized optimization problem defined in (4.4) may still be challenging because the estimators $\hat{F}_j(t_0;\beta)$ are not smooth functions of β, and the resulted solution may be trapped in local minima. Following Jiang et al. (2017), to reduce the bias due to discreteness, we apply kernel smoothing for the regime function $d_\beta(x) = \Phi(\eta'\tilde{x}/h)$ to obtain the smoothed estimators $\tilde{F}_j(t;\beta)$, $j = 1, 2$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and h is a bandwidth pa-

rameter that goes to 0 as $n \to 0$. As suggested in Jiang et al. (2017), we choose $h = c_0 n^{-1/3} \text{sd}(\beta' \tilde{X})$, where $c_0 = 4^{1/3}$.

After smoothing, the objective function can be optimized directly using standard software, such as the "optim" function in R. Because the objective function is not convex, the solution may still be sensitive to the initial values. In practice, we suggest to trying a sequence of different initial values with $||\boldsymbol{\beta}|| = 1$, such as the unit vectors with one of the elements as 1 and the others as 0.

4.3 Simulation Study

In our simulations, we consider two covariates, X_1 and X_2, which are generated independently from the uniform distribution $U(-2, 2)$. The treatment indicator A is generated from a logistic regression model. We consider two cases for the propensity score: $\text{logit}\{\pi(X)\} = X_1 - 0.5X_2$ (case 1) or $\text{logit}\{\pi(X)\} = X_1 - 0.5X_2 + X_1^2$ (case 2). In addition, we consider a proportional hazards model for the cumulative incidence function of risk 1 (Fine and Gray, 1999):

$$F_1(t|X_1, X_2, A) = P(T \leq t, \epsilon = 1|X_1, X_2, A) = 1 - \{1 - q(1 - e^{-t})\}^{\exp\{-X_1 + A(X_1 - X_2)\}},$$

where $q \in (0, 1]$ is a predetermined constant that controls the proportion of risk 1, i.e., $P(\epsilon = 1) = F_1(\infty|X_1, X_2, A) = 1 - (1 - q)^{\exp\{-X_1 + A(X_1 - X_2)\}}$. Given $\epsilon = 1$, the conditional cumulative distribution function for the survival time of risk 1 can be derived as

$$P(T \leq t|\epsilon = 1, X_1, X_2, A) = \frac{F_1(t|X_1, X_2, A)}{P(\epsilon = 1)} = \frac{1 - \{1 - q(1 - e^{-t})\}^{\exp\{-X_1 + A(X_1 - X_2)\}}}{1 - (1 - q)^{\exp\{-X_1 + A(X_1 - X_2)\}}}.$$

Given $\epsilon = 2$, the conditional cumulative distribution function for the survival time of risk 2 is chosen as the exponential with rate $\exp\{-A(2 + X_1 + 2X_2)\}$, i.e.,

$$P(T \leq t|\epsilon = 2, X_1, X_2, A) = 1 - \exp\{-te^{-A(2 + X_1 + 2X_2)}\}.$$

Then the cumulative incidence function for risk 2 is given by

$$\begin{aligned} F_2(t|X_1, X_2, A) &= P(T \leq t, \epsilon = 2|X_1, X_2, A) \\ &= (1-q)^{\exp\{-X_1 + A(X_1 - X_2)\}} \times \left[1 - \exp\{-te^{-A(2+X_1+2X_2)}\}\right]. \end{aligned}$$

To generate competing risk event times, we first generate ϵ (= 1 or 2) from a Bernoulli distribution with the success probability

$$P(\epsilon = 1) = 1 - (1-q)^{\exp\{-X_1 + A(X_1 - X_2)\}}.$$

Then we generate time T by

$$T = \begin{cases} -\log\left[1 - \frac{1}{q}\left\{1 - \left(1 - U\left[1 - (1-q)^{\exp\{-X_1 + A(X_1 - X_2)\}}\right]\right)^{\exp\{X_1 - A(X_1 - X_2)\}}\right\}\right], \epsilon = 1, \\ -\log(1-U) \times \exp\{A(2 + X_1 + 2X_2)\}, \epsilon = 2, \end{cases}$$

where U is generated from the uniform distribution $U(0,1)$. The censoring time C is generated from the uniform distribution $U(0,c)$, where c is chosen to yield 15% and 40% censoring proportions.

We consider three values for q: 0.2, 0.5 and 0.8, which give respectively about 35%, 62% and 76% risk 1 rates under the 15% censoring rate and about 28%, 47% and 55% risk 1 rates under the 40% censoring rate. Moreover, we set $t_0 = 2$. Because of the proportional hazards model formulation for the cumulative incidence function of risk 1, the unrestricted optimal treatment regime for risk 1, i.e., $d_{\beta_1^*}$ that minimizes $F_1^*(t_0; \boldsymbol{\beta})$, is independent of the value of t_0 and q and is given by $\boldsymbol{\beta}_1^* = (0, -0.707, 0.707)$ under the norm-1 constraint. However, the unrestricted optimal treatment regime for risk 2 is very complicated and not linear. Here we use the grid search method to find the unrestricted optimal linear decision rule for risk 2, i.e., $d_{\beta_2^*}$ that minimizes $F_2^*(t_0; \boldsymbol{\beta})$, which is dependent on the value of q. The resulting true parameter values of $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$, and their associated cumulative incidence values $F_j^*(t_0; \boldsymbol{\beta}_k^*)$, $j, k = 1, 2$, are reported in Table 4.1. In addition, a set of α values are selected based on the two cutoff points,

$F_2(t_0, \boldsymbol{\beta}_2^*)$ and $F_2(t_0, \boldsymbol{\beta}_1^*)$, that satisfy $F_2(t_0, \boldsymbol{\beta}_2^*) < F_2(t_0, \boldsymbol{\beta}_1^*)$. The first two values are within the range, whereas the last value is greater than $F_2(t_0, \boldsymbol{\beta}_1^*)$. The true parameter $\boldsymbol{\beta}^*$ in the proposed restricted optimal linear decision rule for different q and α values is also obtained using the grid search method and reported in Table 4.1. It can be seen that when $\alpha > F_2(t_0, \boldsymbol{\beta}_1^*)$, the restricted optimal linear decision rule becomes the unrestricted optimal linear decision rule, i.e., $\boldsymbol{\beta}^* = \boldsymbol{\beta}_1^*$.

Table 4.1 True Parameter Values for Unrestricted and Restricted Optimal Linear Decision Rules

	$\boldsymbol{\beta}_1^*$	$\boldsymbol{\beta}_2^*$	Restricted Regime $\boldsymbol{\beta}^*$		
$q = 0.2$			$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
β_0	0.000	0.794	0.661	0.302	0.000
β_1	-0.707	0.576	-0.050	-0.530	-0.707
β_2	0.707	0.192	0.748	0.792	0.707
$F_1^*(t_0; \boldsymbol{\beta})$	0.134	0.287	0.163	0.137	0.134
$F_2^*(t_0; \boldsymbol{\beta})$	0.470	0.189	0.300	0.400	0.470
$q = 0.5$			$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$
β_0	0.000	0.711	0.609	0.163	0.000
β_1	-0.707	0.702	-0.096	-0.604	-0.707
β_2	0.707	0.028	0.787	0.780	0.707
$F_1^*(t_0; \boldsymbol{\beta})$	0.307	0.586	0.353	0.310	0.307
$F_2^*(t_0; \boldsymbol{\beta})$	0.337	0.055	0.200	0.300	0.337
$q = 0.8$			$\alpha = 0.05$	$\alpha = 0.15$	$\alpha = 0.25$
β_0	0.000	0.569	0.728	0.374	0.000
β_1	-0.707	0.821	0.272	-0.336	-0.707
β_2	0.707	-0.035	0.629	0.864	0.707
$F_1^*(t_0; \boldsymbol{\beta})$	0.475	0.759	0.613	0.497	0.475
$F_2^*(t_0; \boldsymbol{\beta})$	0.216	0.011	0.050	0.150	0.216

When estimating the restricted optimal treatment regime using the proposed method, we consider three propensity score fittings: the true propensity score (denoted by "true score"), a standard logistic regression fit with X_1 and X_2 included as predictors (denoted by "logistic") and a tree-based fit (denoted by "tree"). Therefore, for case 1, the fitted logistic regression model for the propensity score is correctly specified, whereas for case 2, it is misspecified. The tree-based method is a nonparametric fit and

estimates the true propensity score consistently for both cases.

For each setting, we consider sample size $n = 500$ and run 1000 replications. The simulation results for $q = 0.2$, 0.5 and 0.8 are presented from Table 4.2 to Table 4.7, with 15% and 40% right censoring cases, respectively. The following values are reported: (1) the mean and empirical standard deviation (SD) of the estimated $\boldsymbol{\beta}$ coefficients; (2) the mean and SD of the estimated t_0-year cumulative incidence of risk 1, $\hat{F}_1(t_0, \hat{\beta}^{opt})$, under the estimated restricted optimal treatment regime; (3) the mean and SD of the true t_0-year cumulative incidence of risk 1, $F_1(t_0, \hat{\beta}^{opt})$, under the estimated restricted optimal treatment regime computed using the Monte Carlo method based on a large simulated dataset; (4) the mean and SD of the proportion of making the correct decision (PCD) when comparing the estimated restricted optimal treatment regime with the true restricted optimal treatment regime; and (5) the proportion of the estimated t_0-year cumulative incidence of risk 2, $\hat{F}_2(t_0, \hat{\beta}^{opt})$, under the estimated restricted optimal treatment regime being controlled under the pre-determined level α.

From the results, we can see that in most scenarios, the estimated $\boldsymbol{\beta}$ coefficients and the estimated and true t_0-year cumulative incidence of risk 1 under the estimated optimal treatment regime are close to their true values. In addition, the bias in the estimated $\boldsymbol{\beta}$ coefficients gets smaller as the sample size increases and the censoring proportion decreases. The PCD is relatively high, ranging from 80% to 90%, and it also increases as the sample size increases and the censoring proportion decreases. In addition, the proportion of the estimated t_0-year cumulative incidence of risk 2 under the estimated restricted optimal treatment regime being controlled under the pre-determined level α is close to 1 in all of the settings. These results show that the restricted optimal treatment regime obtained by the proposed method can minimize the t_0-year cumulative incidence of risk 1 while controlling the t_0-year cumulative incidence of risk 2 under a predetermined level, as designed.

Comparing the three estimated restricted optimal treatment regimes obtained based on true propensity scores and the estimated propensity scores using logistic regression and the tree classification approach, we find that all the results are generally comparable for both cases 1 and 2, especially for small α values. When α is large, the estimated $\boldsymbol{\beta}$ coefficients based on the misspecified logistic regression of the propensity score under case 2 have relatively larger biases compared with the other two estimators, and the corresponding estimated restricted optimal treatment regimes have slightly smaller PCDs. This suggests that the proposed estimation method may have some robustness to the misspecification of the propensity score model, especially in terms of treatment decision.

4.4 Real Data Analysis

We apply the proposed method to an HIV dataset obtained from Health Sciences South Carolina (HSSC). Two combinations of HIV treatment are considered: "PIs+NRTIs" (regime 1) and "NNRTIs+NRTIs" (regime 2). There are a total of 426 HIV patients included in this study. Among them, 333 patients are assigned to regime 1 and 93 patients are in regime 2. In addition, 178 (41.8%) patients have either CD4 counts drop below 500 cells/mm^3 or HIV viral loads greater than 200 copies/mL (referred to as "risk 1"), and 151 (35.4%) patients have serious drug-induced side effects, such as liver damage, kidney problems or depression (referred to as "risk 2").

The survival time of interest is defined as days after drug administration to the occurrence of either risk, whichever comes first. The risk-type indicator is recorded as 1 ("risk 1") or 2 ("risk 2"). If neither risk occurred during the study period, the survival time is censored. The baseline characteristic for HIV patient includes standardized age, gender (male=1 or female=0), insurance type (government/commercial (GC) = 1 or other = 0) and race (black = 1 or white = 0).

Table 4.2 Simulation Results for Precision Medicine (q=0.2, 15% Censoring)

			Case 1			Case 2		
α	Parameter	Truth	True Score	Logistic	Tree	True Score	Logistic	Tree
0.3	β_0	0.66	0.58(0.165)	0.58(0.160)	0.53(0.167)	0.52(0.195)	0.55(0.163)	0.48(0.206)
	β_1	-0.05	-0.08(0.250)	-0.05(0.234)	-0.08(0.219)	-0.20(0.273)	-0.03(0.250)	-0.15(0.220)
	β_2	0.75	0.74(0.129)	0.75(0.122)	0.79(0.102)	0.74(0.163)	0.76(0.133)	0.80(0.117)
	$\hat{F}_1(t_0, \hat{\beta}^{opt})$	0.16	0.16(0.030)	0.16(0.029)	0.17(0.030)	0.15(0.030)	0.16(0.024)	0.15(0.026)
	$F_1(t_0, \hat{\beta}^{opt})$	-	0.19(0.017)	0.18(0.017)	0.18(0.015)	0.19(0.021)	0.19(0.020)	0.18(0.016)
	PCD	-	0.86(0.108)	0.85(0.113)	0.85(0.126)	0.87(0.099)	0.87(0.097)	0.87(0.112)
	$\hat{F}_2(t_0, \hat{\beta}^{opt}) \leq \alpha$	-	1	0.996	1	1	1	1
0.4	β_0	0.3	0.31(0.231)	0.31(0.216)	0.27(0.221)	0.25(0.249)	0.25(0.248)	0.18(0.265)
	β_1	-0.53	-0.45(0.192)	-0.45(0.187)	-0.45(0.170)	-0.46(0.232)	-0.31(0.198)	-0.45(0.209)
	β_2	0.79	0.77(0.119)	0.78(0.114)	0.80(0.109)	0.77(0.138)	0.85(0.121)	0.80(0.138)
	$\hat{F}_1(t_0, \hat{\beta}^{opt})$	0.14	0.13(0.026)	0.13(0.026)	0.14(0.027)	0.13(0.029)	0.14(0.025)	0.13(0.026)
	$F_1(t_0, \hat{\beta}^{opt})$	-	0.15(0.009)	0.15(0.008)	0.15(0.008)	0.15(0.013)	0.15(0.009)	0.15(0.011)
	PCD	-	0.92(0.073)	0.93(0.074)	0.94(0.078)	0.91(0.088)	0.90(0.084)	0.91(0.086)
	$\hat{F}_2(t_0, \hat{\beta}^{opt}) \leq \alpha$	-	1	1	1	1	1	1
0.5	β_0	0	0.08(0.261)	0.08(0.255)	0.04(0.266)	0.08(0.295)	0.17(0.294)	-0.01(0.316)
	β_1	-0.71	-0.62(0.181)	-0.62(0.178)	-0.61(0.172)	-0.57(0.230)	-0.37(0.220)	-0.58(0.233)
	β_2	0.71	0.70(0.142)	0.70(0.134)	0.72(0.128)	0.71(0.150)	0.82(0.154)	0.70(0.173)
	$\hat{F}_1(t_0, \hat{\beta}^{opt})$	0.13	0.13(0.027)	0.13(0.027)	0.13(0.026)	0.12(0.031)	0.14(0.026)	0.13(0.026)
	$F_1(t_0, \hat{\beta}^{opt})$	-	0.14(0.009)	0.14(0.007)	0.14(0.007)	0.14(0.010)	0.15(0.010)	0.14(0.011)
	PCD	-	0.86(0.114)	0.86(0.112)	0.86(0.125)	0.85(0.118)	0.81(0.099)	0.85(0.107)
	$\hat{F}_2(t_0, \hat{\beta}^{opt}) \leq \alpha$	-	1	1	1	1	1	1

Table 4.3 Simulation Results for Precision Medicine (q=0.2, 40% Censoring)

			Case 1			Case 2		
α	Parameter	Truth	True Score	Logistic	Tree	True Score	Logistic	Tree
0.3	β_0	0.66	0.57(0.191)	0.57(0.181)	0.53(0.184)	0.49(0.218)	0.54(0.183)	0.46(0.228)
	β_1	-0.05	-0.08(0.279)	-0.06(0.264)	-0.10(0.244)	-0.19(0.290)	-0.04(0.273)	-0.16(0.239)
	β_2	0.75	0.73(0.144)	0.74(0.128)	0.78(0.124)	0.76(0.151)	0.76(0.159)	0.80(0.129)
	$\hat{F}_1(t_0, \hat{\beta}^{opt})$	0.16	0.16(0.034)	0.16(0.033)	0.16(0.032)	0.15(0.034)	0.16(0.028)	0.15(0.029)
	$F_1(t_0, \hat{\beta}^{opt})$	-	0.19(0.019)	0.19(0.019)	0.18(0.017)	0.19(0.019)	0.19(0.021)	0.18(0.019)
	PCD	-	0.85(0.111)	0.85(0.115)	0.83(0.120)	0.85(0.101)	0.85(0.100)	0.86(0.113)
	$\hat{F}_2(t_0, \hat{\beta}^{opt}) \leq \alpha$	-	1	1	1	0.998	1	0.998
0.4	β_0	0.3	0.31(0.257)	0.32(0.236)	0.26(0.228)	0.25(0.270)	0.25(0.278)	0.18(0.287)
	β_1	-0.53	-0.43(0.204)	-0.44(0.202)	-0.44(0.198)	-0.46(0.236)	-0.30(0.223)	-0.44(0.228)
	β_2	0.79	0.77(0.132)	0.77(0.121)	0.79(0.118)	0.76(0.155)	0.84(0.147)	0.79(0.147)
	$\hat{F}_1(t_0, \hat{\beta}^{opt})$	0.14	0.13(0.028)	0.13(0.028)	0.14(0.029)	0.13(0.032)	0.14(0.028)	0.13(0.028)
	$F_1(t_0, \hat{\beta}^{opt})$	-	0.15(0.010)	0.15(0.010)	0.15(0.009)	0.15(0.014)	0.15(0.011)	0.15(0.012)
	PCD	-	0.92(0.075)	0.92(0.074)	0.92(0.090)	0.90(0.083)	0.88(0.094)	0.90(0.089)
	$\hat{F}_2(t_0, \hat{\beta}^{opt}) \leq \alpha$	-	1	1	0.998	0.998	1	1
0.5	β_0	0	0.09(0.285)	0.09(0.278)	0.05(0.274)	0.08(0.314)	0.16(0.314)	-0.00(0.322)
	β_1	-0.71	-0.60(0.195)	-0.60(0.200)	-0.61(0.182)	-0.55(0.238)	-0.37(0.240)	-0.57(0.244)
	β_2	0.71	0.70(0.154)	0.70(0.146)	0.71(0.143)	0.71(0.164)	0.80(0.189)	0.69(0.189)
	$\hat{F}_1(t_0, \hat{\beta}^{opt})$	0.13	0.12(0.029)	0.12(0.028)	0.13(0.029)	0.12(0.033)	0.14(0.029)	0.12(0.028)
	$F_1(t_0, \hat{\beta}^{opt})$	-	0.14(0.008)	0.14(0.009)	0.14(0.008)	0.14(0.011)	0.15(0.011)	0.14(0.011)
	PCD	-	0.85(0.114)	0.85(0.116)	0.85(0.121)	0.84(0.115)	0.81(0.109)	0.85(0.114)
	$\hat{F}_2(t_0, \hat{\beta}^{opt}) \leq \alpha$	-	1	1	1	1	1	1

Table 4.4 Simulation Results for Precision Medicine (q=0.5, 15% Censoring)

			Case 1			Case 2		
α	Parameter	Truth	True Score	Logistic	Tree	True Score	Logistic	Tree
0.2	β_0	0.61	0.50(0.201)	0.51(0.189)	0.42(0.194)	0.42(0.247)	0.40(0.225)	0.35(0.218)
	β_1	-0.1	-0.14(0.271)	-0.12(0.254)	-0.18(0.230)	-0.26(0.295)	-0.17(0.238)	-0.25(0.225)
	β_2	0.79	0.77(0.144)	0.78(0.137)	0.83(0.112)	0.76(0.180)	0.82(0.138)	0.84(0.124)
	$\hat{F}_1(t_0, \hat{\beta}^{opt})$	0.35	0.35(0.037)	0.35(0.036)	0.36(0.037)	0.34(0.041)	0.36(0.031)	0.34(0.035)
	$F_1(t_0, \hat{\beta}^{opt})$	-	0.38(0.018)	0.38(0.017)	0.38(0.019)	0.39(0.032)	0.38(0.021)	0.37(0.015)
	PCD	-	0.89(0.055)	0.90(0.051)	0.90(0.047)	0.87(0.075)	0.90(0.046)	0.92(0.040)
	$\hat{F}_2(t_0, \hat{\beta}^{opt}) \leq \alpha$	-	0.998	1	0.996	0.996	1	0.998
0.3	β_0	0.16	0.21(0.228)	0.20(0.220)	0.11(0.211)	0.17(0.279)	0.16(0.259)	0.09(0.260)
	β_1	-0.6	-0.52(0.190)	-0.54(0.175)	-0.53(0.158)	-0.50(0.233)	-0.35(0.182)	-0.50(0.216)
	β_2	0.78	0.76(0.135)	0.76(0.122)	0.79(0.111)	0.75(0.159)	0.86(0.135)	0.77(0.172)
	$\hat{F}_1(t_0, \hat{\beta}^{opt})$	0.31	0.31(0.034)	0.31(0.034)	0.32(0.037)	0.31(0.049)	0.34(0.036)	0.32(0.038)
	$F_1(t_0, \hat{\beta}^{opt})$	-	0.32(0.011)	0.32(0.009)	0.32(0.008)	0.33(0.023)	0.33(0.011)	0.32(0.011)
	PCD	-	0.90(0.049)	0.90(0.048)	0.91(0.041)	0.88(0.071)	0.88(0.045)	0.90(0.050)
	$\hat{F}_2(t_0, \hat{\beta}^{opt}) \leq \alpha$	-	1	1	1	1	1	1
0.4	β_0	0	0.02(0.232)	0.02(0.223)	-0.06(0.242)	0.06(0.275)	0.15(0.266)	-0.03(0.294)
	β_1	-0.71	-0.66(0.173)	-0.66(0.158)	-0.64(0.164)	-0.58(0.224)	-0.36(0.192)	-0.59(0.234)
	β_2	0.71	0.68(0.160)	0.68(0.149)	0.69(0.150)	0.71(0.166)	0.85(0.155)	0.68(0.205)
	$\hat{F}_1(t_0, \hat{\beta}^{opt})$	0.31	0.30(0.037)	0.30(0.036)	0.32(0.038)	0.30(0.056)	0.34(0.036)	0.31(0.040)
	$F_1(t_0, \hat{\beta}^{opt})$	-	0.31(0.008)	0.31(0.007)	0.31(0.007)	0.32(0.013)	0.32(0.011)	0.32(0.011)
	PCD	-	0.90(0.049)	0.91(0.047)	0.90(0.049)	0.88(0.061)	0.83(0.051)	0.87(0.056)
	$\hat{F}_2(t_0, \hat{\beta}^{opt}) \leq \alpha$	-	1	1	1	1	1	1

Table 4.5 Simulation Results for Precision Medicine (q=0.5, 40% Censoring)

			Case 1			Case 2		
α	Parameter	Truth	True Score	Logistic	Tree	True Score	Logistic	Tree
0.2	β_0	0.61	0.46(0.250)	0.46(0.241)	0.40(0.228)	0.38(0.284)	0.38(0.262)	0.32(0.258)
	β_1	-0.1	-0.15(0.299)	-0.15(0.297)	-0.20(0.264)	-0.28(0.315)	-0.18(0.280)	-0.25(0.276)
	β_2	0.79	0.76(0.176)	0.77(0.163)	0.81(0.147)	0.74(0.213)	0.80(0.183)	0.82(0.164)
	$\hat{F}_1(t_0, \hat{\beta}^{opt})$	0.35	0.34(0.046)	0.34(0.045)	0.35(0.047)	0.34(0.046)	0.35(0.040)	0.34(0.045)
	$F_1(t_0, \hat{\beta}^{opt})$	-	0.39(0.024)	0.38(0.023)	0.38(0.023)	0.39(0.032)	0.38(0.027)	0.38(0.020)
	PCD	-	0.88(0.067)	0.88(0.061)	0.88(0.065)	0.86(0.077)	0.89(0.057)	0.90(0.054)
	$\hat{F}_2(t_0, \hat{\beta}^{opt}) \leq \alpha$	-	1	0.998	1	0.998	0.998	1
0.3	β_0	0.16	0.21(0.274)	0.19(0.269)	0.12(0.252)	0.15(0.311)	0.16(0.285)	0.09(0.301)
	β_1	-0.6	-0.49(0.227)	-0.49(0.228)	-0.51(0.203)	-0.51(0.253)	-0.36(0.242)	-0.50(0.240)
	β_2	0.78	0.75(0.168)	0.76(0.156)	0.78(0.141)	0.72(0.196)	0.81(0.212)	0.74(0.214)
	$\hat{F}_1(t_0, \hat{\beta}^{opt})$	0.31	0.30(0.043)	0.30(0.043)	0.32(0.046)	0.30(0.054)	0.34(0.044)	0.31(0.046)
	$F_1(t_0, \hat{\beta}^{opt})$	-	0.33(0.016)	0.32(0.015)	0.32(0.012)	0.33(0.023)	0.33(0.015)	0.33(0.014)
	PCD	-	0.88(0.062)	0.88(0.059)	0.89(0.053)	0.87(0.072)	0.87(0.057)	0.89(0.054)
	$\hat{F}_2(t_0, \hat{\beta}^{opt}) \leq \alpha$	-	1	1	1	1	1	1
0.4	β_0	0	0.04(0.274)	0.03(0.274)	-0.06(0.274)	0.03(0.306)	0.14(0.293)	-0.03(0.317)
	β_1	-0.71	-0.63(0.206)	-0.64(0.197)	-0.62(0.184)	-0.59(0.233)	-0.39(0.251)	-0.59(0.249)
	β_2	0.71	0.67(0.189)	0.67(0.188)	0.68(0.179)	0.68(0.197)	0.79(0.238)	0.66(0.234)
	$\hat{F}_1(t_0, \hat{\beta}^{opt})$	0.31	0.29(0.045)	0.29(0.045)	0.31(0.047)	0.29(0.059)	0.33(0.044)	0.30(0.049)
	$F_1(t_0, \hat{\beta}^{opt})$	-	0.32(0.011)	0.32(0.011)	0.32(0.011)	0.32(0.014)	0.33(0.015)	0.32(0.013)
	PCD	-	0.88(0.059)	0.88(0.060)	0.89(0.060)	0.87(0.063)	0.82(0.061)	0.86(0.061)
	$\hat{F}_2(t_0, \hat{\beta}^{opt}) \leq \alpha$	-	1	1	1	1	1	1

Table 4.6 Simulation Results for Precision Medicine (q=0.8, 15% Censoring)

			Case 1			Case 2		
α	Parameter	Truth	True Score	Logistic	Tree	True Score	Logistic	Tree
0.05	β_0	0.73	0.67(0.196)	0.66(0.185)	0.65(0.183)	0.60(0.256)	0.65(0.219)	0.59(0.253)
	β_1	0.27	0.17(0.304)	0.19(0.289)	0.12(0.297)	-0.06(0.368)	-0.00(0.399)	-0.01(0.345)
	β_2	0.63	0.58(0.230)	0.60(0.219)	0.64(0.196)	0.61(0.248)	0.56(0.227)	0.64(0.233)
	$\hat{F}_1(t_0, \hat{\beta}^{opt})$	0.61	0.60(0.044)	0.60(0.042)	0.61(0.042)	0.57(0.051)	0.59(0.036)	0.58(0.044)
	$F_1(t_0, \hat{\beta}^{opt})$	-	0.64(0.017)	0.64(0.016)	0.64(0.013)	0.63(0.011)	0.64(0.015)	0.63(0.016)
	PCD	-	0.88(0.069)	0.89(0.065)	0.90(0.058)	0.90(0.066)	0.90(0.053)	0.91(0.054)
	$\hat{F}_2(t_0, \hat{\beta}^{opt}) \leq \alpha$	-	0.994	0.994	0.992	0.996	1	0.996
0.15	β_0	0.37	0.31(0.267)	0.30(0.263)	0.18(0.244)	0.20(0.319)	0.18(0.289)	0.14(0.276)
	β_1	-0.34	-0.34(0.277)	-0.34(0.262)	-0.38(0.215)	-0.42(0.289)	-0.38(0.208)	-0.43(0.247)
	β_2	0.86	0.78(0.157)	0.80(0.144)	0.84(0.125)	0.74(0.218)	0.81(0.207)	0.79(0.189)
	$\hat{F}_1(t_0, \hat{\beta}^{opt})$	0.5	0.49(0.038)	0.50(0.037)	0.51(0.041)	0.50(0.051)	0.53(0.036)	0.50(0.044)
	$F_1(t_0, \hat{\beta}^{opt})$	-	0.53(0.022)	0.52(0.020)	0.52(0.014)	0.54(0.039)	0.51(0.011)	0.52(0.016)
	PCD	-	0.88(0.063)	0.89(0.055)	0.90(0.074)	0.84(0.095)	0.91(0.061)	0.91(0.052)
	$\hat{F}_2(t_0, \hat{\beta}^{opt}) \leq \alpha$	-	1	0.998	1	0.998	1	1
0.25	β_0	0	-0.00(0.260)	-0.01(0.245)	-0.09(0.255)	0.04(0.281)	0.16(0.289)	-0.02(0.280)
	β_1	-0.71	-0.65(0.185)	-0.66(0.173)	-0.63(0.166)	-0.57(0.218)	-0.40(0.211)	-0.63(0.250)
	β_2	0.71	0.66(0.179)	0.66(0.168)	0.69(0.167)	0.72(0.185)	0.80(0.218)	0.64(0.246)
	$\hat{F}_1(t_0, \hat{\beta}^{opt})$	0.48	0.46(0.043)	0.46(0.043)	0.48(0.045)	0.47(0.076)	0.53(0.037)	0.48(0.051)
	$F_1(t_0, \hat{\beta}^{opt})$	-	0.48(0.007)	0.48(0.006)	0.48(0.007)	0.49(0.011)	0.50(0.011)	0.49(0.011)
	PCD	-	0.89(0.052)	0.89(0.052)	0.88(0.056)	0.87(0.057)	0.82(0.051)	0.85(0.057)
	$\hat{F}_2(t_0, \hat{\beta}^{opt}) \leq \alpha$	-	1	1	1	1	1	1

Table 4.7 Simulation Results for Precision Medicine (q=0.8, 40% Censoring)

			Case 1			Case 2		
α	Parameter	Truth	True Score	Logistic	Tree	True Score	Logistic	Tree
0.05	β_0	0.73	0.59(0.261)	0.59(0.265)	0.55(0.267)	0.48(0.334)	0.54(0.297)	0.48(0.325)
	β_1	0.27	0.10(0.360)	0.10(0.356)	0.08(0.349)	-0.12(0.405)	-0.05(0.412)	-0.03(0.396)
	β_2	0.63	0.61(0.266)	0.60(0.273)	0.67(0.234)	0.62(0.298)	0.62(0.267)	0.66(0.280)
	$\hat{F}_1(t_0, \hat{\beta}^{opt})$	0.61	0.55(0.068)	0.55(0.067)	0.56(0.066)	0.52(0.075)	0.54(0.057)	0.53(0.069)
	$F_1(t_0, \hat{\beta}^{opt})$	-	0.64(0.027)	0.64(0.029)	0.65(0.030)	0.65(0.035)	0.65(0.022)	0.65(0.036)
	PCD	-	0.85(0.079)	0.85(0.082)	0.88(0.081)	0.86(0.093)	0.88(0.064)	0.87(0.092)
	$\hat{F}_2(t_0, \hat{\beta}^{opt}) \leq \alpha$	-	0.996	0.988	0.99	0.994	0.994	0.996
0.15	β_0	0.37	0.25(0.341)	0.24(0.336)	0.10(0.325)	0.18(0.377)	0.18(0.350)	0.08(0.347)
	β_1	-0.34	-0.39(0.305)	-0.38(0.312)	-0.40(0.277)	-0.45(0.306)	-0.38(0.300)	-0.43(0.308)
	β_2	0.86	0.73(0.223)	0.73(0.222)	0.78(0.204)	0.68(0.271)	0.73(0.277)	0.73(0.255)
	$\hat{F}_1(t_0, \hat{\beta}^{opt})$	0.5	0.45(0.059)	0.45(0.058)	0.47(0.059)	0.46(0.066)	0.49(0.050)	0.46(0.058)
	$F_1(t_0, \hat{\beta}^{opt})$	-	0.53(0.026)	0.53(0.024)	0.52(0.021)	0.54(0.035)	0.52(0.022)	0.52(0.024)
	PCD	-	0.85(0.082)	0.86(0.079)	0.88(0.066)	0.84(0.097)	0.88(0.072)	0.88(0.085)
	$\hat{F}_2(t_0, \hat{\beta}^{opt}) \leq \alpha$	-	1	1	1	0.996	0.998	1
0.25	β_0	0	0.00(0.327)	0.01(0.324)	-0.12(0.319)	0.03(0.341)	0.13(0.356)	-0.04(0.341)
	β_1	-0.71	-0.63(0.234)	-0.64(0.238)	-0.60(0.216)	-0.57(0.266)	-0.42(0.295)	-0.59(0.299)
	β_2	0.71	0.63(0.225)	0.61(0.235)	0.66(0.215)	0.66(0.251)	0.71(0.292)	0.61(0.277)
	$\hat{F}_1(t_0, \hat{\beta}^{opt})$	0.48	0.42(0.060)	0.42(0.059)	0.44(0.059)	0.43(0.079)	0.48(0.051)	0.44(0.062)
	$F_1(t_0, \hat{\beta}^{opt})$	-	0.49(0.014)	0.49(0.015)	0.49(0.014)	0.49(0.018)	0.50(0.019)	0.50(0.019)
	PCD	-	0.86(0.069)	0.86(0.070)	0.86(0.071)	0.84(0.072)	0.80(0.066)	0.83(0.071)
	$\hat{F}_2(t_0, \hat{\beta}^{opt}) \leq \alpha$	-	1	1	1	1	1	1

First, we fit two models for the propensity score: a logistic regression and a tree classification approach. The ROC curves for the logistic and tree regression approaches are plotted in Figure 4.1 with the corresponding AUCs as 0.621 and 0.584 respectively. The logistic regression gives slightly better fit for the propensity score. We compute the proposed restricted optimal treatment regime based on both the logistic and tree regression fits of the propensity score, and the results are similar. Here, to save space, we present the results based on the logistic regression only.

We consider estimation of the restricted optimal treatment regime at one year, two years, three years and four years. That is, time $t_0 = 365$ days, 730 days, 1095 days and 1460 days after drug administration. In Table 4.8, we report the estimated coefficients $\hat{\boldsymbol{\beta}}_1^*$ and $\hat{\boldsymbol{\beta}}_2^*$ in the unrestricted optimal treatment regime, which minimize the t_0-year cumulative incidence of risk 1 and risk 2, respectively. Then, based on the range from $\hat{F}_2(t_0; \hat{\boldsymbol{\beta}}_1^*)$ to $\hat{F}_2(t_0; \hat{\boldsymbol{\beta}}_2^*)$, we select $\alpha = 0.4$ as a common value that is included in this range, and report the corresponding estimated coefficients $\hat{\boldsymbol{\beta}}^{opt}$ in the restricted optimal treatment regime. It can be seen that the t_0-year cumulative incidences of risk 2 under the estimated restricted optimal treatment regime are all controlled at level $\alpha = 0.4$ as desired; however, the t_0-year cumulative incidence of risk 2 ranges between 0.427 and 0.488 under the estimated unrestricted optimal treatment regime that minimizes the t_0-year cumulative incidence of risk 1. In addition, with the constraint on the cumulative incidence of risk 2, the estimated restricted optimal treatment regime increases the cumulative incidence of risk 1 compared with the estimated unrestricted optimal treatment regime. The magnitude of inflation depends on how stringent the posited constraint is on risk 2. For example, when $t_0 = 365$, $\hat{F}_2(t_0; \hat{\boldsymbol{\beta}}_1^*) = 0.427$ and the cumulative incidence of risk 1 increases from 0.355 to 0.366, whereas when $t_0 = 1460$, $\hat{F}_2(t_0; \hat{\boldsymbol{\beta}}_1^*) = 0.488$ and the cumulative incidence of risk 1 increases from 0.426 to 0.481.

In Table 4.9, the number of patients assigned to the two treatment regimes are

Table 4.8 Estimated Regimes for HIV study

	One Year($t_0 = 365$)			Two Year($t_0 = 730$)		
	$\hat{\boldsymbol{\beta}}_1^*$	$\hat{\boldsymbol{\beta}}_2^*$	$\hat{\boldsymbol{\beta}}^{opt}(\alpha = 0.4)$	$\hat{\boldsymbol{\beta}}_1^*$	$\hat{\boldsymbol{\beta}}_2^*$	$\hat{\boldsymbol{\beta}}^{opt}(\alpha = 0.4)$
Intercept	0.775	0.009	0.825	0.539	-0.477	0.831
Age	-0.129	0.646	-0.085	-0.389	0.389	0.200
Race	-0.508	0.472	-0.366	-0.642	0.703	-0.211
Insurance	-0.170	-0.528	-0.298	0.377	-0.351	-0.373
Gender	-0.310	-0.283	-0.298	-0.064	-0.056	-0.292
$\hat{F}_1(t_0; \boldsymbol{\beta})$	0.355	0.457	0.366	0.396	0.486	0.416
$\hat{F}_2(t_0; \boldsymbol{\beta})$	0.427	0.327	0.400	0.477	0.347	0.400
	Three Year($t_0 = 1095$)			Four Year($t_0 = 1460$)		
	$\hat{\boldsymbol{\beta}}_1^*$	$\hat{\boldsymbol{\beta}}_2^*$	$\hat{\boldsymbol{\beta}}^{opt}(\alpha = 0.4)$	$\hat{\boldsymbol{\beta}}_1^*$	$\hat{\boldsymbol{\beta}}_2^*$	$\hat{\boldsymbol{\beta}}^{opt}(\alpha = 0.4)$
Intercept	0.502	-0.505	-0.421	0.484	-0.505	0.567
Age	-0.399	0.381	0.243	-0.404	0.378	-0.218
Race	-0.657	0.707	0.212	-0.645	0.708	0.568
Insurance	0.396	-0.307	0.483	0.430	-0.310	0.114
Gender	-0.021	-0.071	0.697	-0.047	-0.069	-0.544
$\hat{F}_1(t_0; \boldsymbol{\beta})$	0.417	0.523	0.461	0.426	0.542	0.481
$\hat{F}_2(t_0; \boldsymbol{\beta})$	0.483	0.350	0.400	0.488	0.353	0.400

calculated and compared with the actual treatment received. The proportions where the treatment dictated by the estimated optimal treatment regime is consistent with the received treatment (cons%) are also reported. Patients received regime 1 more frequently than regime 2 in our dataset (333 vs 93), because PIs is considered to have slightly greater CD4 cell count recovery and lower antiretroviral drug resistance evolution with virologic failure in practice (Organization, 2016). Similar arguments about allocations of these two drugs were discussed in literature (Jiang et al., 2017). However, the picture can be quite different when we consider side effects and apply restricted optimal treatment regimes. The estimated unrestricted optimal treatment regime that minimizes the t_0-year cumulative incidence of risk 2 is found to be closer to the actual treatment assignment compared with the estimated unrestricted optimal treatment regime that minimizes the t_0-year cumulative incidence of risk 1 and the estimated restricted optimal treatment regimes. This also suggests that doctors are

likely to be conservative in practice and tend to assign drugs with lower risk of side effects.

Table 4.9 Comparing the Restricted Optimal Treatment Regime with Received Treatments

		Optimal Regimes					
Received Treatments		$\hat{\boldsymbol{\beta}}_1^*$		$\hat{\boldsymbol{\beta}}_2^*$		$\hat{\boldsymbol{\beta}}^{opt}(\alpha = 0.4)$	
		Regime 1	Regime 2	Regime 1	Regime 2	Regime 1	Regime 2
One Year	Regime 1	134	199	188	145	91	242
	Regime 2	35	58	44	49	20	73
	Cons(%)	45.1		55.6		38.5	
Two Year	Regime 1	105	228	201	132	67	266
	Regime 2	36	57	55	38	8	85
	Cons(%)	38.0		56.1		35.7	
Three Year	Regime 1	109	224	204	129	59	274
	Regime 2	36	57	55	38	14	79
	Cons(%)	39.0		56.8		32.4	
Four Year	Regime 1	107	226	204	129	13	320
	Regime 2	38	55	55	38	5	88
	Cons(%)	38.0		56.8		23.7	

We also estimate the restricted optimal treatment regime with a sequence of α values in the range from $\hat{F}_2(t_0; \hat{\boldsymbol{\beta}}_1^*)$ to $\hat{F}_2(t_0; \hat{\boldsymbol{\beta}}_2^*)$ and plot the number of patients assigned to each treatment by the estimated restricted optimal treatment regime in Figure 4.2. In the plot, the dashed vertical line corresponds to $\alpha = \hat{F}_2(t_0; \hat{\boldsymbol{\beta}}_2^*)$, beyond which the restricted optimal treatment regime is equivalent to the unrestricted optimal treatment regime. By comparing the two curves, we can see how treatments are distributed among patients under the restricted optimal treatment regime for different α values. In summary, the restricted optimal treatment regime tends to assign less patients to regime 2 (NNRTIs+NRTIs) than regime 1 (PIs+NRTIs) compared with the unrestricted optimal treatment regime for risk 1, especially when α is small. However, as the constraint level α increases, the number of patients assigned to regime 2 by the restricted optimal treatment regime increases and reaches its maximum for some α value, and

then drops to that of the unrestricted optimal treatment regime.

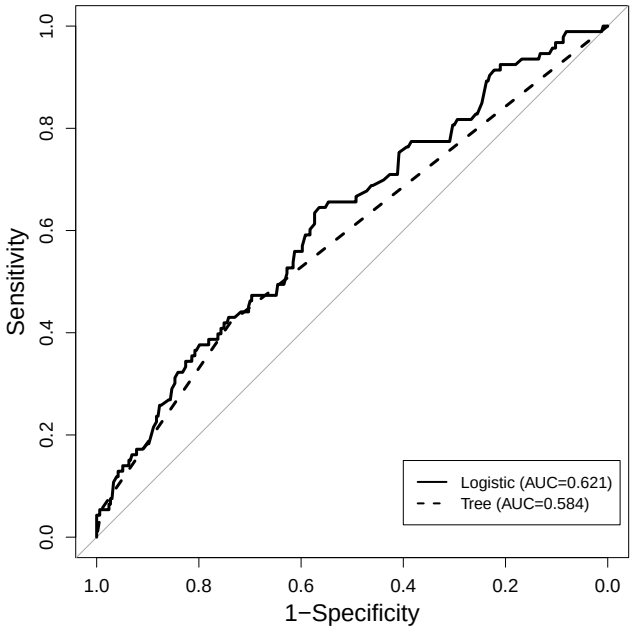

Figure 4.1 Receiver Operating Characteristic Curve for HIV Data

4.5 Discussions and Conclusions

In this work, we propose a new method for estimating the restricted optimal treatment regime for competing risks data, which minimizes the cumulative incidence of the primary risk at a fixed time point while controlling the corresponding cumulative incidence of the second risk under a pre-specified level. A penalization method is developed for obtaining an approximate solution for the challenging restricted optimization problem.

The proposed method is applied to the HSSC HIV dataset to obtain a restricted optimal treatment regime that minimizes the t-year cumulative incidence function of the risk of treatment or virologic failures while controlling the t-year cumulative incidence of serious drug-induced side effects under a predetermined level.

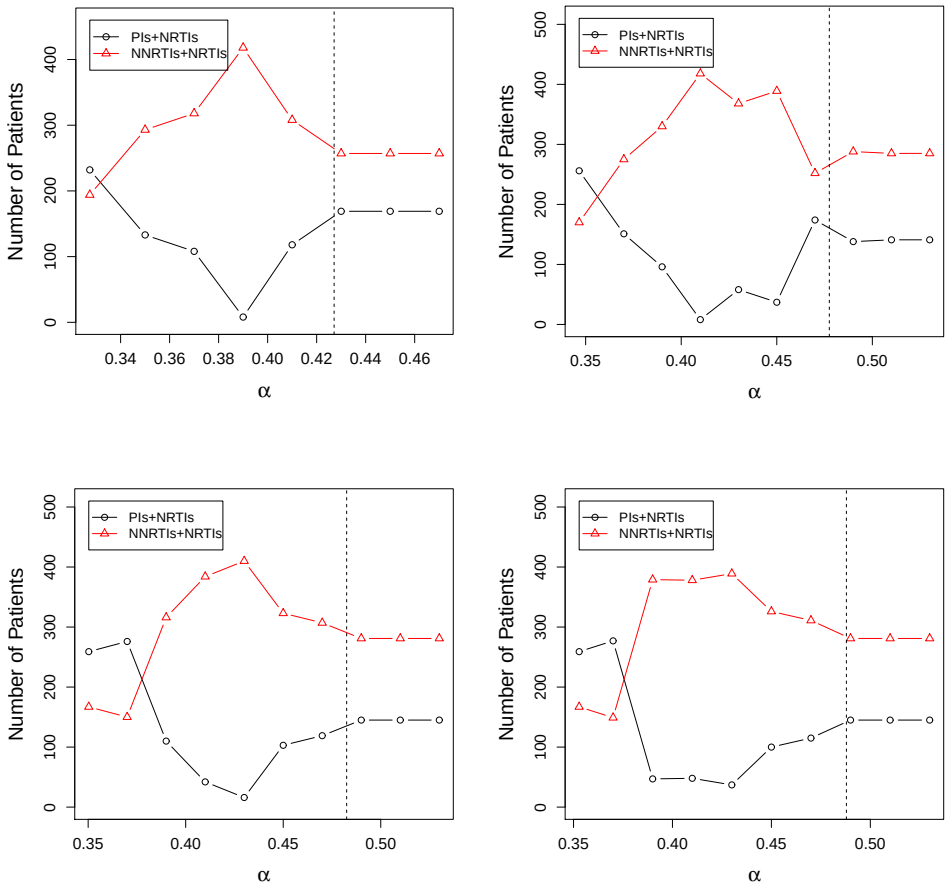

Figure 4.2 Treatment Distribution for HIV Data

Chapter 5

Summary and Future Studies

We completed three projects in this dissertation. The first two projects are based on the ACLS database, and the last project is based on the HSSC database. In the ACLS datasets, we study the longitudinal effect of fitness on CVD mortality through generalized odds rate joint models with varying-coefficients in Project 1. The estimated age-dependent varying coefficient curve for the longitudinal fitness in the survival model clearly represented the change of the fitness effect on CVD mortality over age. Aging is the most important factor with lots of chronic diseases. The age-related behavior change play an important role in disease development and the corresponding disease-related mortality. The proposed model can be broadly used in modeling survival outcomes with age-varying effect of longitudinal predictors, and helps improving the understanding of the real impact of some age-related chronic behaviors on the survival outcomes.

The transitions to the CVD and all-cause mortality states in the ACLS are studied through the Markov illness-death regression models in Project 2. Estimation methods are derived for the proposed model where CVD incidence time is considered as interval censored. Covariates' effect on the transitions among different states in the illness-death model can be evaluated through the regression coefficients in each transition intensity functions. Based on the estimated coefficients, we are able to know which factors are important for subjects to transit to the CVD and death states. Current research requires that the mortality reasons, especially the one related to the disease of

interest, are available. This additional information is available in the ACLS data, but not always obtainable in observational studies. In the case when such information is not available, the accuracy of estimation based on the proposed method also relies on the length of visit intervals. That is, if the average length between two visits is relatively small compared with the time to develop the disease of interest, we can also ignore the possibility that the patient dies with the disease when there's no previous diagnosed records. Future extensions of current research can focus on the semi-markovian or the non-markovian models, where the transition from disease to death depend on both the disease transition time and the duration of having disease.

In Project 3, we study the precision medicine problem under a competing risks framework based on the HSSC HIV dataset. We define a restricted optimal treatment regime that minimizes the t-year cumulative incidence function of the main risk while controlling the t-year cumulative incidence of the other risk under a predetermined level. In the proposed method, we only consider the inverse propensity score-weighted estimators for the regime-specific overall survival function and cumulative incidence functions. However, it can be extended to accommodate the augmented inverse propensity score-weighted estimators, similar to in Jiang et al. (2017). Such an extension requires modeling and estimation of the cumulative incidence functions of both risks simultaneously, for example, as in Lu and Peng (2008) and Mao and Lin (2017), which may be difficult to implement in practice. In addition, the proposed method can be extended to derive the restricted optimal dynamic treatment regime for multiple treatment decision time points. These warrant future research.

Bibliography

Andersen, P. K. and N. Keiding (2002). Multi-state models for event history analysis. *Statistical methods in medical research 11*(2), 91–115.

Andrinopoulou, E.-R., D. Rizopoulos, J. J. Takkenberg, and E. Lesaffre (2014). Joint modeling of two longitudinal outcomes and competing risk data. *Statistics in medicine 33*(18), 3167–3178.

Bai, X., A. A. Tsiatis, W. Lu, and R. Song (2017). Optimal treatment regimes for survival endpoints using a locally-efficient doubly-robust estimator from a classification perspective. *Lifetime data analysis 23*(4), 585–604.

Balke, B. and R. W. Ware (1959). An experimental study of physical fitness of air force personnel. *United States Armed Forces Medical Journal 10*(6), 675–688.

Barrett, J. K., F. Siannis, and V. T. Farewell (2011). A semi-competing risks model for data with interval-censoring and informative observation: An application to the mrc cognitive function and ageing study. *Statistics in medicine 30*(1), 1–10.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in medicine 2*(2), 273–277.

Blair, S. N., J. B. Kampert, H. W. Kohl, C. E. Barlow, C. A. Macera, R. S. Paffenbarger, and L. W. Gibbons (1996). Influences of cardiorespiratory fitness and other precursors on cardiovascular disease and all-cause mortality in men and women. *Jama 276*(3), 205–210.

Borges, Á. H., A. Lundh, B. Tendal, J. A. Bartlett, N. Clumeck, D. Costagliola, E. S. Daar, P. Echeverría, M. Gisslén, T. B. Huedo-Medina, et al. (2016). Nonnucleoside reverse-transcriptase inhibitor-vs ritonavir-boosted protease inhibitor–based regimens for initial treatment of hiv infection: A systematic review and metaanalysis of randomized trials. *Clinical Infectious Diseases 63*(2), 268–280.

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.

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.

Cai, Z. and Y. Sun (2003). Local linear estimation for time-dependent coefficients in cox's regression models. *Scandinavian Journal of Statistics 30*(1), 93–111.

Dabrowska, D. M. and K. A. Doksum (1988). Estimation and testing in a two-sample generalized odds-rate model. *Journal of the American Statistical Association 83*(403), 744–749.

Datta, S., G. A. Satten, and S. Datta (2000). Nonparametric estimation for the three-stage irreversible illness–death model. *Biometrics 56*(3), 841–847.

David, C. R. et al. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society 34*, 187–220.

de Uña-Álvarez, J. and L. Meira-Machado (2015). Nonparametric estimation of transition probabilities in the non-markov illness-death model: A comparative study. *Biometrics 71*(2), 364–375.

Dybul, M., A. S. Fauci, J. G. Bartlett, J. E. Kaplan, and A. K. Pau (2002). Guidelines for using antiretroviral agents among hiv-infected adults and adolescents. recommendations of the panel on clinical practices for treatment of hiv. *Annals of Internal Medicine 137*(5), 381–433.

Elashoff, R. M., G. Li, and N. Li (2008). A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics 64*(3), 762–771.

Fine, J. P. and R. J. Gray (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association 94*(446), 496–509.

Fine, J. P., H. Jiang, and R. Chappell (2001). On semi-competing risks data. *Biometrika 88*(4), 907–919.

Fisher, L. D. and D. Y. Lin (1999). Time-dependent covariates in the cox proportional-hazards regression model. *Annual review of public health 20*(1), 145–157.

Frydman, H. (1992). A nonparametric estimation procedure for a periodically observed three-state markov process, with application to aids. *Journal of the Royal Statistical Society. Series B (Methodological)*, 853–866.

Frydman, H. (1995a). Nonparametric estimation of a markov Śillness-deathŠprocess from interval-censored observations, with application to diabetes survival data. *Biometrika 82*(4), 773–789.

Frydman, H. (1995b). Semiparametric estimation in a three-state duration-dependent markov model from interval-censored observations with application to aids data. *Biometrics*, 502–511.

Frydman, H. and M. Szarek (2009). Nonparametric estimation in a markov Şillness–deathT̃ process from interval censored observations with missing intermediate transition status. *Biometrics 65*(1), 143–151.

Gray, R. J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of statistics 16*(3), 1141–1154.

Günthard, H. F., M. S. Saag, C. A. Benson, C. Del Rio, J. J. Eron, J. E. Gallant, J. F. Hoy, M. J. Mugavero, P. E. Sax, M. A. Thompson, et al. (2016). Antiretroviral drugs for treatment and prevention of hiv infection in adults: 2016 recommendations of the international antiviral society–usa panel. *Jama 316*(2), 191–210.

Haubrich, R. H., S. A. Riddler, A. G. DiRienzo, L. Komarow, W. G. Powderly, K. Klingman, K. W. Garren, D. L. Butcher, J. F. Rooney, D. W. Haas, et al. (2009).

Metabolic outcomes in a randomized trial of nucleoside, nonnucleoside and protease inhibitor-sparing regimens for initial hiv treatment. *AIDS (London, England) 23*(9), 1109.

Hsieh, J.-J. and Y.-T. Huang (2012). Regression analysis based on conditional likelihood approach under semi-competing risks data. *Lifetime data analysis 18*(3), 302–320.

Huang, X., G. Li, R. M. Elashoff, and J. Pan (2011). A general joint model for longitudinal measurements and competing risks survival data with heterogeneous random effects. *Lifetime data analysis 17*(1), 80–100.

Jiang, R., W. Lu, R. Song, and M. Davidian (2017). On estimation of optimal treatment regimes for maximizing t-year survival probability. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(4), 1165–1185.

Jiang, R., W. Lu, R. Song, M. Hudgens, and S. Naprvavnik (2017). Doubly robust estimation of optimal treatment regimes for survival data - with application to an hiv/aids study. *The Annals of Applied Statistics 11*(3), 1763–1786.

Joly, P., D. Commenges, C. Helmer, and L. Letenneur (2002). A penalized likelihood approach for an illness–death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics 3*(3), 433–443.

Kay, R. (1986). A markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, 855–865.

Klein, J. P. and P. K. Andersen (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics 61*(1), 223–229.

Klein, J. P., J. H. Klotz, and M. R. Grever (1984). A biological marker model for predicting disease transitions. *Biometrics*, 927–936.

Kohl 3rd, H. (2001). Physical activity and cardiovascular disease: evidence for a dose response. *Medicine and science in sports and exercise 33*(6 Suppl), S472–83.

Li, R. and Y. Cheng (2016). Flexible association modelling and prediction with semi-competing risks data. *Canadian Journal of Statistics 44*(3), 361–374.

Lu, W. and L. Peng (2008). Semiparametric analysis of mixture regression models with competing risks data. *Lifetime data analysis 14*(3), 231–252.

Mao, L. and D. Lin (2017). Efficient estimation of semiparametric transformation models for the cumulative incidence of competing risks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(2), 573–587.

Meira-Machado, L., J. de Uña-Álvarez, and C. Cadarso-Suárez (2006). Nonparametric estimation of transition probabilities in a non-markov illness–death model. *Lifetime Data Analysis 12*(3), 325–344.

Meira-Machado, L., J. de Uña-Álvarez, C. Cadarso-Suárez, and P. K. Andersen (2009). Multi-state models for the analysis of time-to-event data. *Statistical methods in medical research 18*(2), 195–222.

Mora, S., N. Cook, J. E. Buring, P. M. Ridker, and I.-M. Lee (2007). Physical activity and reduced risk of cardiovascular events. *Circulation 116*(19), 2110–2118.

Moreno-Betancur, M., J. B. Carlin, S. L. Brilleman, S. K. Tanamas, A. Peeters, and R. Wolfe (2017). Survival analysis with time-dependent covariates subject to missing data or measurement error: Multiple imputation for joint modeling (mijm). *Biostatistics*, kxx046.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(2), 331–355.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in medicine 24*(10), 1455–1481.

Nocon, M., T. Hiemann, F. Müller-Riemenschneider, F. Thalau, S. Roll, and S. N. Willich (2008). Association of physical activity with all-cause and cardiovascular mortality: a systematic review and meta-analysis. *European Journal of Cardiovascular Prevention & Rehabilitation 15*(3), 239–246.

of Sports Medicine, A. C. (2013). *ACSM's guidelines for exercise testing and prescription*. Lippincott Williams & Wilkins.

Organization, W. H. (2016). *Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach*. World Health Organization.

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.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*(5), 688–701.

Scharfstein, D. O., A. A. Tsiatis, and P. B. Gilbert (1998). Semiparametric efficient estimation in the generalized odds-rate class of regression models for right-censored time-to-event data. *Lifetime data analysis 4*(4), 355–391.

Siannis, F., V. Farewell, and J. Head (2007). A multi-state model for joint modelling of terminal and non-terminal events with application to whitehall ii. *Statistics in medicine 26*(2), 426–442.

Simpson, K. N., K. A. Hanson, G. Harding, S. Haider, M. Tawadrous, A. Khachatryan, C. L. Pashos, and A. W. Wu (2014). Review of the impact of nnrti-based hiv treatment regimens on patient-reported disease burden. *AIDS care 26*(4), 466–475.

Smith, K. Y., P. Patel, D. Fine, N. Bellos, L. Sloan, P. Lackey, P. N. Kumar, D. H. Sutherland-Phillips, C. Vavro, L. Yau, et al. (2009). Randomized, double-blind, placebo-matched, multicenter trial of abacavir/lamivudine or tenofovir/emtricitabine with lopinavir/ritonavir for initial hiv treatment. *Aids 23*(12), 1547–1556.

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.

Staszewski, S., J. Morales-Ramirez, K. T. Tashima, A. Rachlis, D. Skiest, J. Stanford, R. Stryker, P. Johnson, D. F. Labriola, D. Farina, et al. (1999). Efavirenz plus zidovudine and lamivudine, efavirenz plus indinavir, and indinavir plus zidovudine and lamivudine in the treatment of hiv-1 infection in adults. *New England Journal of Medicine 341*(25), 1865–1873.

Sun, L., J. Liu, J. Sun, and M.-J. Zhang (2006). Modeling the subdistribution of a competing risk. *Statistica Sinica 16*(4), 1367–1385.

Tian, L., D. Zucker, and L. Wei (2005). On the cox model with time-varying regression coefficients. *Journal of the American statistical Association 100*(469), 172–183.

Tsiatis, A. A. and M. Davidian (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica 14*(3), 809–834.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 290–295.

Watkins, C. J. and P. Dayan (1992). Q-learning. *Machine learning 8*(3-4), 279–292.

Worm, S. W., C. Sabin, R. Weber, P. Reiss, W. El-Sadr, F. Dabis, S. De Wit, M. Law, A. D. Monforte, N. Friis-Møller, et al. (2010). Risk of myocardial infarction in patients with hiv infection exposed to specific individual antiretroviral drugs from the 3 major drug classes: the data collection on adverse events of anti-hiv drugs (d: A: D) study. *The Journal of infectious diseases 201*(3), 318–330.

Wulfsohn, M. S. and A. A. Tsiatis (1997). A joint model for survival and longitudinal data measured with error. *Biometrics 53*(1), 330–339.

Xu, J., J. D. Kalbfleisch, and B. Tai (2010). Statistical analysis of illness–death processes and semicompeting risks data. *Biometrics 66*(3), 716–725.

Yu, M. (2016). Improving estimation efficiency for semi-competing risks data with partially observed terminal event. *Journal of Nonparametric Statistics 28*(4), 860–874.

Yu, M., N. J. Law, J. M. Taylor, and H. M. Sandler (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica 14*(3), 835–862.

Zeng, D. and J. Cai (2005). Simultaneous modelling of survival and longitudinal data with an application to repeated quality of life measures. *Lifetime Data Analysis 11*(2), 151–174.

Zeng, D., J. Cai, et al. (2005). Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *The Annals of Statistics 33*(5), 2132–2163.

Zeng, D. and D. Lin (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika 93*(3), 627–640.

Zhang, B., A. A. Tsiatis, E. B. Laber, and M. Davidian (2012). A robust method for estimating optimal treatment regimes. *Biometrics 68*(4), 1010–1018.

Zhang, B., A. A. Tsiatis, E. B. Laber, and M. Davidian (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika 100*(3), 681–694.

Zhao, Y., D. Zeng, E. B. Laber, and M. R. Kosorok (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association 110*(510), 583–598.

Zhao, Y., D. Zeng, E. B. Laber, R. Song, and M. R. Kosorok (2015). Doubly robust learning for estimating individualized treatment with censored data. *Biometrika 102*(1), 151–168.

Zhao, Y., D. Zeng, A. J. Rush, and M. R. Kosorok (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association 107*(499), 1106–1118.

Zhou, J., J. Zhang, and W. Lu (2017). An expectation maximization algorithm for fitting the generalized odds-rate model to interval censored data. *Statistics in medicine 36*(7), 1157–1171.

Zhou, R., H. Zhu, M. Bondy, and J. Ning (2017). Analyzing semi-competing risks data with missing cause of informative terminal event. *Statistics in medicine 36*(5), 738–753.