JOINT MODELING OF BIVARIATE TIME TO EVENT DATA WITH SEMI-COMPETING RISK

Ran Liao

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the Department of Biostatistics,

Indiana University

February, 2017

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

Sujuan Gao, Ph.D., Chair

Barry Katz, Ph.D.

Doctoral Committee

Ying Zhang, Ph.D.

September 8, 2016

Shanshan Li, M.D.

Jianjun Zhang, Ph.D.

DEDICATION

To My Beloved Family.

ACKNOWLEDGMENTS

While a completed dissertation bears the single name of the student, the process that leads to its completion is always accomplished in combination with the dedicated work of other people. I owe my gratitude to all those people who made this dissertation possible and because of whom my doctoral experience has been one that I will cherish forever.

First and foremost I want to thank my advisor and committee chair, Professor Sujuan Gao, who continually and convincingly conveyed a spirit of adventure in regard to research and scholarship. Without her guidance and persistent help, this dissertation would not have been possible. It has been an honor to be her Ph.D. student. She has taught me, both intentionally and unintentionally, how good a biostatistician and a researcher could be. I appreciate all her contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating. The joy and enthusiasm she has for her research and projects was appealing and motivational for me, even during tough times in the Ph.D. pursuit. I am also thankful for the excellent example she has provided as a successful female biostatistician and researcher. Her guidance will accompany me along my life, especially my journey in biostatistics, and I will continue to benefit from it.

I would like to gratefully and sincerely thank my committee members: Dr.Barry Katz, Dr. Ying Zhang, Dr.Shanshan Li and Dr. Jianjun Zhang. First, I would like to give my special thanks to Dr. Barry Katz, Chair of Biostatistics Department and Dr. Ying Zhang, Director of Biostatistics Program, for their constant effort in improving our program and providing so many wonderful opportunities for students

and junior researchers. Additionally, I want to thank Dr. Katz for his instruction and advice during classes and discussion. He showed me the way to make significant contributions to society by conducting biostatistics research that is critical in fighting disease. Thanks to Dr. Ying Zhang for his extensive knowledge in survival analysis and counting process; our discussion regarding likelihood, frailty model, and twostage model gave me a clear picture of methodologies in these areas. Thanks to Dr. Shanshan Li for her creative and cutting edge knowledge in joint modeling of survival data and longitudinal data analysis, and also in dynamic prediction. It has been great opportunity for me to learn these methodologies when I sit in the LiveAD project meeting. Thanks to Dr. Jianjun Zhang who is also my minor advisor; his instructions and knowledge in cancer epidemiology and nutrition epidemiology inspired me and also provoked my interest in oncology, which led me to my future career direction.

It has been wonderful experience to study and work in the Department of Biostatistics at Indiana University. I would like to express my sincere thanks to all the faculty members, biostatisticians and staff in our department, for their input, expertise, valuable discussions, patience, and accessibility. In particular, I would like to thank our department for offering me the chance to collaborate with the Chinese Center of Disease Control and the Regenstrief Institute, which has been a truly precious opportunity for me, where I developed collaboration skills and learned how to apply my knowledge to real world problems. Also I would like to thank Dr. Zhangsheng Yu and Dr.Wanzhu Tu for their assistance and insightful comments to my research work in frailty model. I learned a lot from their previous work and through our research discussions. I also feel thankful for daily help from all the current

and former students in the biostatistics program as well as the students outside the program.

I would like to extend my thanks to Dr.Tianle Hu at Eli Lilly for his support to my "bivariate time to event data association" research. Dr.Hu offered me his insightful knowledge and expertise in this area. I am so honored to work with Dr.Hu and have him as coauthor for some of my research work. And Dr.Hu also enlightened me with the latest progress and methodologies in pharmaceutical industrial and clinical trials, which has been valuable information to direct me in my career path.

I wish to express my thanks to my family and friends. Thanks to my parents and grandparents who raised me up as a positive and energetic person and gave me enough support to let me explore the world. Thanks to my mom Rong Zhang for her unconditional love and endless support. Thanks to my parents in law and grandparents in law for their understanding and encouragement. Thanks to my friends who can accompany me when I am so frustrated with my life and family.

Last, thanks to my dearest husband: Jie Xue. Jie and I are working on our doctorate degree at same time. We went through this journey together; we made each other much braver and stronger in facing challenges and obstacles in life. In the meantime, we are much humbler and more prepared for the uncertainty of the future. And to my daughter, Qixuan Xue (Emma), and possible future kids, let mom share this statistician's secrete with you: If you determined to get some significance, but it just doesn't work out. Don't give up. Try harder and try again, play smarter and play again, you will be pleased by what the persistence and the possibility bring to you during the procedure, and also in the end.

Ran Liao

JOINT MODELING OF BIVARIATE TIME TO EVENT DATA WITH SEMI-COMPETING RISK

Survival analysis often encounters the situations of correlated multiple events including the same type of event observed from siblings or multiple events experienced by the same individual. In this dissertation, we focus on the joint modeling of bivariate time to event data with the estimation of the association parameters and also in the situation of a semi-competing risk.

This dissertation contains three related topics on bivariate time to event models. The first topic is on estimating the cross ratio which is an association parameter between bivariate survival functions. One advantage of using cross-ratio as a dependence measure is that it has an attractive hazard ratio interpretation by comparing two groups of interest. We compare the parametric, a two-stage semiparametric and a nonparametric approaches in simulation studies to evaluate the estimation performance among the three estimation approaches.

The second part is on semiparametric models of univariate time to event with a semi-competing risk. The third part is on semiparametric models of bivariate time to event with semi-competing risks. A frailty-based model framework was used to accommodate potential correlations among the multiple event times. We propose two estimation approaches. The first approach is a two stage semiparametric method where cumulative baseline hazards were estimated by nonparametric methods first and used in the likelihood function. The second approach is a penalized partial

likelihood approach. Simulation studies were conducted to compare the estimation accuracy between the proposed approaches. Data from an elderly cohort were used to examine factors associated with times to multiple diseases and considering death as a semi-competing risk.

Sujuan Gao, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Overview

This dissertation is devoted to develop new methodologies in survival analysis of joint models of bivariate time to events data with a semi-competing risk. This research work is primarily motivated by some interesting problem emerging from observational study of chronic diseases in aging cohort, in which a better understanding of chronic diseases natural history is needed to better understand and identify risk factor, to learn diseases relations, to design better health care and intervention for optimal treatment.The data support through out this dissertation come from electronic medical records (EMRs) in a longitudinal cohort of elderly African Americans enrolled in the Indianapolis-Ibadan Dementia Project (IIDP)(Hendrie et al., 2001).

Multivariate survival data arises when one encounters the situation of correlated multiple types of events including the same type of events observed from siblings or multiple events experienced by the same individual. A naive approach analyzing these survival data separately for each survival outcome by ignoring the association among the multiple events may produce biased results. Furthermore, investigating on how these multiple events relate to each other may offer important information on the underlying mechanisms for these events. In this dissertation, we focus on the joint modeling of bivariate time to event data with the estimation of the association parameters and also in the situation of a semi-competing risk.

By studying and reviewing some current and well established techniques in survival analysis realm, the proportional hazards model, proposed by Cox (1972), is certainly one of the most widely, used and studied regression model for time-to-event data, the cox model focus on exploring the relationship between the baseline hazard and treatment effect with the adjustment for the other explanatory variables. Extend from cox model, in order to take into account the heterogeneity due to the unobserved risk factor, Clayton (1978) and Vaupel et al. (1979) proposed to use frailty model or mixed proportional hazards model. The frailty term which can be understand as random effects, if these effects are subject-specific and unobserved heterogeneity stands for overdispersion and the model is called univariate frailty model (Wienke, 2010). In the case when random effects are shared by groups of subjects, a clustering effect is there, i.e. observations belonging to the same group are dependent. This is the case of shared frailty models (Hougaard, 2012).

The integration of frailty and multi-event models can provide powerful survival models to study the risk of many interrelated events while accounting for dependence among multiple events. Many practical situations can be thought of in which such integration is of interest. The main problem motivating our research arises from observational study of elder cohort, in which the participant usually was facing multiple diseases due to aging. Thus, we need to take into account the dependence between events when we conduct analysis and make inference .

The dissertation contains three related topics on bivariate time to event models. The first topic is on estimating the association parameter between bivariate survival functions. The second is on semiparametric pseudo-likelihood and semiparametric penalized partial likelihood models of univariate time to event with a

semi-competing risk. The third part is on semiparametric models of bivariate times to event with a semi-competing risk.

1.2 Covariate Dependent Cross Ratio of Bivariate Survival Times

Most current methods used in estimating the association parameter in bivariate time to event data have used either the copula approach or a frailty approach where the association parameter is treated as a constant parameter or as a nuisance. Cross-ratio is an association parameter which measures the dependence structure between two correlated failure times. One advantage of using cross-ratio as a dependence measure is that it has an attractive hazard ratio interpretation by comparing two groups of interest. In shared frailty models for bivariate survival data the frailty is identifiable through the cross ratio function, which provides a convenient measure of association for correlated survival variables. The cross ratio function may be used to compare patterns of dependence across models and data sets.

To estimate the cross ratio as a function, Nan et al. (2006) partitioned the sample space of the bivaratiate survival time into rectangular regions with edges parallel to the time axes and assumed that the cross ratio is constant in each retangular region. Shih and Louis (1995) and Shih and Albert (2010) proposed a two stage semiparametric likelihood based method to estimate constant cross ratio and piecewise cross ratio under competing risk setup, respectively. In the context of competing risks and nonparametric appraoches,Cheng and Fine (2008), Bandeen-Roche and Ning (2008)and Ning and Bandeen-Roche (2014) proposed a nonparametric method for estimating the piecewise constant time-varying cause-specific cross ratio using the binned survival data based on the same partitioning idea for the sample space,

counted the concurrence events pair and discordinate events pair can formed up a logistics-form of regression procedure for the estimation procedure. Recent years, Li and Lin (2012), Othus and Li (2010) and Hsu and Moodie (2007) characterized the dependence of bivariate survival data through the correlation coefficient of normally transformed bivariate survival times. Such methods, however, require assumptions of specific copula models for the joint survival function, for which appropriate model checking techniques are lacking.

Hu et al. (2011) proposed an estimation approach for time dependent cross ratio using a pseudo-partial likelihood approach. Build on Hu et al. (2011)'s methodology, we propose a cross ratio set-up which allows the modeling of covariate effects on the association parameter. The advantage of such a model is that covariate effect is linked with cross ratio explicitly. In addition, the non-parametric estimation approach does not require the specification of either the joint or the marginal survival functions and thus is robust against model misspecification. A simulation study is conducted to evaluate the estimation performance of this nonparametric estimation approaches. The proposed estimation approach is used to estimate gender effects on the association between time to coronary artery disease (CAD) and time to depression using data from an elderly cohort.

1.3 Frailty based Semiparametric Models for Time to Event Data with a Semi-competing Risk

Semi-competing risk often arises in biomedical research, in particular, in studies of aging when individuals at risk of a particular disease die from other causes. As the two-types of events are usually correlated, models for semi-competing risks should

properly take account of the dependence. In the literature, copula models are popular approaches for modeling of such data. However, the copula model postulates latent failure times and marginal distributions for the non-terminal event that may not be easily interpretable in reality. Further, the development of regression models is complicated for copula models. To overcome these issues, the well-known illness-death models have been recently proposed for more flexible modeling of semi-competing risks data.

In the second part of this dissertation, we proposed a frailty model approach for a survival outcome with a semi-competing risk. The standard likelihood based approach for multivariate lognormal frailty models involves multi-dimensional integrals over the distribution of the multivariate frailties, which almost always do not have analytical solutions. Numerical solutions such as Gaussian quadrature rules, Monte Carlo sampling have been routinely used in literature.However, as the dimension increases, these approaches still remain computationally demanding.

In order to retain the nice interpretation of frailty model and overcome the computational challenge, two estimation approaches are proposed and compared. The first is a two-stage pseudo-likelihood approach where cumulative baseline hazards were first estimated by a nonparametric method. Parameter estimation is then achieved by maximizing the pseudo-likelihood functions where the estimated cumulative hazards from stage one were used. In the second approach, we propose a penalized partial likelihood function for parameter estimation and inference similarly to the concept used in the Cox's partial likelihood. An estimation procedure based on penalized pseudo-partial likelihood is used for estimating covariate effects. The penalized partial likelihood is obtained by Laplace approximation to the true likelihood. Penalized

Cox PH model discussed by Gray (1992); Perperoglou (2014) provided methods of parameter estimation. A simulation study is conducted to compare the estimation performance of these two approaches. The proposed estimation approach is used to estimate gender effects on the time to coronary artery disease (CAD) with time to death as a semi-competing risk.

1.4 Frailty-based Multi-event Semiparametric Models for Failure Time Data with Semi-competing Risks

The third topic of this dissertation extends the models considered in the second part to bivariate survival outcomes with a semi-competing risk.

In medical research, multi-event and multi-stage data arises when a individual was at risk of multiple disease, or a certain disease progressed in several states. It was crucial to study the inner structure and dependence between multiple diseases or multiple states.

In this part, we propose to use frailty based semparametric model introducing random effects to account for unobserved risk factors, possibly shared by multiple diseases or multiple states. For model estimation, we developed and evaluated parametric, two stage semiparametric estimation and penalized partial likelihood approach.The two stage pseudo-likelihood approach and the penalized pseudo-partial likelihood approach are also be used and compared in simulation studies.

In many epidemiological studies of the elderly population, it has been observed that individuals at risk of one chronic condition tend to have increased risk of other medical conditions with a substantial numbers having multiple chronic conditions. Studying the co-occurrence of these conditions may identify common biological

pathways linking these disorders and ultimately lead to effective treatment and prevention strategies. Another complication facing the studies in aging is death due to other causes which can be indirectly related to the conditions under study through genetic or environmental exposures related to the individual's susceptibility to both disease and death. The proposed approaches were applied to data from the elderly cohort to determine risk factors associated with CAD (Coronary Artery Disease) and depression with death as the semi-competing risk.

1.5 Main Contribution and Structure of Dissertation

The work presented in this thesis contributes to research in survival analysis in following areas: modeling methodology and data applications, and simulation techniques.

The main contribution to modeling methodology consists of proposing covariate dependent cross ratio estimation methods, and frailty based semiparametric model in the presence of semi-competing risk data. Up to now, there existed no covariate dependent association measure for bivariate time to event data. Capturing the dependent structure between multiple survival events is a challenging topic. Our proposed cross ratio model offered a feasible approach to measure the dependence between bivariate survival times. In addition, we propose to use frailty based model approach to handle semi-competing risk data and multiple event, which captures the transition between multiple events and also account for informative censoring caused by the other event and death.

Two estimation approaches have been developed and investigated in the dissertation: a parametric and a semiparametric approach. First, fully parametric inference, based on maximum marginal likelihood, is considered. Then, a semiparametric

estimation approach, based on maximum penalized partial likelihood, is proposed and investigated(Rotolo and Legrand, 2012; Rotolo et al., 2013).

Another contribution of our work is that we developed a general method in multi-event research for simulating data according to a given scenario. Dependence can be added between time variables of grouped subjects, to study the effect of clustering. Moreover, the simulation method is able to introduce, using copulas, dependence between times of different transitions while fixing the marginal distributions according to a given scenario. This is a useful tool to study, for instance, the robustness of (frailty) multi-event models .

The structure of this dissertation is as follows. In Chapter 2, we focus on the estimation of dependent association parameter: cross ratio between bivariate survival times. In Chapter 3, we present two estimation approaches for semi-competing risk models. In Chapter 4, we extent the semi-competing risk model presented in chapter 3 to bivariate time to event data. Chapter 5 gives concluding marks.

Chapter 2

Covariate Dependent Cross Ratio of Bivariate Survival Times

2.1 Abstract

Cross ratio is formed as the ratio of two conditional hazard rates for one events given the other event. Inherited the nice interpretation of hazard ratio from survival analysis setup, cross ratio can be interpreted as hazard ratio of one event conditional the status of the other event. It is very meaningful to investigate the covariate effect on the cross ratio, which can be a useful tool to explain contribution of the certain component to the dependent between two time to events. In this paper, first, we extended two methodologies in constant cross ratio estimation into covariate adjust cross ratio with multiplicative covariate effect set up, which are Clayton copula model and Shih and Louis(1995) two stage semiparametric model. Then, we conducted a simulation study comparing Hu et al. (2011)'s non-parametric estimator with parametric estimator from copula approach and semi-parametric estimator from Shih and Louis (1995)'s two stage approach. In the mean time, we presented a comprehensive review and discussion of these three methodologies. To illustrate three estimation methodologies, we analyzed data from Indianapolis-Ibadan Dementia Project (IIDP) to investigate the gender effect between cardiovascular event and depression.

2.2 Introduction

Bivariate survival outcomes are often collected in medical studies. In many cases, the two failure times may be correlated. Earlier interests have focused on determining the correlations between disease occurrence times of family members in genetic epidemiology such as the age of onsets to asthma or type I diabetes in twin studies (Hyttinen et al., 2003; Thomsen et al., 2011). However, there is also an increasing interest in examining times to two related diseases observed from the same individual in order to identify common pathways and potential risk factors underlying both diseases. For instance, there have been considerable research efforts focusing on the link between coronary artery disease (CAD) and depression. CAD and depression are both common in late life and have been shown to be associated with increased risk of disability and mortality(Callahan et al., 1998). A "vascular depression hypothesis" was first proposed by Alexopoulos et al. (1997) when the authors proposed that cardiovascular disease may predispose, precipitate, or perpetuate some geriatric depressive syndromes. However, the vascular depression hypothesis was recently replaced by a new model describing the association between CAD and depression as the outcome of "two intertwined, mutually reinforcing disorders"(de Jonge and Roest, 2012).Evidence supporting this new bi-directional model between CAD and depression includes the increased risk of CAD in people suffering from depression and that late-life depression has been found to be associated with neuroimaging findings for subclinical cerebrovascular disease(de Groot et al., 2000; Hawkins et al., 2014). A gender difference in CAD and comorbid depression has been observed prompting a search for a common immunological basis including the role of inflammation in both

diseases (Möller-Leimkühler, 2010; Wright et al., 2014). Therefore, analysis of bivariate survival outcomes includes estimating the dependence between the two times to events and determining the contributions from common risk factors as the two primary objectives.

The dependence between two survival times has been discussed previously in the literature (Diva et al., 2008; Li and Lin, 2012; Li et al., 2008; Rondeau et al., 2012) . One naive approach is to use global rank measures such as Kendall's τ and Spearman's coefficient ρ(Hougaard, 2012; Kendall, 1948). However, two major issues were not addressed using these estimators: first, both estimators cannot incorporate censoring information leading to potentially biased and inefficient estimates; second, both estimators do not account for covariate contribution to the association of the two event times.

In contrast to the global rank based association measures, cross ratios, formulated as the ratio of two conditional hazard functions, offer a direct measure of dependence between two survival times that can account for censoring and accommodate potential covariates(Kalbfleisch and Prentice, 2002). There are three broad classes of estimation approaches for cross ratio estimation.

The first is a full parametric approach. Clayton (1978) introduced the Clayton copula model as an explicit closed-form bivariate survival function model with a constant cross ratio. Oakes (1982) demonstrated that the Clayton copula model can be derived using a frailty framework, where a common latent variable induces a correlation between events. A parametric approach will require the specification of a bivariate survival model, such as the Clayton model, and the simultaneous estimation of the marginal survival functions and the cross ratio parameter. The second

is a semi-parametric approach developed by Shih and Louis (1995). The marginal survival functions were first estimated by Kaplan-Meier estimators and used in the bivariate survival function to derive the cross ratio estimate. Shih and Louis (1995) showed that the two stage semi-parametric approach is efficient when the marginal survival functions were unknown. Lawless and Yilmaz (2011) compared a one stage semiparametric maximum likelihood (ML) approach and a two stage semi-parametric pseudo maximum likelihood (PML) approach for the Clayton model and Frank copula. In the one stage semi-parametric ML approach, the marginal functions, and the association parameter were estimated using non-parametric methods simultaneously. Lawless and Yilmaz (2011) concluded that that the two-stage semi-parametric PML was the preferable approach for marginal distribution estimation in most situations that do not involve covariates. When covariates were presented in the marginal distributions, however, the one stage ML method can be substantially better in some settings. When the bivariate survival model is misspecified, Lawless and Yilmaz (2011) showed that the two stage PML can perform worse than the one stage ML for cross ratio estimates. They also pointed out that one stage semiparametric approach was more computationally intensive compare to two stage method.

Both the parametric and the semi-parametric approaches assume a constant cross ratio.Nan et al. (2006) considered a piece-constant cross ratio set up by partitioning the sample space of bivariate survival function into rectangles each of which was assumed to have a constant cross ratio. Hu et al. (2011) proposed a nonparametric estimation approach which allowed cross ratio to be modeled as a time varying function. For estimation, Hu et al. (2011) constructed an objective function by mimicking the partial likelihood in the David (1972) propertional hazard model.

No previously published studies have considered modeling covariate effect in the cross ratio. Given the interpretation of cross ratio as a conditional hazard ratio for one event given the other event, It will be interesting to determine the effect of covariates on the cross ratio in order to account for the change in the association between two events. In this paper, we extend Clayton (1978) 's copula model and Shih and Louis (1995)'s two stage semiparametric model into covariate adjusted cross ratio setup with multiplicative covariate effect similar to Hu et al. (2011). We present a simulation study comparing Hu et al. (2011) 's non-parametric estimator with the parametric estimator from the copula approach and a two stage semi-parametric estimator from Shih and Louis (1995). The proposed method is illustrated using data from the Indianapolis-Ibadan Dementia Project (IIDP) to determine gender effect on the associating between time to coronary artery disease (CAD) and time to depression (Gao et al., 1998; Hendrie et al., 2001; Unverzagt et al., 2001).

In the following sections, we present the notations and model set up in Section 2. We describe estimation approaches in Section 3 and results from a simulation study in Section 4. We present results from the IIDP data analysis in Section 5 and conclude the article with a discussion in Section 6.

2.3 Notation, Definition and Model Setup

In this section, we introduce some common notation and definition in survival analysis and cross ratio analysis

Consider a pair of correlated continuous failure times (T_1, T_2) that are subject to right censoring by a pair of censoring times (C_1, C_2). Let (S_1, S_2) and (f_1, f_2) denote the corresponding marginal survival functions and density functions, respec-

tively. Let (h_1, h_2) and (H_1, H_2) denote the corresponding marginal hazard and cumulative hazard, respectively. We assume that censoring times are independent of failure times. Suppose we observe n independent and identically distributed vectors of $(X_1, X_2, \Delta_1, \Delta_2)$, where $X_1 = min(T_1, C_1), X_2 = min(T_2, C_2), \Delta_1 = I(T_1 \leq C_1)$ and $\Delta_2 = I(T_2 \leq C_2)$. Here $I(\cdot)$ denotes the indicator function. We further assume that there are no ties among the two observed times.

Cross ratio function of T_1 and T_2 at time (t_1, t_2) is defined as

$$\alpha(t_1, t_2) = \frac{h_2(t_2|T_1 = t_1)}{h_2(t_2|T_1 > t_1)} = \frac{h_1(t_1|T_2 = t_2)}{h_1(t_1|T_2 > t_2)} \tag{2.1}$$

The function can be interpreted as the ratio of the hazard rate of the conditional distribution of T_1, given T_2, to that of T_1. given $T_2 \geq t_2$.(Oakes, 1989) We have

$$\begin{aligned} h(t_1|T_2 = t_2) &= -\frac{\partial_1 S_1(t_1|T_2 = t_2)}{S_1(t_1|T_2 = t_2)} \\ &= -\frac{\partial_{1,2} S(t_1, t_2)}{\partial_2 S(t_1, t_2)} \end{aligned}$$

and

$$h(t_1|T_2 \geq t_2) = -\frac{\partial_1 S(t_1, t_2)}{S(t_1, t_2)}$$

Then,

$$\begin{aligned} \alpha(t_1, t_2) &= \frac{\partial_{1,2} S(t_1, t_2) \times S(t_1, t_2)}{\partial_1 S(t_1, t_2) \times \partial_2 S(t_1, t_2)} \\ &= \frac{f(t_1, t_2) \times S(t_1, t_2)}{\partial_1 S(t_1, t_2) \times \partial_2 S(t_1, t_2)} \end{aligned}$$

Where,

$$\partial_1 f(t_1, t_2) = \frac{\partial f(t_1, t_2)}{\partial t_1}$$
$$\partial_2 f(t_1, t_2) = \frac{\partial f(t_1, t_2)}{\partial t_2}$$
$$\partial_{1,2} f(t_1, t_2) = \frac{\partial f(t_1, t_2)}{\partial t_1 \partial t_2}$$

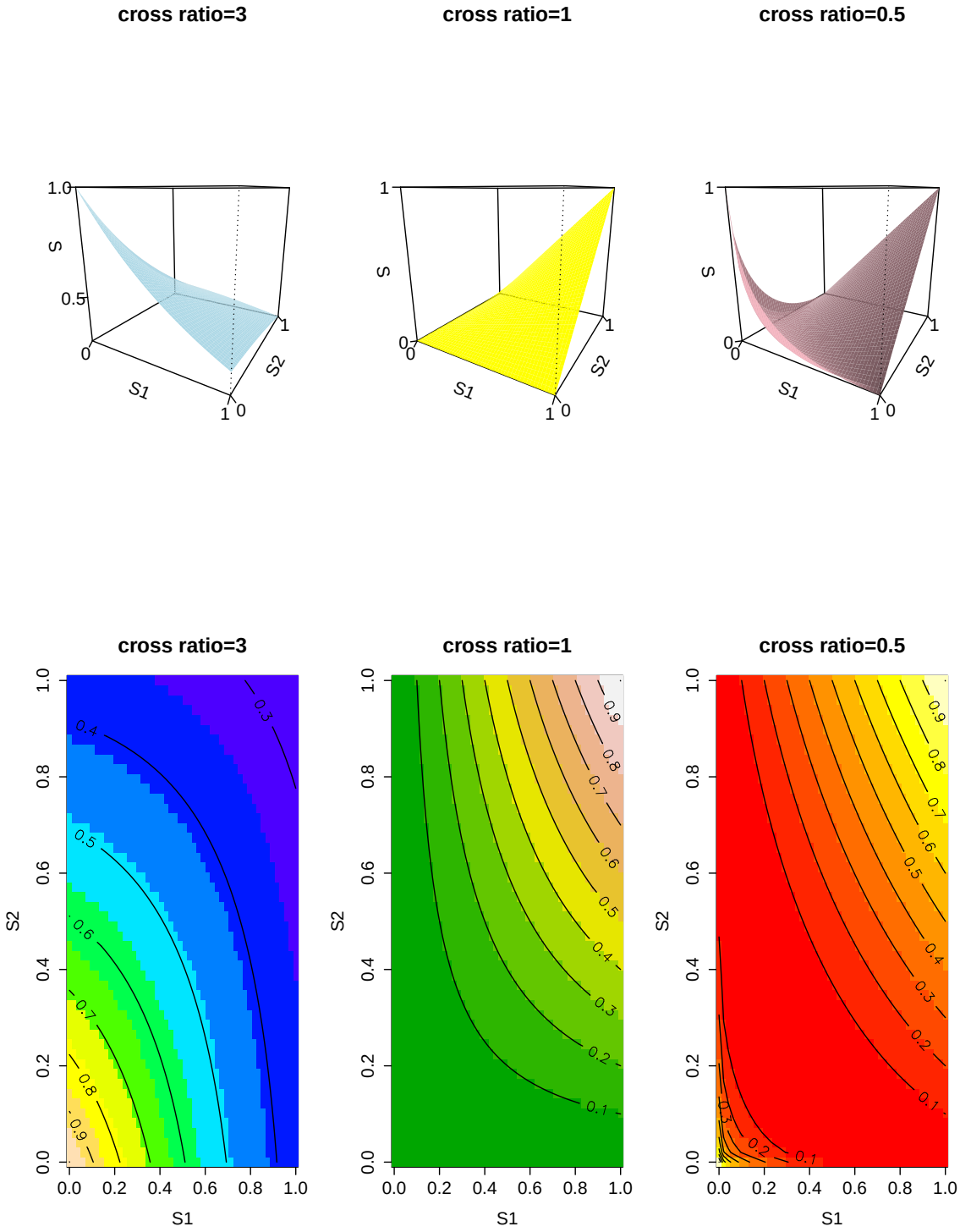
$f(t_1, t_2)$ is a function of t_1 and t_2.

When $\alpha(t_1, t_2) = 1$, the two events are independent; when $\alpha(t_1, t_2) > 1$, the two events are positively correlated; when $\alpha(t_1, t_2) < 1$, the two events are negatively correlated. Hu et al. (2011); Oakes (1982, 1986, 1989) Figure (2.1) demonstrate the joint survival distribution of bivariate survival model under different cross ratio using perspective 3D surface plot and 2D contour plot. When $crossratio = 3$, the two time to event were positively correlated, the joint survival will increase as two marginal survival increase, the contour plot (Figure (2.1) bottom left) showed concave feature. When $crossratio = 1$, the two events were independent, the joint survival is the direct product of two marginal survival $S = S_1(t) \cdot S_2(t)$. When $crossratio = 0.5$, the two events were negatively correlated, the joint survival showed twisted structure over the space.

Let W be a set of covariates. Cross ratio function conditional on covariates can be defined as:

$$\alpha(t_1, t_2; \mathbf{w}) = \frac{h_2(t_2|T_1 = t_1, W = \mathbf{w})}{h_2(t_2|T_1 > t_1, W = \mathbf{w})} = \frac{h_1(t_1|T_2 = t_2, W = \mathbf{w})}{h_1(t_1|T_2 > t_2, W = \mathbf{w})} \tag{2.2}$$

Figure 2.1: The Joint Distribution of Bivariate Survival Model Plot under Varying Cross Ratio: the first row is the 3-D plot, the height represents joint survival probability and the marginal of survival distribution of T_1 and T_2 are identical; the second row is the contour plot, the number on the black line represent the value of joint survival probability



Here, we further assume that the covariate W has a multiplicative effect on the cross ratio:

$$\alpha(\boldsymbol{\beta}; t_1, t_2) = \alpha(\boldsymbol{\beta}; t_1, t_2, \mathbf{w}) = \alpha_0(t_1, t_2) \cdot exp(\mathbf{w} \cdot \boldsymbol{\beta}) \tag{2.3}$$

where $\alpha_0(t_1, t_2)$ is the cross ratio for a reference value defined by $\mathbf{w}$, and $exp(\mathbf{w} \cdot \boldsymbol{\beta})$ is an exponential function of $\mathbf{w}$. For example, if $W = 0$ is to be used as a reference for the effect of W, then

$$\alpha_0(t_1, t_2) = \frac{h_2(t_2|T_1 = t_1, W = 0)}{h_2(t_2|T_1 > t_1, W = 0)} = \frac{h_1(t_1|T_2 = t_2, W = 0)}{h_1(t_1|T_2 > t_2, W = 0)} \tag{2.4}$$

Model (2.3) effectively separates the reference cross ratio function and the covariate effect thus providing an opportunity to model each piece separately.

2.4 Estimation Approaches

In this section, we describe three estimation approaches. The first two approaches are based on the parametric formation of Clayton copula. Thus, these two approaches can only accommodate discrete covariates in order to achieve constant cross ratio within each level of the covariate thus retaining the Clayton copula form. The third approach is nonparametric following the spirit of (Hu et al., 2011) where both discrete and continuous covariates can be handled(Hu, 2011).

2.4.1 Bivariate Clayton Copula Approach

The definition of cross ratio (2.1) is equivalent to the following second-order partial differential equation:

$$\frac{\partial^2 - \log(S(t_1,t_2))}{\partial t_1 \partial t_2} + (\alpha(\boldsymbol{\beta}; t_1, t_2) - 1)\frac{\partial - \log(S(t_1,t_2))}{\partial t_1}\frac{\partial - \log(S(t_1,t_2))}{\partial t_2} = 0 \quad (2.5)$$

where $S(t_1, t_2)$ is the joint survival function of (T_1, T_2) at (t_1, t_2).

When $\alpha(\boldsymbol{\beta}; t_1, t_2) = \alpha$ is constant, it can be shown that equation (2.5) has a unique solution of the form

$$C_\alpha(t_1,t_2) = [S_1(t_1)^{-(\alpha-1)} + S_2(t_2)^{-(\alpha-1)} - 1]^{-\frac{1}{\alpha-1}} \quad (2.6)$$

where $C_\alpha(t_1, t_2)$ is called Clayton copula (Clayton, 1978). The formation of Clayton copula is differed by the value of cross ratio α.

$$C_\alpha(t_1,t_2) = \begin{cases} [S_1(t_1)^{-(\alpha-1)} + S_2(t_2)^{-(\alpha-1)} - 1]^{-\frac{1}{\alpha-1}} & \alpha > 1 \\ S_1(t_1) \cdot S_2(t_2) & \alpha = 1 \\ \max([S_1(t_1)^{-(\alpha-1)} + S_2(t_2)^{-(\alpha-1)} - 1]^{-\frac{1}{\alpha-1}}, 0) & \alpha < 1 \end{cases}$$

where S_1 and S_2 are the marginal survival functions of T_1 and T_2.

Clayton Copula and Archimedean Family

In fact the Clayton copula belongs to an important family of copulas known as Archimedean copulas which have a simple form with a variety of dependence struc-

tures. The copula function $\mathcal{C}$ is generally define as multivariate function which can couples the joint survival function to its univariate margins in a manner completely analogous to the way in which a copula connects the joint distribution function to its margins. (Nelsen, 2007) The support of copula approach is supported by Sklar's canonical representation theorem.

Theorem 1 (*Sklar's Canonical Representation*) *Let $\mathcal{S}$ be an N-dimensional survival function with margins $S_1, \ldots, S_N$. Then, $\mathcal{S}$ has a copula representation:*

$$\mathcal{S}(t_1, \ldots, t_N) = \mathcal{C}(S_1(t_1), \ldots, S_N(t_N))$$

The copula $\mathcal{C}$ is unique if the margins are continuous.

Archimedean copula model has the following representation:

$$H(u, v) = \phi^{-1}(\phi(u) + \phi(v)), \quad (u, v) \in [0, 1]^2$$

where $\phi : [0, 1] \to [0, +\infty]$ is a function satisfying $\phi(1) = 0, \phi(0) = \infty, \phi'(x) < 0$ and $\phi''(x) > 0$. Then $H(u, v)$ is a distribution function on$[0, 1]^2$ with uniform marginals.

Commonly used Archimedean copula models include:

- Clayton copula, where $\phi(u, \alpha) = u^{-(\alpha-1)} - 1$,
- Frank copula, where $\phi(u, \theta) = \log \frac{1-\theta}{1-\theta^u}$,
- Gumbel copula, where $\phi(u, \theta) = (-\log u)^\theta$.

Parametric Clayton Copula Likelihood

The joint likelihood of bivariate time to events data can be written as

$$L = \prod_i f(t_1, t_2)^{\delta_1 \cdot \delta_2} \cdot -S_{12}(t_1, t_2)^{\delta_1 \cdot (1-\delta_2)} \cdot -S_{21}(t_1, t_2)^{(1-\delta_1) \cdot \delta_2} \cdot S(t_1, t_2)^{(1-\delta_1) \cdot (1-\delta_2)} \quad (2.7)$$

where $S(t_1, t_2)$is the joint survival function and

$$S_{12}(t_1, t_2) = \frac{\partial S(t_1, t_2)}{\partial t_1} \quad (2.8)$$

$$S_{21}(t_1, t_2) = \frac{\partial S(t_1, t_2)}{\partial t_2} \quad (2.9)$$

$$f(t_1, t_2) = \frac{\partial^2 S(t_1, t_2)}{\partial t_1 \partial t_2} \quad (2.10)$$

Under the Clayton copula structure. The likelihood can be written as

$$\begin{aligned} L_i = f(t_{i1}, t_{i2})^{\Delta_{i1}\Delta_{i2}} \cdot [-\frac{\partial C_{\alpha(\boldsymbol{\beta};t_1,t_2)}(t_{i1}, t_{i2})}{\partial t_{i1}}]^{\Delta_{i1}(1-\Delta_{i2})} \\ \times [-\frac{\partial C_{\alpha(\boldsymbol{\beta};t_1,t_2)}(t_{i1}, t_{i2})}{\partial t_{i2}}]^{(1-\Delta_{i1})\Delta_{i2}} \cdot C_{\alpha(\boldsymbol{\beta};t_1,t_2)}(t_{i1}, t_{i2})^{(1-\Delta_{i1})(1-\Delta_{i2})} \end{aligned} \quad (2.11)$$

Based on (2.6) and $\frac{\partial S(t)}{\partial t} = -h(t) \cdot S(t)$, we have

By symmetry,

$$\begin{aligned} &\frac{\partial C_{\alpha(\boldsymbol{\beta};t_1,t_2)}(t_1, t_2)}{\partial t_2} \\ &= -S(t_1, t_2) \cdot \{S_1(t_1)^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} + S_2(t_2)^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} - 1\}^{-1} \\ &\quad \cdot S_2(t_2)^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} \cdot h_2(t_2) \end{aligned} \quad (2.12)$$

Then

$$\begin{aligned}
&\frac{\partial C_{\alpha(\boldsymbol{\beta};t_1,t_2)}(t_1,t_2)}{\partial t_1} \\
&= -\{S_1(t_1)^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} + S_2(t_2)^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} - 1\}^{-\frac{1}{\alpha(\boldsymbol{\beta};t_1,t_2)-1}-1} \\
&\qquad \cdot S_1(t_1)^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} \cdot h_1(t_1) \\
&= -S(t_1,t_2) \cdot \{S_1(t_1)^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} + S_2(t_2)^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} - 1\}^{-1} \\
&\qquad \cdot S_1(t_1)^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} \cdot h_1(t_1)
\end{aligned} \tag{2.13}$$

$$\begin{aligned}
f(t_1,t_2) &= \frac{\partial^2 C_{\alpha(\boldsymbol{\beta};t_1,t_2)}(t_1,t_2)}{\partial t_1 \partial t_2} \\
&= (1+\alpha(\boldsymbol{\beta};t_1,t_2)) \cdot h_1(t_1) \cdot S_1(t_1)^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} \\
&\quad \cdot \{S_1(t_1)^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} + S_2(t_2)^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} - 1\}^{-\frac{1}{\alpha(\boldsymbol{\beta};t_1,t_2)-1}-2} \\
&\quad \cdot h_2(t_2) \cdot S_2(t_2)^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)}
\end{aligned} \tag{2.14}$$

The joint likelihood for all the observation is $L = \prod_{i=1}^{n} L_i$. Let $\phi = (\boldsymbol{\gamma_1}', \boldsymbol{\gamma_2}', \boldsymbol{\beta})$, where $\boldsymbol{\gamma_1}'$,$\boldsymbol{\gamma_2}'$ are the parameters in the marginal survival distribution $S_1(t_1)$ and $S_2(t_2)$), respectively. $\boldsymbol{\beta}$ is the parameter for the covariate in the cross ratio function. $U_{\boldsymbol{\gamma_1}'}(\phi), U_{\boldsymbol{\gamma_2}'}(\phi), U_{\boldsymbol{\beta}}(\phi)$ are the score functions which are essentially the first derivative of the log of (2.11) for $\boldsymbol{\gamma_1}', \boldsymbol{\gamma_2}', \boldsymbol{\beta}$. Maximum likelihood estimate $\hat{\phi}$ is the solution to $U_{\boldsymbol{\gamma_1}'}(\phi) = 0, U_{\boldsymbol{\gamma_2}'}(\phi) = 0, U_{\boldsymbol{\beta}}(\phi) = 0$. Under Cox and Hinkley (1979) regularity conditions , $n^{\frac{1}{2}}(\hat{\phi} - \phi_0)$ converges to multivariate normal with mean vector zero and

variance-covariance matrix $\boldsymbol{I}^{-1}$, where $\boldsymbol{I}$ is the information matrix obtained from the second derivative of likelihood equation (2.11)(Cox and Oakes, 1984). Given full parametric functions of the marginal survival functions, maximum likelihood estimates of $\boldsymbol{\beta}$ as well as parameters in the marginal survival functions can be obtained.

2.4.2 Two Stage Semiparametric Estimation Approach

In the parametric approach described above, the two marginal survival functions are assumed to be fully specified and the joint survival model follows a Clayton copula. In a two-stage semiparametric estimation approach, the marginal survival functions are estimated by the nonparametric Kaplan-Meier approach as $\hat{S}_1$ and $\hat{S}_2$ first. The cross ratio parameter, $\hat{\boldsymbol{\beta}}$, is then estimated at the second stage by maximizing the pseudolikelihood function $L(\hat{S}_1, \hat{S}_2, \boldsymbol{\beta})$.

Write (u_i, v_i) for the non parametric estimator of $(S_1(X_{1i}), S_2(X_{2i}))$. Then given $(u_i, v_i),\ j = 1, \ldots, n$, the likelihood of $\boldsymbol{\beta}$ is

$$L^{pseudo}(\boldsymbol{\beta}, u_i, v_i) = \prod_i f_{\alpha(\boldsymbol{\beta};t_1,t_2)}(u_i, v_i)^{\Delta_{1i}\cdot\Delta_{2i}} \cdot \frac{\partial C_{\alpha(\boldsymbol{\beta};t_1,t_2)}(u_i, v_i)}{\partial u_i}^{\Delta_{1i}\cdot(1-\Delta_{2i})} \cdot \frac{\partial C_{\alpha(\boldsymbol{\beta};t_1,t_2)}(u_i, v_i)}{\partial v_i}^{(1-\Delta_{1i})\cdot\Delta_{2i}} \cdot C_{\alpha(\boldsymbol{\beta};t_1,t_2)}(u_i, v_i)^{(1-\Delta_{1i})\cdot(1-\Delta_{2i})} \tag{2.15}$$

where

$$C(u, v; \alpha(\boldsymbol{\beta}; t_1, t_2)) = \{u^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} + v^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} - 1\}^{-\frac{1}{\alpha(\boldsymbol{\beta};t_1,t_2)-1}} \tag{2.16}$$

$$\frac{\partial C}{\partial u} = \{u^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} + v^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} - 1\}^{-\frac{1}{\alpha(\boldsymbol{\beta};t_1,t_2)-1}-1} \cdot u^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)-1} \tag{2.17}$$

$$\frac{\partial C}{\partial v} = \{u^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} + v^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} - 1\}^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)-1} \cdot v^{-\frac{1}{\alpha(\boldsymbol{\beta};t_1,t_2)-1}-1} \tag{2.18}$$

$$\begin{aligned}\frac{\partial^2 C}{\partial u \partial v} =&(1 + \frac{1}{\alpha(\boldsymbol{\beta};t_1,t_2)-1})\{u^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} + v^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)} - 1\}^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)-2} \\ &\cdot u^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)-1} \cdot v^{-(\alpha(\boldsymbol{\beta};t_1,t_2)-1)-1}\end{aligned} \tag{2.19}$$

Let $l(\boldsymbol{\beta}, S_1, S_2)$ be the log likelihood function in equation (4.8) and $U(\boldsymbol{\beta}, S_1, S_2)$ the score function of $\boldsymbol{\beta}$, then

$$U(\boldsymbol{\beta}, \hat{S_1}, \hat{S_2}) = \frac{\partial l(\boldsymbol{\beta}, \hat{S_1}, \hat{S_2})}{\partial \boldsymbol{\beta}} \tag{2.20}$$

The pseudo likelihood estimator $\boldsymbol{\beta}^\star$ is the solution to score function.

To estimate standard error, we extend the results from Theorem 2 in Shih and Louis (1995) by chain rule of derivation.We use the notation from Shih and Louis (1995). Let cross ratio α be a function of covariate of interest $\boldsymbol{\beta}$, i.e. $\alpha(\boldsymbol{\beta})$. Then, according to chain rule, we have

$$\begin{aligned}
W_{\boldsymbol{\beta}} &= \frac{\partial l(\alpha(\boldsymbol{\beta};t_1,t_2),u,v)}{\partial \boldsymbol{\beta}} = \frac{\partial l(\alpha(\boldsymbol{\beta};t_1,t_2),u,v)}{\partial \alpha(\boldsymbol{\beta};t_1,t_2)} \cdot \frac{\partial \alpha(\boldsymbol{\beta};t_1,t_2)}{\partial \boldsymbol{\beta}} \\
V_{\boldsymbol{\beta}} &= \frac{\partial^2 l(\alpha(\boldsymbol{\beta};t_1,t_2),u,v)}{\partial \boldsymbol{\beta}^2} \\
&= \frac{\partial^2 l(\alpha(\boldsymbol{\beta};t_1,t_2),u,v)}{\partial \alpha(\boldsymbol{\beta};t_1,t_2)^2} \cdot (\frac{\partial \alpha(\boldsymbol{\beta};t_1,t_2)}{\partial \boldsymbol{\beta}})^2 + \frac{\partial l(\alpha(\boldsymbol{\beta};t_1,t_2),u,v)}{\partial \alpha} \cdot \frac{\partial^2 \alpha(\boldsymbol{\beta};t_1,t_2)}{\partial \boldsymbol{\beta}^2} \\
V_{\boldsymbol{\beta},1} &= \frac{\partial l(\alpha(\boldsymbol{\beta};t_1,t_2),u,v)}{\partial \boldsymbol{\beta} \partial u} = \frac{\partial l(\alpha(\boldsymbol{\beta};t_1,t_2),u,v)}{\partial \alpha(\boldsymbol{\beta};t_1,t_2) \partial u} \cdot \frac{\partial \alpha(\boldsymbol{\beta};t_1,t_2)}{\partial \boldsymbol{\beta}} \\
V_{\boldsymbol{\beta},2} &= \frac{\partial l(\alpha(\boldsymbol{\beta};t_1,t_2),u,v)}{\partial \boldsymbol{\beta} \partial v} = \frac{\partial l(\alpha(\boldsymbol{\beta};t_1,t_2),u,v)}{\partial \alpha(\boldsymbol{\beta};t_1,t_2) \partial v} \cdot \frac{\partial \alpha(\boldsymbol{\beta};t_1,t_2)}{\partial \boldsymbol{\beta}}
\end{aligned} \tag{2.21}$$

The estimator for standard error can be expressed as

$$\hat{\tau} = \frac{\hat{\tau}_1^2 + \hat{\tau}_2^2}{\hat{\tau}_1^4}, \tag{2.22}$$

where $\hat{\tau}_1^2$ is the model based variance estimator and can be obtained from the second derivative of pseudo likelihood

$$\hat{\tau}_1^2 = \frac{1}{2}\sum_{i=1}^{n} -V_{\boldsymbol{\beta}}(\boldsymbol{\beta}^\star, \hat{S}_1(X_{1i}, X_{2i})),$$
$$\hat{\tau}_2^2 = \frac{1}{n}\sum_{i=1}^{n}[\hat{I}_1(X_{1i}, \Delta_{1i}, \boldsymbol{\beta}^\star) + \hat{I}_2(X_{2i}, \Delta_{2i}, \boldsymbol{\beta}^\star)]^2$$

where

$$\hat{I}_1(X_{1k}, \delta_{1k}, \boldsymbol{\beta}^\star) = \frac{1}{n}\sum_{k} V_{\boldsymbol{\beta},1}(\boldsymbol{\beta}^\star, \hat{S}_1(X_{1k}, X_{2k}))\hat{I}_1^0(X_{1k}, \delta_{1k})(X_{1k}),$$
$$\hat{I}_2(X_{2k}, \delta_{2k}, \boldsymbol{\beta}^\star) = \frac{1}{n}\sum_{k} V_{\boldsymbol{\beta},2}(\boldsymbol{\beta}^\star, \hat{S}_1(X_{1k}, X_{2k}))\hat{I}_2^0(X_{2k}, \delta_{2k})(X_{2k})$$

and

$$\hat{I}_1^0(X_{1i}, \delta_{1i})(X_{1k}) = -\hat{S}_1(X_{1k})\{\frac{IX_{1i} \le X_{1k}, \Delta_{1i} = 1}{\hat{p}_{1i}} - \sum_{X_{1l} \le X_{1i}, X_{1k}} \frac{\Delta\hat{\Lambda}_1(X_{1l})}{\hat{p}_{1l}}\},$$
$$\hat{I}_2^0(X_{2i}, \delta_{2i})(X_{2k}) = -\hat{S}_2(X_{2k})\{\frac{IX_{2i} \le X_{2k}, \Delta_{2i} = 1}{\hat{p}_{2i}} - \sum_{X_{2l} \le X_{2i}, X_{2k}} \frac{\Delta\hat{\Lambda}_2(X_{2l})}{\hat{p}_{2l}}\}$$

$\Delta\hat{\Lambda}_i(t)$ is a Nelson's estimator, which can be calculated as $\Delta\hat{\Lambda}_i(t) = \frac{I\{\bar{Y}_i(t)>0\}}{\bar{Y}_i(t)} d\bar{N}_i(t)$, where $\bar{Y}_i(t) = \sum_j I\{X_{ij} \ge t\}$ and $N_i(t) = \sum_j N_{ij}(t)$. $\boldsymbol{\beta}^\star$ is the solution for score equation in (2.20). Shih and Louis (1995) showed that if cross ratio $\alpha(\boldsymbol{\beta}; t_1, t_2) = \alpha$ is constant, under regularity conditions $\hat{\tau}^2$ is a consistent estimator for standard error.

2.4.3 Nonparametric Pseudo-Partial Likelihood Estimation Approach

The nonparametric approach from Hu et al. (2011) is motivated by the Cox proportional hazards model, which is use to capture the local feature of the dependence structure (Hu, 2011). The idea is to group observations into distinct strata by covariate values, then using one survival time as exposure and using the other survival time as the outcome in order to construct a pseudo-partial likelihood function.

The procedure to construct the pseudo partial likelihood followed the interpretation of conditional hazard ratio in epidemiology terminology. If we treat $\{j : T_{1j} = t_1\}$ and $\{j : T_{1j} > t_1\}$ as"exposure" and"non-exposure" groups, respectively, then from (2.2) , the cross ratio can be interpreted as the hazard ratio of T_2 between these two groups within the stratum $W = w$. Given $t_1 = X_{1i}$, by mimicking the partial likelihood used in the Cox's models, Hu et al. (2011) proposed the following pseudo-partial likelihood function:

$$\prod_{j=1}^{n}[\frac{h_2(X_{2j}|X_{1j} = X_{1i}, \mathbf{w}_j = \mathbf{w}_i)^{I(X_{1j}=X_{1i})}}{\sum_{X_{2k}\geq X_{2j}} I(X_{1k} \geq X_{1i})h_2(X_{2j}|X_{1j} = X_{1i}, \mathbf{w}_j = \mathbf{w}_i)^{I(X_{1k}=X_{1i})}}]^{I(X_{1j}\geq X_{1i})\Delta_{2j}\Delta_{1i}} \tag{2.23}$$

$$\prod_{j=1}^{n}[\frac{h_2(X_{2j}|X_{1j} > X_{1i}, \mathbf{w}_j = \mathbf{w}_i) \cdot \alpha(X_{1i}, X_{2j}, \mathbf{w}_j)^{I(X_{1j}=X_{1i})}}{\sum_{X_{2k}\geq X_{2j}} I(X_{1k} \geq X_{1i})h_2(X_{2j}|X_{1j} > X_{1i}, \mathbf{w}_j = \mathbf{w}_i) \cdot \alpha(X_{1i}, X_{2j}, \mathbf{w}_j)^{I(X_{1k}=X_{1i})}}]^{I_{ij}} \tag{2.24}$$

Where $I_{ij} = I(X_{1j} \geq X_{1i})\Delta_{2j}\Delta_{1i}$. With some simplification, the above equation can be write as

$$\prod_{j=1}^{n}[\frac{\alpha(X_{1i}, X_{2j}, \mathbf{w}_j)^{I(X_{1j}=X_{1i})}}{\sum_{X_{2k}\geq X_{2j}} I(X_{1k} \geq X_{1i})\alpha(X_{1i}, X_{2j}, \mathbf{w}_j)^{I(X_{1k}=X_{1i})}}]^{I(X_{1j}\geq X_{1i})\Delta_{2j}\Delta_{1i}} \tag{2.25}$$

Denote (4.9) as $L_i^{(1)}$. Considering the symmetric structure of the definition of $\theta(t_1, t_2, w)$, given $t_2 = X_{2i}$, we have :

$$L_i^{(2)} = \prod_{j=1}^{n} [\frac{\alpha(X_{2i}, X_{1j}, \mathbf{w}_j)^{I(X_{2j}=X_{2i})}}{\sum_{X_{1k} \geq X_{1j}} I(X_{2k} \geq X_{2i}) \alpha(X_{2i}, X_{1j}, \mathbf{w}_j)^{I(X_{2k}=X_{2i})}}]^{I(X_{2j} \geq X_{2i}) \Delta_{1j} \Delta_{2i}} \quad (2.26)$$

The finial pseudo-partial likelihood function can be obtained, by multiplying these two objective functions from all subjects,

$$L_n = \prod_{i=1}^{n} L_i^{(1)} \cdot L_i^{(2)} \quad (2.27)$$

The estimator obtained by maximizing (3.28) is then called the pseudo-partial likelihood estimator.

Hu et al.(2011) proved that, under some regularity conditions, the maximum pseudo-partial likelihood estimator $\boldsymbol{\beta}$ have $n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a normal random variable with mean zero and variance $I(\boldsymbol{\beta})^{-1} \Sigma_(\boldsymbol{\beta}) I(\boldsymbol{\beta})^{-1}$, where $I(\boldsymbol{\beta}) = 2E(\Delta_1 \cdot \Delta_2 \cdot \mathbf{w}^2)$ and $\Sigma(\boldsymbol{\beta})$ is the asymptotic variance for $U_n(\boldsymbol{\beta}) = \frac{\partial \log L_n}{\partial \boldsymbol{\beta}}$, which can be estimated using sample variance.

For continuous covariates, the observation with "relative close" covariate value can be combined into the same "group". This can be achieved by replacing the indicator function $I(W_j = W_i)$ by a kernel function $K_h(W_j - W_i)$ in (4.9) and (4.10).

2.5 Simulation Study

We conducted a simulation study to evaluate the performance of these three estimation approaches under covariate dependent cross ratio setup. Since simulating data

from a bivariate distribution with an arbitrary cross ratio function is most possible because there may not be a corresponding closed form survival function, we simulated data from a Clayton model with piecewise constant cross ratio following Nan et al. (2006). This simulation setup had also been used in He and Lawless (2003).

Two simulation scenarios were considered, the first scenario was used to demonstrate the performance of each estimation approach under model is correctly specified and with identical marginal distribution; the second scenario was used to demonstrate the performance of each estimation approach under model is misspecified. For all three scenarios both equal and unequal censoring percentage of T_1 and T_2 were considered.

2.5.1 Data Setup

Bivariate data (T_{1i}, T_{2i}), $i = 1, \ldots, n$, were generated one component at a time. First, T_{1i} was generated from the uniform variate $u_{i1} \sim U[0,1]$ by

$$T_{i1} = [\frac{-\log(u_{i1})}{\lambda_1}]^{\frac{1}{p_1}} \tag{2.28}$$

Then, T_{i2} was generated from the independent variate $u_{i2} \sim U[0,1]$ by

$$T_{i2} = [\frac{1}{-(\alpha_i(\boldsymbol{\beta}; t_1, t_2) - 1)} \cdot \frac{\log(1 - u_{i1}^{1-\alpha_i(\boldsymbol{\beta};t_1,t_2)} + u_{i1}^{1-\alpha_i(\boldsymbol{\beta};t_1,t_2)} \cdot u_{i2}^{-\frac{\alpha_i(\boldsymbol{\beta};t_1,t_2)}{\alpha_i(\boldsymbol{\beta};t_1,t_2)-1}})}{\lambda_2}]^{\frac{1}{p_2}} \tag{2.29}$$

In our simulation study, the cross ratio function was setup as $\alpha_i(\boldsymbol{\beta}; t_1, t_2) = \alpha_0 \cdot exp(\boldsymbol{\beta} \cdot \mathbf{w}_i)$ and $\alpha_0 = 3$ and $\beta = 0.5$, and we used $w_i \sim Bernoulli(0.5)$. We considered both uncensored and censored samples. For the censored cases, bivariate censoring

times C_{i1} and C_{i2} were generated independently from uniform distributions.$C \sim Uniform(0, 4.8)$ or $C \sim Uniform(0, 4.1)$, with probability of censoring 10% or 30%, respectively. For each scenario, we generated 1000 simulate samples; sizes $n = 100$, $n = 400$ and $n = 800$ were considered.

For each estimation approach, we calculated relative bias, standard error estimate, and the estimated coverage probability rates of 95% confidence intervals using the asymptotic normal distribution assumption for each of the estimates. The relative bias were calculated as the difference between estimates and true value divided by the true value, $\frac{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}}{\boldsymbol{\beta}}$.

2.5.2 Simulation Results when the Model is Correctly Specified

Table (2.1) summarized the simulation results for model is correctly specified scenario, the relative bias, model based standard error, empirical standard error of parametric Clayton copula approach, semiparametric two stage approach and nonparametric pseudo partial likelihood estimators of the association are given. The data was generated from Clayton copula with two levels of cross ratio and identical exponential distribution as marginal survival. The marginal survival were generated from identical exponential distribution, $S_i(t) = \exp(-t)$, where $i = 1, 2$. And the true value of β was 0.5. The Table (2.1) presented the results for equal percentage censoring scenario and unbalanced censoring senior.

For no censoring case, the bias of parametric Clayton copula estimates was the smallest among three estimation approach, this results was as expected since the data were generated from Clayton copula and Clayton copula estimation approach can revival all the information in the simulated data by using correct likelihood. We

also found that, as expected, the bias and error estimates were decreasing as sample size increasing for all three estimates.

When censoring percentages were equal, the bias and error estimates were increasing as censoring percentage increasing. We also found that under moderate percentage of censoring Clayton copula approach performed adequately well, but if the censoring percentage were considerable, the performance of Clayton copula approach was not ideal compare to two stage semiparametric approach and nonparametric pseudo partial likelihood approach. And the two stage semiparametic estimates and non-parametric pseudo partial likelihood estimates performed more robust results against censoring compare to Clayton copula approach. But among three estimates, the nonparametric pseudo partial likelihood(PPL) estimates had smallest inflammation percentage of the bias, which indicated that the non parametric PPL estimate was the most robust estimates against censoring.

In unbalanced censoring scenario, we found that nonparametric PPL estimates was the most robust and accurate estimate among three estimates. Both parametric Clayton copula estimate and semiparametric two stage estimate approach were very sensitive to unbalanced censoring scenario. Especially, the performance of these two estimates were quite poor if the censoring percentage was quite different between two event, this drawback may resulted by using the Clayton copula structure during the estimation for these two approaches, since both parametric Clayton copula approach and semiparametric two stage approach were highly relied on Clayton copula structure. From the likelihood formula in (2.11), the unbalanced censoring won't influence much on the joint survival $S(t_1, t_2) = C_\alpha(t_1, t_2)$, but the $\frac{\partial S(t_1,t_2)}{\partial t1} = \frac{\partial C_\alpha(t_1,t_2)}{\partial t1}$ and $\frac{\partial S(t_1,t_2)}{\partial t2} = \frac{\partial C_\alpha(t_1,t_2)}{\partial t2}$ will be influenced due to the correlation between two events.

Table 2.1: Simulation Result for Covariate Dependent Cross Ratio Estimation with Correctly Specified Model Scenario: Marginal Survival of T_1 and T_2 are Exponential Distribution, the true $\beta = 0.5$

		Parametric(Clayton Copula)				Semiparametric(Two Stage)				Nonparametric(Pseudo Partial Likelihood)			
T_1	T_2	$R.Bias_\beta$	$M.SE_\beta$	$E.SE_\beta$	$M.CP_\beta$	$R.Bias_\beta$	$M.SE_\beta$	$E.SE_\beta$	$M.CP_\beta$	$R.Bias_\beta$	$M.SE_\beta$	$E.SE_\beta$	$M.CP_\beta$
sample size=100													
0 censor	0 censor	0.034	0.185	0.177	0.95	-0.019	0.171	0.195	0.96	0.064	0.173	0.169	0.93
10% censor	10% censor	0.063	0.203	0.216	0.92	-0.009	0.184	0.211	0.94	0.087	0.224	0.219	0.93
30% censor	30% censor	0.056	0.244	0.251	0.95	0.109	0.239	0.238	0.92	0.146	0.530	0.518	0.9
0 censor	10% censor	0.054	0.195	0.190	0.94	-0.097	0.180	0.209	0.94	0.082	0.199	0.195	0.94
10% censor	30% censor	0.052	0.234	0.238	0.92	-0.115	0.213	0.234	0.96	0.100	0.373	0.385	0.97
30% censor	0 censor	0.066	0.231	0.239	0.92	0.029	0.207	0.216	0.96	0.109	0.347	0.349	0.95
sample size=400													
0 censor	0 censor	0.001	0.091	0.100	0.93	-0.029	0.086	0.082	0.96	0.026	0.078	0.078	0.95
10% censor	10% censor	0.032	0.100	0.105	0.92	0.016	0.097	0.087	0.92	0.039	0.101	0.120	0.96
30% censor	30% censor	-0.043	0.121	0.129	0.97	0.127	0.195	0.114	0.76	0.046	0.242	0.237	0.93
0 censor	10% censor	0.014	0.096	0.101	0.94	-0.060	0.106	0.088	0.88	0.032	0.089	0.087	0.95
10% censor	30% censor	0.012	0.116	0.132	0.91	-0.036	0.134	0.105	0.84	0.051	0.172	0.169	0.92
30% censor	0 censor	-0.018	0.114	0.122	0.94	-0.099	0.131	0.102	0.86	0.039	0.159	0.156	0.9
sample size=800													
0 censor	0 censor	0.013	0.064	0.064	0.95	-0.028	0.057	0.073	0.88	0.001	0.053	0.054	0.96
10% censor	10% censor	0.044	0.071	0.074	0.92	-0.021	0.061	0.079	0.86	0.014	0.069	0.068	0.94
30% censor	30% censor	-0.011	0.086	0.086	0.95	0.032	0.080	0.101	0.88	0.028	0.165	0.161	0.93
0 censor	10% censor	0.030	0.068	0.066	0.97	-0.129	0.062	0.078	0.72	0.006	0.061	0.060	0.93
10% censor	30% censor	0.022	0.082	0.076	0.97	-0.117	0.074	0.100	0.78	0.022	0.118	0.115	0.94
30% censor	0 censor	0.026	0.081	0.087	0.91	-0.137	0.072	0.082	0.8	0.017	0.108	0.106	0.95

R.Bias: Relative bias

M.SE: Model based standard error.

E.SE: Empirical standard error.

M.CP: 95 % coverage probability based on M.SE.

2.5.3 Simulation Results when the Model is Misspecified

Table (2.2) summarized the simulation results for model is misspecified specified, the relative bias, model based standard error, empirical standard error of parametric Clayton copula approach, semiparametric two stage approach and nonparametric pseudo partial likelihood estimators of the association are given.The data were generated from Clayton copula model with Weibull distribution as identical marginal, $S_i(t) = \exp(-2 \cdot t^{\frac{1}{3}})$, where $i = 1, 2$.

From the Table (2.2), we found that the compare to the semiparametric two stage estimate and nonparametric PPL estimate, the Clayton copula estimate was more sensitive to the structure of the marginal survival. If the marginal is misspecified, the estimate from Clayton copula approach performed poorly compare to the others. The two stage estimation was much less affected by the misspecification of the marginal model in contrast to the parametric approach, since in semiparametric two stage approach, the marginal distributions were estimated in the first stage using nonparametric estimates, so it can handle the misspecification of marginal in the first stage and lead to a relative accurate cross ratio estimate in the second stage. The nonparametric pseudo partial likelihood approach provides superior and robust estimation compare to the other two approaches, since the nonparametric approach was not rely on the information of marginal distribution.

Table 2.2: Simulation Result for Covariate Dependent Cross Ratio Estimation with Mis-specified Model Scenario: Marginal Survival: Marginal Survival of T_1 and T_2 are Weibull Distribution with $\lambda = 2, p = 3$, the true $\beta = 0.5$.

		Parametric(Clayton Copula)				Semiparametric(Two Stage)				Nonparametric(Pseudo Partial Likelihood)			
T_1	T_2	$R.Bias_\beta$	$M.SE_\beta$	$E.SE_\beta$	$M.CP_\beta$	$R.Bias_\beta$	$M.SE_\beta$	$E.SE_\beta$	$M.CP_\beta$	$R.Bias_\beta$	$M.SE_\beta$	$E.SE_\beta$	$M.CP_\beta$
sample size=100													
0 censor	0 censor	-0.241	0.165	0.183	0.83	-0.087	0.170	0.215	0.9	0.060	0.172	0.170	0.93
10% censor	10% censor	-0.246	0.176	0.197	0.87	-0.090	0.184	0.230	0.88	0.071	0.209	0.207	0.94
30% censor	30% censor	-0.156	0.234	0.392	0.9	-0.172	0.250	0.260	0.98	0.105	0.494	0.488	0.93
0 censor	10% censor	-0.243	0.170	0.189	0.85	-0.108	0.176	0.248	0.84	0.074	0.191	0.189	0.92
10% censor	30% censor	-0.255	0.202	0.202	0.88	-0.306	0.216	0.268	0.82	0.068	0.324	0.340	0.96
30% censor	0 censor	-0.231	0.197	0.228	0.86	-0.418	0.208	0.261	0.72	0.080	0.296	0.292	0.91
sample size=400													
0 censor	0 censor	-0.284	0.082	0.093	0.55	-0.054	0.082	0.072	0.96	0.026	0.077	0.078	0.95
10% censor	10% censor	-0.285	0.087	0.098	0.58	-0.045	0.087	0.075	0.96	0.036	0.094	0.093	0.93
30% censor	30% censor	-0.275	0.114	0.121	0.75	0.070	0.113	0.143	0.9	0.063	0.220	0.218	0.94
0 censor	10% censor	-0.282	0.085	0.094	0.58	-0.093	0.086	0.084	0.94	0.029	0.085	0.084	0.93
10% censor	30% censor	-0.277	0.100	0.113	0.66	-0.307	0.111	0.125	0.6	0.063	0.146	0.144	0.92
30% censor	0 censor	-0.302	0.097	0.101	0.62	-0.401	0.104	0.108	0.56	0.030	0.132	0.131	0.92
sample size=800													
0 censor	0 censor	-0.298	0.058	0.059	0.28	-0.085	0.186	0.186	0.96	0.001	0.053	0.054	0.96
10% censor	10% censor	-0.293	0.061	0.061	0.31	-0.119	0.200	0.212	0.88	0.009	0.064	0.064	0.93
30% censor	30% censor	-0.260	0.081	0.082	0.63	-0.109	0.270	0.289	0.96	0.037	0.150	0.148	0.93
0 censor	10% censor	-0.297	0.060	0.058	0.3	-0.101	0.192	0.210	0.92	0.006	0.058	0.058	0.92
10% censor	30% censor	-0.287	0.070	0.068	0.44	-0.259	0.234	0.306	0.84	0.017	0.099	0.099	0.95
30% censor	0 censor	-0.285	0.068	0.072	0.46	-0.333	0.227	0.258	0.82	0.012	0.091	0.091	0.95

R.Bias: Relative bias

M.SE: Model based standard error.

E.SE: Empirical standard error.

M.CP: 95 % coverage probability based on M.SE.

2.6 Data Application: Estimate Gender Effect in Cross Ratio between Time to CAD and Depression

In this section, we demonstrate our proposed method in real data application example.

2.6.1 Indianapolis-Ibadan African American Cohort

To illustrate the three estimation approaches in covariate dependent cross ratios, we present a data analysis exploring potential gender differences in the association between time to coronary artery disease (CAD) and time to depression using data from the Indianapolis-Ibadan Dementia Project (IIDP). The Indianapolis-Ibadan dementia project (IIDP) was a 20 year National Institute on Aging funded a longitudinal study of dementia and its risk factors in elderly community-dwelling African Americans living in Indianapolis, Indiana and elderly community-dwelling Yoruba living in Ibadan, Nigeria. Recently, data from the African-American participants in the study were merged with data from the Indiana Network for Patient Care, a regional health information exchange, allowing us to examine medical conditions such as CAD and depression, using electronic medical records(EMR) obtained in the routine care of older adults.

For our analysis, the study population consisted of African American participants of the IIDP. All were age 65 or older residing in Indianapolis, Indiana. Recruitment was conducted at two-time points. During the first recruitment in 1992, 2212 African Americans age 65 or older living in Indianapolis were enrolled in the study. In 2001, the project enrolled 1893 additional African American community-dwelling participants 70 years and older. All participants agreed to undergo regular follow-up

cognitive assessment and clinical evaluations. Details on the assembling of the original cohort and the enrichment cohort were described elsewhere.Hall et al. (2009); Hendrie et al. (2001) Electronic medical records from 1992 to December 31, 2014, were retrieved as a re-identified data set to examine cardiovascular diseases and other risk factors. There were 4105 participants enrolled. After excluding 854 participants who did not have EMR and 28 participants who had CAD or depression before enrollment, there were 3223 participants free of CAD and depression at baseline. Mean age at baseline was 75.6 (standard deviation=6.38) and 68.04 % were women. In Table 2.3 we present the number of participants with incident CAD, depression and both events by gender. Female participants had a higher percentage of depression events and male participant have a higher percentage of CAD events, in addition, the female participants had higher percentages of experience both depression and CAD events compare to male group. Figure (2.2) is the survival plot of CAD event and depression by gender group, from which we can see that the two events are somewhat correlated and there is a slight difference by gender.

Table 2.3: Demographic Characteristic of IIDP Data with Number of Event and Incidence Rate by Each Gender Group

Gender	Total	CAD	Depression	CAD and Depression
Female	2192	822 (37.5%)	479 (21.85%)	193 (9.55%)
Male	1031	396(38.4%)	138 (13.38%)	62 (6.01%)
Total	3223	1191(36.95%)	617(19.14%)	271(8.4%)

In Table 2.4, we present the median age for each disease onset by gender and the status of the other disease. For the male group, we found that participants who had one disease had an earlier onset of the other disease. However, female participants

with depression actually had a slightly later onset of CAD while female participants with CAD had the similar age of onset for depression as those without CAD. The Figure 2.2 is the cumulative hazard plot for time one disease onset give the other disease status for each gender. From Figure 2.2, we observed that the male group tend to have more close association compare to female group. This result showed that there maybe gender difference in the association between the two diseases.

Table 2.4: Median Age Onset for Each Disease by Gender and the Status of the Other Disease

	CAD				Depression		
	Total	Female	Male		Total	Female	Male
Total	79.01	80.21	79.03	Total	80.23	80.06	80.4
Depression	80.19	80.89	76.39	CAD	80.48	81.05	79.81
No Depression	79.58	79.94	79.2	No CAD	79.01	78.91	81.05

CAD: Cardiovascular event

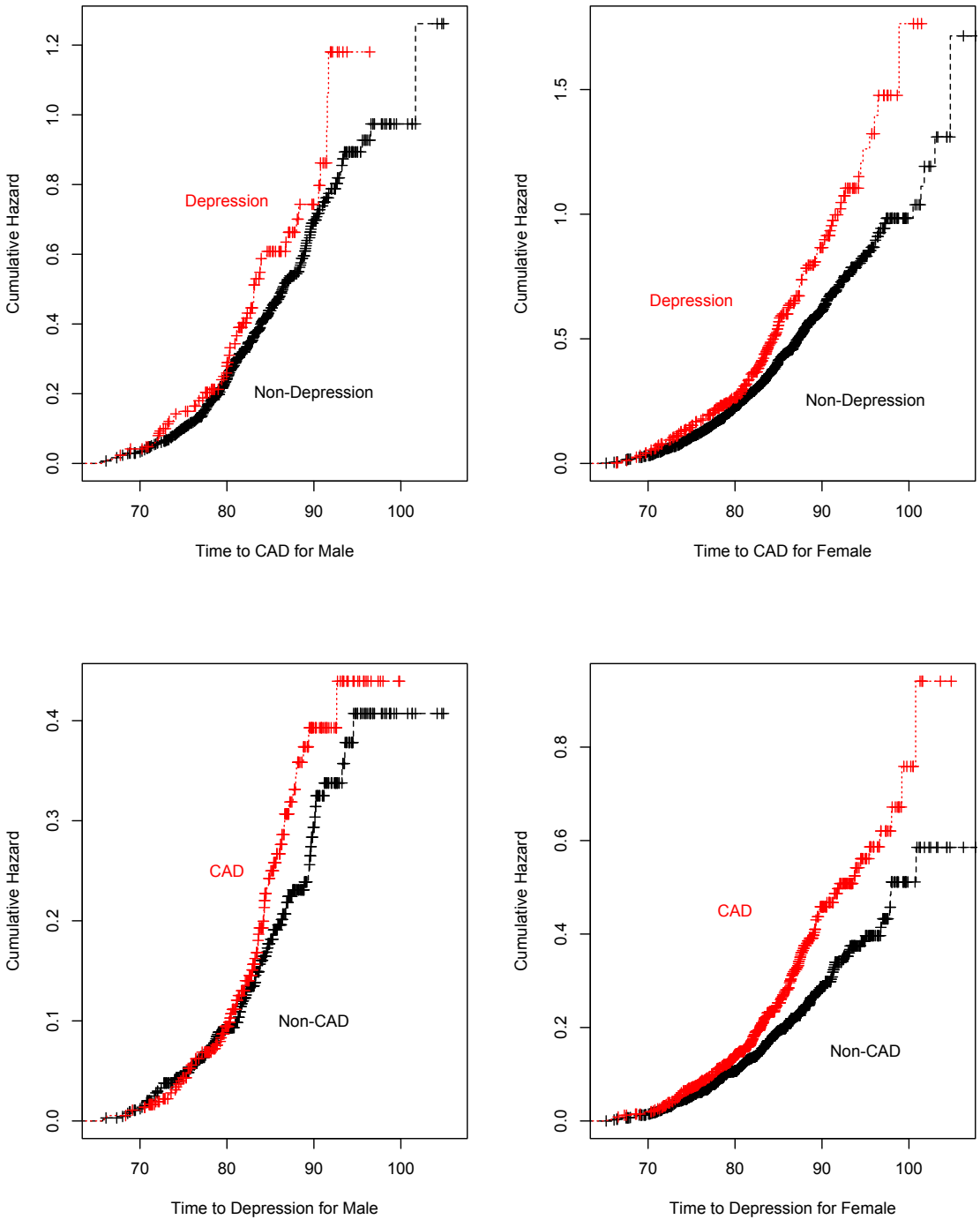
DP: Depression

2.6.2 Estimate Gender Effect in Cross Ratio between Time to CAD and Depression

Denote t_{CAD} as time to CAD and t_{DP} as time to depression; M as male group and F as female group; $h_{CAD}(\cdot)$ as hazard for CAD and h_{DP} as hazard for depression. To estimate the cross ratio as a function of gender, we use following multiplicative model

$$\begin{aligned}\theta(t_{CAD}, t_{DP}; Gender = M) &= \frac{h_{DP}(t_{DP}|T_{CAD} = t_{CAD}, Gender = M)}{h_{DP}(t_{DP}|T_{CAD} > t_{CAD}, Gender = M)} \\ &= \frac{h_{CAD}(t_{CAD}|T_{DP} = t_{DP}, Gender = M)}{h_{CAD}(t_{CAD}|T_{DP} > t_{DP}, Gender = M)} \end{aligned} \tag{2.30}$$

Figure 2.2: Cumulative Hazard Plot of Two Diseases by Gender Group: The first row is the cumulative hazard of time to CAD by Male and Female group respectively, the red line represents depression group, the back line is non-depression group; the second row is the cumulative hazard of time to depression by Male and Female group respectively, the red line is CAD group and black line is Non-CAD group.



and

$$\theta(t_{CAD}, t_{DP}; Gender = M) = \theta_0(t_{CAD}, t_{DP}; I(Gender_i = F)) \cdot exp(I(Gender_i = M) \cdot \beta) \tag{2.31}$$

where $I(Gender_i = M)$is an indicator function for male and $\theta_0(t_1, t_2; I(Gender_i = F))$ is the reference cross-ratio in females, i.e.

$$\begin{aligned} \theta_0(t_{CAD}, t_{DP}; I(Gender_i = F)) &= \frac{h_{DP}(t_{DP}|T_{CAD} = t_{CAD}, Gender = F)}{h_{DP}(t_{DP}|T_{CAD} > t_{CAD}, Gender = F)} \\ &= \frac{h_{CAD}(t_{CAD}|T_{DP} = t_{DP}, Gender = F)}{h_{CAD}(t_{CAD}|T_{DP} > t_{DP}, Gender = F)} \end{aligned} \tag{2.32}$$

Where M indicate male and F indicate female.

Three estimation approaches were used in this data set. Results were presented in Table 2.5. All three estimation approaches showed the estimated cross ratio larger than 1 in the reference group indicating that women who had early onset of one disease were more likely to have an onset of the other disease. The coefficient β for the gender indicator variable in equation (2.31) was estimated to be greater than 0 suggesting that the association between the two disease onsets is stronger in males than the association in females, but this difference is not statistically significant. In order to verify these results, we also conducted stratified analyzes by estimating constant cross ratio in each gender group separately. In Table 2.5, $\widehat{\theta_F}$ and $\widehat{\theta_M}$ represent the cross ratio estimation for each group. All three approaches still showed greater cross ratio estimate in male participants than in female participants. However, only log-ratio of the two nonparametric cross ratio estimates of male over the female was close to the coefficient estimate produced using the nonparametric approach. Both parametric

and two stage semiparametric estimates using the gender specific cross ratio estimates have deviated from regression estimates $\widehat{\beta}$. This may be caused by an invalid Clayton copula distribution assumption. Since nonparametric approach does not rely on any parametric form, it is expected to be more robust to model misspecification.

Table 2.5: Estimates of Covariate Dependent Cross Ratio and Gender Effect in IIDP Data

	Regression		Stratified		
	$\hat{\beta}$(SE)	$\hat{\theta}$ (SE)	$\hat{\theta_M}$ (SE)	$\hat{\theta_F}$(SE)	$\log \frac{\hat{\theta_M}}{\hat{\theta_F}}$
Clayton	0.428(0.05)	1.13(0.04)	1.535(0.14)	1.519(0.12)	0.0104
Two Stage	1.33(0.19)	1.21(0.20)	2.9(0.31)	2.55(0.28)	0.128
Nonparametric	0.828(0.09)	1.05(0.08)	2.42(0.14)	1.06(0.12)	0.827

SE: Standard Error.

2.7 Discussion

We considered covariate dependent cross ratios of bivariate survival times in order to identify covariates that are associated with the co-occupance of two events. We compared three estimation approaches for parameter estimation including a parametric copula approach, a two-stage semi-parametric pseudo-likelihood approach, and a nonparametric pseudo-partial likelihood approach in simulation studies. The nonparametric pseudo-partial likelihood approach proposed by Hu et al. (2011) is shown to perform well under various censoring scenarios and it is also robust against model misspecification.Hu (2011) The parametric copula and the two stage semi-parametric approaches both relied on the correct specification of a joint survival distribution and can produce biased results when such an assumption is violated.

There are several limitations in our proposed setup of the covariate dependent cross ratios. The first limitation is the assumption of a multiplicative covariate effect, *i.e.*, covariates' effect on the cross ratio is multiplicative of the cross-ratio in the reference group. The nonparametric pseudo-partial likelihood approach by Hu et al. (2011) allows the cross ratio function to be modeled as a time-dependent function. Thus, in our extension to allow covariates in the cross ratio set up, the cross ratio function for the reference group can also be modeled as time-dependent. However, our model setup requires that the effect of time be separated from the effect of the covariates, analogous to the proportional hazard assumption. Additional research will be needed for appropriate methods to verify these assumptions in data analyzes.

The second limitation is that our approach considered uninformative censoring. In medical research, informative censoring is often encountered. Thus extending the models for the cross ratio to competing risk or semi-competing risk models is necessary. A number of authors have proposed estimation method under a competing risk by modeling the ratio of concordant and discordant pairs (Bandeen-Roche and Liang, 2002; Bandeen-Roche and Ning, 2008; Ning and Bandeen-Roche, 2014; Shih and Albert, 2010) . This new approach offers a different way to model the association between multiple survival times under informative censoring. It is not clear whether the nonparametric approach of Hu et al. (2011) can be easily extended to accommodate competing risk or semi-competing risk and how such an extension compares to the methods of Ning and Bandeen-Roche (2014). These additional interests can be explored in future research.

It will be an interesting investigation to compare Lawless and Yilmaz (2011)'s two stage semiparametric approach with Shih and Louis (1995)'s two stage semipara-

metric approach. Compare to Shih and Louis (1995)'s approach, Lawless and Yilmaz (2011) proposed to use likelihood (2.33)

$$L^{Lawless} = \prod_i (\frac{\partial^2 C}{\partial t_1 \partial t_2})^{\delta_1 \cdot \delta_2} \cdot (-\frac{\partial C}{\partial t_1})^{\delta_1 \cdot (1-\delta_2)} \cdot (-\frac{\partial C}{\partial t_2})^{(1-\delta_1) \cdot \delta_2} \cdot C^{(1-\delta_1) \cdot (1-\delta_2)} \tag{2.33}$$

Compare to Shih and Louis (1995)'s likelihood

$$L^{Shil} = \prod_i (\frac{\partial^2 C}{\partial u \partial v})^{\delta_1 \cdot \delta_2} \cdot (-\frac{\partial C}{\partial u})^{\delta_1 \cdot (1-\delta_2)} \cdot (-\frac{\partial C}{\partial v})^{(1-\delta_1) \cdot \delta_2} \cdot C^{(1-\delta_1) \cdot (1-\delta_2)} \tag{2.34}$$

The difference is

$$\begin{aligned} L^{Lawless} = L^{Shil} \cdot \prod_i & [h_1(t_1) \cdot S_1(t_1) \cdot h_2(t_2) \cdot S_2(t_2)]^{\delta_1 \cdot \delta_2} \\ & \cdot [h_1(t_1) \cdot S_1(t_1)]^{\delta_1 \cdot (1-\delta_2)} \cdot [h_2(t_2) \cdot S_2(t_2)]^{(1-\delta_1) \cdot \delta_2} \end{aligned} \tag{2.35}$$

It would be meaningful to conducted a simulation study to investigate the difference of those two approaches, especially in finding standard error.

In the data application part, we also considered to test the underline joint survival for bivariate events, but it isn't a trivial test and we hope we can address the test of joint survival for bivariate events in the future work.Also more simulation study could be conducted to investigate the sensitivity of three methodologies. In our simulation, we consider true $\beta = 0.5$, conditional on baseline cross ratio $\alpha_0 = 3$, which is positively correlated. Thus, it would be interested to conduct a simulation study under negative association scenario. And in our simulation, we considered

identical marginal, thus, it would be interested to conduct a study for non identical marginal distribution.

In summary, we demonstrated that a nonparametric pseudo-partial likelihood approach can be used to estimate covariates' effect on cross-ratios between bivariate survival outcomes. We have also shown that the proposed approach performed well and is robust under model misspecification in simulation studies. Given the increasing trends in medical research to study common pathways underlying multiple conditions, the proposed methods can be readily applied to these data for the identification of common risk factors in the association of two survival outcomes.

Funding

This research is supported by funding from the National Institute of Health Grants R01 AG019181, R01 AG0145350, and P30 AG10133.

Chapter 3

Frailty-based Semiparametric Models for Time to Event Data with Semi-competing Risk

3.1 Abstract

Survival analysis of time to events data often encounters the situations of correlated multiple events including the same type of event observed from siblings or multiple events experienced by the same individual. In addition, survival analysis in biomedical research can be further complicated by semi-competing risk when individuals at risk of a particular disease die from other causes. Motivated from illness-death model, we propose a frailty model based approach for survival outcomes with a semi-competing risk to account for the dependence between disease progression time, survival time. Two estimation approaches are proposed and compared. The first is a two-stage semiparametric approach where the cumulative baseline hazard was first estimated by a nonparametric method. Parameter estimation was then achieved by maximizing the pseudo-likelihood functions. In the second approach, we propose to use a penalized partial likelihood approach for parameter estimation and inference similar to the concept of the Cox's partial likelihood. Simulation studies are conducted to compare the performances of these two approaches. The proposed model is applied to data from a longitudinal study of an elderly population.

3.2 Introduction

Semi-competing risk data, first proposed by Fine et al. (2001), refers to the situation in which a terminal event censors a nonterminal event but not vice versa, thus violating the uninformative censoring assumption for traditional survival data. Semi-competing risk data are often encountered in biomedical research including studies of chronic diseases in elderly cohorts and cancer or AIDS trials (Putter et al. (2007) and Wang (2003)).

In a semi-competing risk setting, the terminal and non-terminal events are often correlated and are both of interest. Survival method ignoring the semi-competing risk may yield biased results due to the violation of the independent censoring assumption. In order to handle the informative censoring, a copula approach has been used to jointly model the terminal and the nonterminal events simultaneously(Ding et al., 2009; Fine et al., 2001; Lakhal et al., 2008; Peng and Fine, 2007; Wang, 2003). Peng and Fine (2007) proposed a regression model of semi-competing risks data with a novel time-dependent copula using the proportional hazard model with time-varying coefficient as the marginal models. Lakhal et al. (2008) proposed to use the copula-graphic estimator of Zheng and Klein (1995) for estimating the marginal survival functions of the nonterminal event and use an Archimedean copula for the joint model of both events. However, the applications of the copula model approach have been limited due to model identification issues, the lack of model flexibility and the difficulty in incorporating covariates.

In contrast to the copula approach, Xu et al. (2010) and Han et al. (2014) proposed a frailty model framework for semi-competing risk data. Their general

illness-death models differentiate three types of hazards: hazard of illness, hazard of death without illness and hazard of death with illness. Covariates are incorporated through proportional hazards modeling. In Xu et al. (2010), the two types of events were linked by a gamma frailty and nonparametric maximum likelihood estimation (NPMLE) was used for estimation. Han et al. (2014) proposed a Bayesian Markov Chain Monte Carlo methods (MCMC) for model fitting and a frailty term with normal distribution.

There is an extensive literature on parameter estimation for covariate effects in frailty models, where the event times are assumed to be independent conditional on unobserved frailty terms. The frailties are unobserved random variable assumed to follow a probability distribution, the shape of which is described with a few parameters. In order to handle the additional information introduced by frailty term, the EM algorithm has been widely used in this area. Klein (1992) proposed to use the EM algorithm based on a profile likelihood construction. Since EM algorithm is computational expensive, Cortinas Abrahantes and Burzykowski (2005) proposed an alternative implementation of EM algorithm, in which the expected values were computed with the use of Laplace approximation. McGilchrist and Aisbett (1991) proposed a penalized partial likelihood approach in a Gaussian frailty model setting. Following Breslow and Clayton (1993),Ripatti and Palmgren (2000) used the Laplace approximation for the integrated likelihood. A comprehensive review of methodologies in handling frailty model can be find in Therneau and Grambsch (2000).

In this work, we followed the model framework in Xu et al. (2010) and Han et al. (2014). However, instead of the nonparametric maximize likelihood approach used in Xu et al. (2010) and the Bayesian Markov Chain Monte Carlo methods

(MCMC) used in Han et al. (2014), we propose to use a penalized partial likelihood(PPL) approach for parameter estimation and inferences. The penalized partial likelihood theory has been addressed by Green (1987) in general semiparametric regression frame work where he compared the performance of penalized approach to the composite likelihood approach to show the stabilty and accuracy of the penalized approach by choosing the turning parameter λ using cross validation method. Therneau and Grambsch (2000) showed an exact connection between the shared gamma frailty model and a penalized likelihood procedure. Therneau et al. (2003) also mentioned the closed linked to penalized models and illustrated that the fitting from frailty models with penalized likelihoods can be made quite efficient by taking advantage of computational methods available for penalized models.

In this paper, we demonstrate the advantage of penalized partial likelihood approach on semi-competing risk data through simulation study and application example. Compared to other modeling and estimating approaches, our semiparametric model setup and PPL estimation approach have three advantages.

1. Modeling structure: our model dis-tangles the baseline hazard and covariates in the same spirit as the Cox model. Modeling covariate effects using semiparametric additive function allows for both parametric and nonparametric covariate effects, such as spline covariate, with extensions to multiple covariate and time-varing covariate.
2. Parameter estimation: our approach connects the frailty model with the penalized partial likelihood estimation, which is parallel to the connection between mixed models with penalized least square estimation in Bates and DebRoy (2004).

3. Computation: our approach can be conducted in both SAS and R using current packages respectively by formatting the data accordingly.

In the following sections, we present the notations and model setup in Section 2. We describe estimation approaches in Section 3 and results from a simulation study in Section 4. We present results from the Indiana-Ibadan Dementia Project (IIDP) data analysis in Section 5 and conclude the chapter with a discussion in Section 6.

3.3 Frailty Model in Competing Risk Data and Semi-competing Risk Data

Frailty model provides a convenient way to introduce random effects, association and unobserved heterogeneity into models for survival data. A frailty can be interpreted as an unobserved random proportionality factor that modifies the hazard function of an individual, or of related individuals. The term frailty itself was introduced by Vaupel et al. (1979) in univariate survival models. The frailty model is defined in terms of the conditional hazard

$$\lambda_{ij}(t|u_i) = \lambda_0(t) \cdot u_i \cdot exp(x_{ij}^T \beta)$$

with $i \in I = \{1, \ldots, G\}$ and $j \in J_i = \{1, \ldots, n_i\}$, where $h_0(\cdot)$ is the baseline hazard function, u_i is the frailty term in group i, x_{ij} is the vector of covariates for subject j in group i, and β is the vector of regression coefficients.

Normally, in most clinical applications, survival analysis implicitly assumes a homogeneous population to be studied. This means that all individuals sampled into that study are subjects under the same risk (e.g., risk of death, risk of disease

recurrence). In many applications, the study population can not be assumed to be homogeneous. The frailty approach is a statistical modeling concept which aims to account for heterogeneity, caused by unmeasured covariates. Generally, frailty models can be distinguished into two broad classes:

1. Models with an univariate survival time as the endpoint.
2. Models which describe multivariate survival endpoints (e.g; competing risks, recurrence of events in the same individual, occurrence of a disease in relatives, semi-competing risks).

In the first case, an univariate (independent) lifetime is used to describe the influence of unobserved covariates in a proportional hazards model (heterogeneity). The variability of survival data is split into a part that depends on risk factors, and is therefore theoretically predictable, and a part that is initially unpredictable, even when all relevant information is known. A separation of these two sources of variability has the advantage that heterogeneity can explain some unexpected results or give an alternative interpretation of some results.

In the second case when multivariate survival times are considered, the aim is to account for the dependence in clustered event times. A natural way to model dependence of clustered event times is through the introduction of a cluster-specific random effect - the frailty. This random effect explains the dependence in the sense that had we known the frailty, the events would be independent. In other words, the lifetimes are conditional independent, given the frailty. This approach can be used for survival times of related individuals like family members or recurrent observations on the same person.

In this work, we focus on the second use of frailty model, to explain the dependence in the illness and death on same subject. The examples and illustrations of frailty model can be found in numerous literatures (Box-Steffensmeier et al., 2007; Wienke, 2010). Gorfine and Hsu (2011) provided a new class of frailty-based competing risks models for clustered failure times data. Especially, Liu et al. (2004); Xu et al. (2010) proposed to use frailty model in semi-competing risk data, also known as illness-death model. Xue et al. (2008) focused on the use of frailty model in aging study with competing risk.

3.4 Model and Likelihood

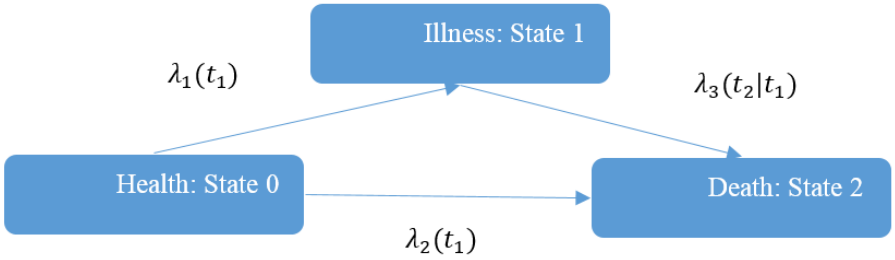
Let T_1 be the time to the non-terminal event (referred to as illness hereafter), T_2 be the time to the terminal event (referred to death hereafter). Let C be an external censoring variable due to patient withdraw or the end of study. We observe the variables $X_1 = T_1 \wedge T_2 \wedge C$ and $X_2 = T_2 \wedge C$. Let $\delta_1 = I(T_1 \leq (T_2 \wedge C))$ and $\delta_2 = I(T_2 \leq C)$, where "$\wedge$ "denotes the minimum and $I(\cdot)$ is the indicator function. Note that T_2 can censor T_1 but not vice visa, whereas C can censor both T_1 and T_2. In addition, a vector of covariate Z is observed. Furthermore, we assume that C is independent of the joint distribution of T_1 and T_2 given Z. Let $\{(T_{1i}, T_{2i}, C_i), i = 1, \ldots, n.\}$ be independent and identically distributed (IID) replications of (T_1, T_2, C). The observed data are IID replications of $(X_1, X_2, \delta_1, \delta_2)$.

3.4.1 Models for Semi-competing Risks Data

Assume individuals begin in an initial healthy state (state 0) from which they may transit to death (state 2) directly or may transit to an illness state (state 1) first

and then to death (state 2) (Figure 3.1). As in Xu et al. (2010), the hazards can be defined as:

Figure 3.1: Two events with a Semi-competing Risk



$$d\Lambda_1(t_1) = \lambda_1(t_1)dt_1 = Pr(t_1 \leq T_1 \leq t_1 + dt_1 | T_1 \geq t_1, T_2 \geq t_1), t_1 > 0 \tag{3.1}$$

$$d\Lambda_2(t_2) = \lambda_2(t_2)dt_2 = Pr(t_2 \leq T_2 \leq t_2 + dt_2 | T_1 \geq t_2, T_2 \geq t_2), t_2 > 0 \tag{3.2}$$

$$d\Lambda_3(t_2|t_1) = \lambda_3(t_2|t_1)dt_2 = Pr(t_2 \leq T_2 \leq t_2 + dt_2 | T_1 = t_1, T_2 \geq t_2), t_2 > t_1 > 0 \tag{3.3}$$

Equations (3.1) and (3.2) are the hazard functions for illness and death without illness, which are the competing risk parts of the model. Equation (3.3) defines the hazard for death following illness. In general, $\lambda_3(t_2|t_1)$ can depend on both t_1 and t_2. To account for the dependency structure between T_1 and T_2, Xu et al. (2010) introduced a single shared gamma frailty term, Han et al. (2014) extended the association model using multivariate random effects as following:

$$\lambda_1(t_1|\mathbf{z}_1, \mathbf{b}) = \lambda_{01}(t_1) \cdot exp(X_1^T \cdot \beta_1 + \mathbf{z}_1^T \cdot \mathbf{b}), t_1 > 0 \tag{3.4}$$

$$\lambda_2(t_2|\mathbf{z}, \mathbf{b}) = \lambda_{02}(t_2) \cdot exp(X_2^T \cdot \beta_2 + \mathbf{z}_2^T \cdot \mathbf{b}), t_2 > 0 \tag{3.5}$$

$$\lambda_3(t_2|t_1, \mathbf{z}, \mathbf{b}) = \lambda_{03}(t_2|t_1) \cdot exp(X_3^T \cdot \beta_3 + \mathbf{z}_3^T \cdot \mathbf{b}), t_2 > t_1 > 0 \tag{3.6}$$

where λ_{0i} is the unspecified baseline hazard; β_i is vectors of regression coefficients associated with each hazard; the covariate X_i has p components and the covariates $\mathbf{z}_i$, usually consists of 1 and a subset of covariates from X_i, is assumed to be associated with random effect $\mathbf{b} = \{b_1, b_2, \ldots, b_q\}^T$. $\mathbf{b}$ represents random effects that account for possible associations among the three hazards. We assume a normal distribution for the random effects, $\mathbf{b} \sim MVN(0, D(\nu))$, with a full rank covariance matrix $D(\nu)$ and ν is a vector of variance components. The zero mean constraint is imposed so that the random effects represent deviations from population averages. Examples of the choices of covariance structures for clusters, hierarchical and spatial survival data can be found, e.g., in Breslow and Clayton (1993). Conditioning on the random effects $\mathbf{b}$, we assume that survival times T_i s is independent of censoring time C_i. We further assume that the censoring times are independent of the random effects $\mathbf{b}$, $i = 1, 2, 3$.

Model (3.4 3.5 3.6) allow multivariate random effects with arbitrary design matrix in the log relative risk. In its simplest form, when $\mathbf{z}_1 = \mathbf{z}_2 = \mathbf{z}_3 = 1$, the frailty term $\mathbf{z}$ is reduced to a univariate random variable that accounts for the subject-specific dependency of three types of hazards. The models in Xu et al. (2010) belong to this simple case where they assume that $exp(\boldsymbol{\beta})$ follows a gamma distribution. However, in many cases, random effects based on covariates, e.g., clinical center or age, may provide better models for the correlation structure. Then the terms $\mathbf{z}_1' \cdot \mathbf{b}$, $\mathbf{z}_2' \cdot \mathbf{b}$ and $\mathbf{z}_3' \cdot \mathbf{b}$ can be used to incorporate these random covariates. For example, clustered semicompeting risks data frequently arises from oncology trials evaluating efficacies of different treatments. A typical model for this type of data is to have both subject-level and cluster-level frailty terms. (Gustafson, 1995, 1997) We assume a normal distribution for the random effects. The zero mean constraint is imposed so that the

random effects represent deviations from population averages. The covariance matrix is assumed to be unconstrained.

Interests on the unknown quantities, $\beta_1, \beta_2, \beta_3, \mathbf{b}, MVN(0, D(\nu)), \lambda_{01}, \lambda_{02}, \lambda_{03}$ can depend on specific analysis. In clinical trial setting, effects of treatment and prognostic factors, $\beta_1, \beta_2, \beta_3$, are usually the focus of primary analysis. For genetic data analysis the focus may be on $MVN(0, D(\nu))$ which captures genetic variability. The baseline hazards are usually treated as nuisance parameters but are needed for the estimation and prediction of survival probabilities for individual subjects.

3.4.2 Likelihood

For a subject j, we observe $(t_{1j}, t_{2j}, \delta_{1i}, \delta_{2i}, c_j, \mathbf{z}_j)$. Using counting process, the three patterns of the event can be represented as the following:

$$N_{1j}(t) = I(t_{1j} \leq t, \delta_{1j} = 1)$$

$$N_{2j}(t) = I(t_{2j} \leq t, \delta_{1j} = 0, \delta_{2j} = 1)$$

$$N_{3j}(t) = I(t_{2j} \leq t, \delta_{1j} = 1, \delta_{2j} = 1).$$

Correspondingly, let the at risk process for the three types of events be represented as following:

$$R_{1j}(t) = I(t_{1j} \geq t)$$

$$R_{2j}(t) = I(t_{1j} \geq t, t_{2j} \geq t)$$

$$R_{3j}(t) = I(t_{2j} \geq t > t_{1j}).$$

We further assume that the censoring time C is independent of X_1, X_2, given covariate Z.

Given frailty, assuming conditional independence of the three events, for subject j, the likelihood is

$$L_i|b = P(T_{1j} = t_{1j}, T_{2j} = t_{2j})^{\delta_{1j}\cdot\delta 2j} \times P(T_{1j} = t_{1j}, T_{2j} \geq t_{2j})^{\delta_{1j}\cdot(1-\delta 2j)} \tag{3.7}$$

$$\times P(T_{1j} \geq t_{1j}, T_{2j} = t_{2j})^{(1-\delta_{1j})\cdot\delta 2j} \times P(T_{1j} \geq t_{1j}, T_{2j} \geq t_{2j})^{(1-\delta_{1j})\cdot(1-\delta 2j)} \tag{3.8}$$

If conditionally on $\mathbf{b}$ the censoring is independent and non-informative also of $\mathbf{b}$, let $\theta = (\beta_1, \beta_2, \beta_3)'$ denote the parameter of interest, then the likelihood can be formed as

$$L_n(\theta) = \int \prod_j L_j(\theta|\mathbf{b}_j) \cdot f(\mathbf{b}_j) d\mathbf{b}_j \tag{3.9}$$

$$= \int \prod_j Pr(T_{1j} = t_{1j}, T_{2j} = t_{2j})^{\delta_{1j}\delta_{2j}} \cdot Pr(T_{1j} = t_{1j}, T_{2j} \geq t_{2j})^{\delta_{1j}(1-\delta_{2j})}$$

$$\cdot Pr(T_{1j} \geq t_{1j}, T_{2j} = t_{2j})^{(1-\delta_{1j})\delta_{2j}} \cdot Pr(T_{1j} \geq t_{1j}, T_{2j} \geq t_{2j})^{(1-\delta_{1j})(1-\delta_{2j})} \cdot f(\mathbf{b}_j) d\mathbf{b}_j$$

According to the definition of hazard (3.1,3.2, 3.3), we have

$$\lambda_3(t_2|t_1) = Pr(t_2 \leq T_2 \leq t_2 + dt_2 | T_1 = t_1, T_2 \geq t_2) = \frac{P(T_1 = t_1, T_2 = t_2)}{P(T_1 = t_1, T_2 \geq t_2)} \tag{3.10}$$

Then,

$$P(T_1 = t_1, T_2 = t_2) = \lambda_3(t_2|t_1) \cdot P(T_1 = t_1, T_2 \geq t_2) \tag{3.11}$$

Similarly,

$$P(T_1 \geq t_1, T_2 = t_2) = \lambda_2(t_2) \cdot P(T_1 \geq t_1, T_2 \geq t_2) \tag{3.12}$$

Since $(T_1 \geq t_1, T_2 \geq t_2)$ can only happen if both events are censored, thus it is equivalent to $(T_1 \geq t_1, T_2 \geq t_1)$,

$$P(T_1 \geq t_1, T_2 \geq t_2) = P(T_1 \geq t_1, T_2 \geq t_1) = e^{-\Lambda_1(t_1)-\Lambda_2(t_1)} \tag{3.13}$$

$$P(T_1 = t_1, T_2 \geq t_2) = \lambda_1(t_1) \cdot e^{-\Lambda_1(t_1)-\Lambda_2(t_2)-\Lambda_3(t_2|t_1)+\Lambda_3(t_1|t_1)} \tag{3.14}$$

With some simplification and the assumption of conditional independence of the hazard function given $\mathbf{b}$, the likelihood can be written as:

$$L_n(\theta) = \int \prod_j \lambda_1(t_{1j})^{\delta_{1j}} \cdot S_1(t_{1j}) \cdot \lambda_2(t_{1j})^{(1-\delta_{1j})\delta_{2j}} \tag{3.15}$$
$$\cdot S_2(t_{1j}) \cdot [\lambda_3(t_{2j})^{\delta_{2j}} \cdot \frac{S_3(t_{2j})}{S_3(t_{1j})}]^{\delta_{1j}} \cdot f(\mathbf{b}_j) \cdot d\mathbf{b}_j$$

where $\theta = (\beta_1, \beta_2, \beta_3)'$ is the parameter of interest.

By the definition of survival function and hazard function, the likelihood can be written as

$$
\begin{aligned}
& L\{\underline{\boldsymbol{\lambda_0}}, \theta, \nu\} \qquad (3.16)\\
&= \int \prod_j \lambda_1(t_{1j}|X_j, \mathbf{z}_j, \mathbf{b})^{\delta_{1j}} \cdot exp\{-\Lambda_1(t_{1j}|X_j, \mathbf{z}_j, \mathbf{b})\} \\
&\quad \cdot \lambda_2(t_{2j}|X_j, \mathbf{z}_j, \mathbf{b})^{(1-\delta_{1j})\delta_{2j}} \cdot exp\{-\Lambda_2(t_{2j}|X_j, \mathbf{z}_j, \mathbf{b})\} \\
&\quad \cdot \lambda_3(t_{2j}|X_j, \mathbf{z}_j, \mathbf{b})^{\delta_{1j}\cdot\delta_{2j}} \cdot exp\{-\delta_{1j}[\Lambda_3(t_{2j}|X_j, \mathbf{z}_j, \mathbf{b}) - \Lambda_3(t_{1j}|X_j, \mathbf{z}_j, \mathbf{b})]\} \\
&\quad \cdot f(\mathbf{b}, D(\nu))d\mathbf{b}
\end{aligned}
$$

where $\underline{\boldsymbol{\lambda_0}} = (\lambda_{01}(t_1), \lambda_{02}(t_2), \lambda_{03}(t_3))^T$, $\Lambda_i(t) = \int_0^t \lambda_i(u)du$, and $\underline{\boldsymbol{\Lambda_0}} = (\Lambda_{01}(t_1), \Lambda_{02}(t_2), \Lambda_{03}(t_3))^T$.The likelihood in (3.15) can also be consider a multi-state model.

With the proportional hazards assumptions and the use of counting process notations, the corresponding likelihood can be rewritten as:

$$
\prod_{j=1}^{n}\prod_{k=1}^{3}\{\prod_{t\geq 0}\lambda_{kj}(t|\mathbf{z}, \mathbf{b})^{dN_{kj}(t)} \cdot exp[-\int_{t=0}^{\infty} R_{kj}(t)\cdot\lambda_{kj}(t|\mathbf{z}, \mathbf{b})dt]\} \qquad (3.17)
$$

According to the definition in (3.4, 3.5, 3.6), the likelihood in (3.17), can be formed up as:

$$L = \int \prod_j [\lambda_{01}(t_{1j}) \cdot exp(x_j \cdot \beta_1 + \mathbf{z}'_{1j} \cdot \mathbf{b}_j)]^{\delta_{1j}} \cdot exp[-exp(x_j \cdot \beta_1 + \mathbf{z}' \cdot \mathbf{b}) \cdot \Lambda_{01}(t_{1j})] \tag{3.18}$$

$$\cdot [\lambda_{02}(t_{2j}) \cdot exp(x_j \cdot \beta_2 + \mathbf{z}'_{2j} \cdot \mathbf{b}_j)]^{(1-\delta_{1j}) \cdot \delta_{2j}} exp[-exp(x_j \cdot \beta_2 + \mathbf{z}'_{2j} \cdot \mathbf{b}_j) \cdot \Lambda_{02}(t_{2j})]$$

$$\cdot [\lambda_{03}(t_{2j}) \cdot exp(x_j \cdot \beta_3 + \mathbf{z}'_{3j} \cdot \mathbf{b}_j)]^{\delta_{1j} \cdot \delta_{2j}} \cdot exp[-\delta_{1j} exp(x_j \cdot \beta_3 + \mathbf{z}'_{3j} \cdot \mathbf{b}_j) \cdot \Lambda_{03}(t_{2j})] \cdot f(\mathbf{b}) d\mathbf{b}$$

3.5 Estimation Approaches

In this section,we introduced two estimation approaches to solve the parameter estimation in model (3.4,3.5,3.6) and likelihood (3.15).

3.5.1 Two Stage Semiparametric Pseudo Likelihood Approach

The motivation for a two stage semiparametric pseudo likelihood was from to estimate of baseline hazard, $\underline{\boldsymbol{\lambda_0}}$. In Zeng and Lin (2007) and Han et al. (2014)'s work, the baseline hazard $\underline{\boldsymbol{\lambda_0}}$ has been treated as discrete function, or $\underline{\boldsymbol{\Lambda_0}}$ as a step function, with increments or jumps occurring at the corresponding observed distinct failure time points. But a limitation of piecewise baseline hazard or cumulative hazard estimator was that it will introduce more parameters into the estimation process, which will be computational expensive and require the sample size to be substantial.

In order to ease the computational burden induced by baseline hazard and to focus on parameters of interest, we propose two stage pseudo-likelihood estimation approach. The two stage estimation methodology has been widely used in bivariate and multivariate data analysis. The advantage of the two stage approach was

computational efficiency. In order to avoid solving the joint likelihood for all parameters simultaneously, the two stage method decomposes the estimation into two steps, which will reduce the computation cost. But the two stage method does have it limitations. Since the two stage method was no longer using the joint likelihood, thus, the estimator was no longer maximum likelihood estimator(MLE), so it can't inherit the asymptotic property of MLE. Thus, computing standard error could be challenging.

In our case, the baseline hazards are estimated by the nonparametric Nelson-Aalen estimates in the first stage denoted as $\underline{\widehat{\mathbf{\Lambda_0}}}$.

$$\widehat{\Lambda_{0i}(t_{ij})} = \int_0^{t_{ij}} \frac{N_{ij}(u)}{R_{ij}(u)} du, \quad i = 1, 2, 3;\ j = 1, \cdots, n. \tag{3.19}$$

In the second stage, $\beta_1, \beta_2, \beta_3$ are estimated by maximizing the pseudo-likelihood function with estimates from the first stage into equation (3.15), as the following:

$$\begin{aligned} L^{Pseudo}(\theta|\underline{\widehat{\mathbf{\Lambda_0}}}) \propto \int \prod_i & [exp(\mathbf{z}_{1i} \cdot \beta_1 + \mathbf{b}_i)]^{\delta_{1i}} exp(-\widehat{\Lambda_{01}(t_{1i})} exp(\mathbf{z}_{1i} \cdot \beta_1 + \mathbf{b}_i)) \\ & \cdot [exp(\mathbf{z}_{2i} \cdot \beta_1 + \mathbf{b}_i)]^{(1-\delta_{1i})\delta_{2i}} exp(-\widehat{\Lambda_{02}(t_{1i})} exp(\mathbf{z}_{2i} \cdot \beta_2 + \mathbf{b}_i)) \\ & \cdot \{[exp(\mathbf{z}_{3i} \cdot \beta_3 + \mathbf{b}_i)]^{\delta_{2i}} exp(-[\widehat{\Lambda_{03}(t_{2i})} - \widehat{\Lambda_{03}(t_{1i})}] exp(\mathbf{z}_{2i} \cdot \beta_3 + \mathbf{b}_i))\}^{\delta_{1i}} \\ & \cdot f(\mathbf{b}_i) \cdot d\mathbf{b}_i \end{aligned} \tag{3.20}$$

Let $l_\theta^{Pseudo}(\theta|\underline{\widehat{\mathbf{\Lambda_0}}})$ and $U_\theta(\theta|\underline{\widehat{\mathbf{\Lambda_0}}})$ the score function of θ which is the derivative of the log of the likelihood in (3.20), the two stage pseudo estimator $\hat{\theta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)^T$

is the solution to the estimating equation:

$$U_\theta(\theta|\widehat{\underline{\mathbf{\Lambda_0}}}) = \sum_j \frac{\partial l_\theta^{Pseudo}(\theta|\widehat{\underline{\mathbf{\Lambda_0}}})}{\partial\theta} \tag{3.21}$$

Standard Error for the Two Stage Estimator

The variance of the Nelson-Aalen estimator is estimated by

$$\hat{\sigma}_i^2(t_{ij}) = \int_0^{t_{ij}} \frac{(R_{ij}(u) - N_{ij}(u))N_{ij}(u)}{(R_{ij}(u)-1)R_{ij}(u)^2}du$$

Expanding the score function $U_\theta(\theta|\widehat{\underline{\mathbf{\Lambda_0}}})$ in a Taylor series around θ_0 and evaluating it at $\theta = \hat{\theta}$, we get

$$\begin{aligned} U_\theta(\hat{\theta}|\widehat{\underline{\mathbf{\Lambda_0}}}) &= 0 \\ &= U_\theta(\theta_0|\widehat{\underline{\mathbf{\Lambda_0}}}) \\ &\quad + (\hat{\theta} - \theta_0)\sum_j V_{\theta_0}(\theta|\widehat{\underline{\mathbf{\Lambda_0}}}) + o_p(n^{1/2}) \end{aligned} \tag{3.22}$$

Where,

$$V_\theta(\theta|\widehat{\underline{\mathbf{\Lambda_0}}}) = \sum_j \frac{\partial^2 l_\theta^{Pseudo}(\theta|\widehat{\underline{\mathbf{\Lambda_0}}})}{\partial\theta^2} \tag{3.23}$$

Since the first stage marginal estimation is embedded with in the pesudo likelihood (3.20), the second stage model contains variables constructed from parameters estimated in the first stage. However, the covariance matrix of the second stage estimator includes noise induced by the first-stage estimates.To correct the standard error from the first stage estimation, we followed Karaca-Mandic and Train (2003)'s deduction. The final standard error estimator should be a sandwich estimator plus a

correction term. But since this is not a trivial extension, we recommend the use of bootstrap method to obtain standard error estimates. The method is demonstrated as follows:

1. Draw bootstrap samples;
2. Run first-stage Nelson Aalen Estimates in (3.19).
3. Maximize the two stage pseudo partial likelihood in (3.20) and obtain parameter estimates $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$
4. Repeat 1–3. Note that the two stage approach need to be performed on the same bootstrap samples; and
5. Compute the standard errors from the sampling distribution of the estimates.

3.5.2 Penalized Partial Likelihood Estimation

We restrict $\mathbf{b}$ to follow a multivariate normal distribution. Under the proportional mean model and the general additive frailty model setup, the likelihood for observed data can be written as:

$$
\begin{aligned}
& L\{\underline{\boldsymbol{\lambda_0}}, \theta, \nu\} \qquad (3.24)\\
&= \frac{1}{D(\nu)^{1/2}} \int \prod_i [\lambda_{01}(t_{1i}) \cdot exp(X_i^T \cdot \beta_1 + \mathbf{z}_i^T \cdot \mathbf{b})]^{\delta_{1i}} \cdot exp\{-\Lambda_{01}(t_{1i}) \cdot e^{X_i^T \cdot \beta_1 + \mathbf{z}_i^T \cdot \mathbf{b}}\} \\
&\quad \cdot [\lambda_{02}(t_{2i}) \cdot exp(X_i^T \cdot \beta_2 + \mathbf{z}_i^T \cdot \mathbf{b})]^{(1-\delta_{1i}) \cdot \delta_{2i}} \cdot exp\{-\Lambda_{02}(t_{2i}) \cdot e^{X_i^T \cdot \beta_2 + \mathbf{z}_i^T \cdot \mathbf{b}}\} \\
&\quad \cdot [\lambda_{03}(t_{2i}) \cdot exp(X_i^T \cdot \beta_3 + \mathbf{z}_i^T \cdot \mathbf{b})]^{(1-\delta_{1i}) \cdot \delta_{2i}} \cdot exp\{-\delta_{1i}[\Lambda_{03}(t_{2i}) - \Lambda_{03}(t_{1i})] \cdot e^{X_i^T \cdot \beta_3 + \mathbf{z}_i^T \cdot \mathbf{b}}\} \\
&\quad \cdot e^{-\frac{1}{2}\mathbf{b}^T D(\nu)^{-1} \mathbf{b}} d\mathbf{b}
\end{aligned}
$$

Since the integrated log likelihood (3.24) does not have a closed form expression, following Breslow and Clayton (1993) and Ripatti and Palmgren (2000), we write

(3.24) as $\int exp\{-S(\mathbf{b})\}d\mathbf{b}$, and apply the Laplace approximation to (3.24). Even though we use multivariate normal distribution in this paper, the derived likelihood approximations can be easily adapted to other frailty distributions as well. Profiling out the baseline hazard in similar way to Appendix B of Ripatti and Palmgren (2000), one can show that, given ν, the parametric regression coefficient θ can be obtained by jointly maximizing following penalized partial likelihood (PPL) with respect to $\theta, \nu, \mathbf{b}$. We follow Breslow and Clayton (1993) in their approximation for the generalized linear mixed model. Laplace's method for integral approximation allows the marginal log likelihood to be approximated by

$$l(\theta, \mathbf{b}, \nu) = \log(L(\theta, \mathbf{b}, \nu)) \approx -\frac{1}{2}\log|D(\nu)| - \frac{1}{2}\log|K''(\tilde{\mathbf{b}})| - K(\tilde{\mathbf{b}}) \tag{3.25}$$

where

$$\begin{aligned} K(\tilde{\mathbf{b}}) = -[\sum_{i=1}^{n} & \delta_{1i}[\log(\lambda_{01}(t)) + X_j^T\beta_1 + \mathbf{z}_i\tilde{\mathbf{b}}] - \Lambda_{01}(t)\exp(X_i^T\beta_1 + \mathbf{z}_i\tilde{\mathbf{b}}) \\ & + (1-\delta_{1i})\delta_{2i}[\log(\lambda_{02}(t)) + X_j^T\beta_2 + \mathbf{z}_i\tilde{\mathbf{b}}] - \Lambda_{02}(t)\exp(X_i^T\beta_2 + \mathbf{z}_i\tilde{\mathbf{b}})] \\ & + \delta_{1i}\delta_{2i}[\log(\lambda_{03}(t)) + X_j^T\beta_3 + \mathbf{z}_i\tilde{\mathbf{b}}] - \Lambda_{03}(t)\exp(X_i^T\beta_3 + \mathbf{z}_i\tilde{\mathbf{b}}) \\ & - \frac{1}{2}\tilde{\mathbf{b}}'D(\nu)^{-1}\tilde{\mathbf{b}}] \end{aligned} \tag{3.26}$$

The set of second partial derivatives of $K(\mathbf{b})$ with respect to $\mathbf{b}$ is denoted $K''(\mathbf{b})$ and has the form

$$K''(\tilde{\mathbf{b}}) = \sum_{i=1}^{n} \Lambda_{01}(t)\exp(X_i^T\beta_1 + \mathbf{z}_i^T\tilde{\mathbf{b}})\mathbf{z}_i\mathbf{z}_i' + \Lambda_{02}(t)\exp(X_i^T\beta_2 + \mathbf{z}_i^T\tilde{\mathbf{b}})\mathbf{z}_i\mathbf{z}_i' \qquad (3.27)$$
$$+ \Lambda_{03}(t)\exp(X_i^T\beta_3 + \mathbf{z}_i^T\tilde{\mathbf{b}})\mathbf{z}_i\mathbf{z}_i' + D(\nu)^{-1}$$

This leads to the approximate marginal log likelihood

$$l(\theta, \mathbf{b}; \nu) \approx \sum_{i=1}^{n}\{\delta_{1i}[(X_i^T \cdot \beta_1 + \mathbf{z}_i^T \cdot \mathbf{b}) - \log \sum_{j\in R(t_{1i})} e^{X_j^T\cdot\beta_1 + \mathbf{z}_j^T\mathbf{b}}] \qquad (3.28)$$
$$+ (1-\delta_{1i})\delta_{2i}[(X_i^T \cdot \beta_2 + \mathbf{z}_i^T \cdot \mathbf{b}) - \log \sum_{j\in R(t_{2i}|\delta_{1i}=0)} e^{X_j^T\cdot\beta_2 + \mathbf{z}_j^T\mathbf{b}}]$$
$$+ \delta_{1i}\delta_{2i}[(X_i^T \cdot \beta_3 + \mathbf{z}_i^T \cdot \mathbf{b}) - \log \sum_{j\in R(t_{2i}|\delta_{1i}=1)} e^{X_j^T\cdot\beta_3 + \mathbf{z}_j^T\mathbf{b}}]\} - \frac{1}{2}\mathbf{b}^T D(\nu)^{-1}\mathbf{b}$$

If both ν were known and $\mathbf{b}$ were considered a fixed effects parameter, then (3.28) would be a penalized log likelihood (Green (1987)), where $-\frac{1}{2}\mathbf{b}^T D(\nu)^{-1}\mathbf{b}$ is the penalty term penalizing for extreme values of $\mathbf{b}$.

For given ν, the estimating equations based on the first partial derivatives of the PPL are following, for $\theta = (\beta_1, \beta_2, \beta_3)$,

$$U_1 = \sum_{i=1}^{n} \delta_{1i}[X_i - \frac{X_i \cdot exp(X_i^T \cdot \beta_1 + \mathbf{z}_i^T \cdot \mathbf{b})}{\sum_{j\in R(t_{1i})} exp(X_j^T \cdot \beta_1 + \mathbf{z}_j^T \cdot \mathbf{b})}] = 0 \qquad (3.29)$$
$$U_2 = \sum_{i=1}^{n} (1-\delta_{1i})\delta_{2i}[X_i - \frac{X_i \cdot exp(X_i^T \cdot \beta_2 + \mathbf{z}_i^T \cdot \mathbf{b})}{\sum_{j\in R(t_{2i}|\delta_{1i}=0)} exp(X_j^T \cdot \beta_2 + \mathbf{z}_j^T \cdot \mathbf{b})}] = 0$$
$$U_3 = \sum_{i=1}^{n} \delta_{1i}\delta_{2i}[X_i - \frac{X_i \cdot exp(X_i^T \cdot \beta_3 + \mathbf{z}_i^T \cdot \mathbf{b})}{\sum_{j\in R(t_{2i}|\delta_{1i}=1)} exp(X_j^T \cdot \beta_3 + \mathbf{z}_j^T \cdot \mathbf{b})}] = 0$$

For $\mathbf{b}$,

$$
\begin{aligned}
U_{\mathbf{b}} = & \sum_{i=1}^{n} \delta_{1i}[\mathbf{z}_i - \frac{\mathbf{z}_i \cdot exp(X_i^T \cdot \beta_1 + \mathbf{z}_i^T \cdot \mathbf{b})}{\sum_{j \in R(t_{1i})} exp(X_j^T \cdot \beta_1 + \mathbf{z}_j^T \cdot \mathbf{b})}] \\
& + (1 - \delta_{1i})\delta_{2i}[\mathbf{z}_i - \frac{\mathbf{z}_i \cdot exp(X_i^T \cdot \beta_2 + \mathbf{z}_i^T \cdot \mathbf{b})}{\sum_{j \in R(t_{2i}|\delta_{1i}=0)} exp(X_j^T \cdot \beta_2 + \mathbf{z}_j^T \cdot \mathbf{b})}] \\
& + \sum_{i=1}^{n} \delta_{1i}\delta_{2i}[\mathbf{z}_i - \frac{\mathbf{z}_i \cdot exp(X_i^T \cdot \beta_3 + \mathbf{z}_i^T \cdot \mathbf{b})}{\sum_{j \in R(t_{2i}|\delta_{1i}=1)} exp(X_j^T \cdot \beta_3 + \mathbf{z}_j^T \cdot \mathbf{b})}] - D(\nu)^{-1}\mathbf{b} = 0
\end{aligned} \tag{3.30}
$$

The estimated standard error can be approximated by the inverse of the minus second partial derivative matrix. The maximization of the approximate likelihood (3.28) can be done using the Newton-Raphson technique.

Variance Component Estimation via Penalized Partial Likelihood

We assuming that the variance component ν are known. In practice, it need to be estimated from the data. If we assign the maximized value $(\hat{\theta}(\nu), \hat{\mathbf{b}}(\nu))$ of the PPL into (3.28), we get an approximate profile likelihood function for ν,

$$
\hat{l}(\hat{\theta}(\nu), \nu) \approx -\frac{1}{2}\log|D(\nu)| - \frac{1}{2}\log|K''(\hat{\mathbf{b}})| - \frac{1}{2}\hat{\mathbf{b}}' D(\nu)^{-1}\hat{\mathbf{b}} \tag{3.31}
$$

where $K''(\hat{\mathbf{b}})$ is derived in (3.27), given $\theta = \hat{\theta}; \mathbf{b} = \hat{\mathbf{b}}$. Follow Ripatti and Palmgren (2000)'s variance estimation procedure, we also use $K''_{PPL}(\hat{\mathbf{b}}) = \frac{\partial^2 PPL}{\partial \mathbf{b} \partial \mathbf{b}'}$ instead of $K''(\hat{\mathbf{b}})$, after differentiation and some simplification, en estimating equation for ν is

$$
U_\nu = \frac{1}{2}[tr(D^{-1}\frac{\partial D}{\partial \nu}) + tr(K''_{PPL}(\hat{\mathbf{b}})^{-1}\frac{\partial D^{-1}}{\partial \nu}) - \hat{\mathbf{b}}' D^{-1}\frac{\partial D}{\partial \nu}D^{-1}\hat{\mathbf{b}}] = 0 \tag{3.32}
$$

The corresponding fisher information matrix, derived by differentiating (3.31) twice and taking the expectation with respect to $\boldsymbol{\beta}$, is

$$\begin{aligned} J =& \frac{1}{2}[tr(D^{-1}\frac{\partial D}{\partial \nu}D^{-1}\frac{\partial D}{\partial \nu} + D^{-1}\frac{\partial^2 D}{\partial \nu \partial \nu'}) \\ &+ tr(K''_{PPL}(\hat{\mathbf{b}})^{-1}\frac{\partial D^{-1}}{\partial \nu}K''_{PPL}(\hat{\mathbf{b}})^{-1}\frac{\partial D^{-1}}{\partial \nu} - K''_{PPL}(\hat{\mathbf{b}})^{-1}\frac{\partial^2 D^{-1}}{\partial \nu \partial \nu'})] \end{aligned} \tag{3.33}$$

3.6 Simulations

In order to illustrate our methodology, we conducted the following simulation study. The performance of the likelihood approximation is evaluated in the Two sets of simulations: (i) a shared Gaussian frailty model with varying frailty variance; (ii) a model misspecification scenario : the frailty term for simulated data follows log-normal distribution but estimation is conducted using normal frailty. We compared the performance of frailty model approach with the Cox model approach by treating each event as independent event.

3.6.1 Simulation Setup

We generated data according to model (3.1, 3.2, 3.3) with the Weibull baseline hazard function $h(t) = \lambda p t^{p-1}$. Specifically, we choose $\lambda = 1, p = 1$ in our simulation. A fixed covariate $Z \sim Uniform(0, 2)$ was used to all three events, with corresponding coefficient $\beta_i = 1, i = 1, 2, 3$. Random effect were incorporated using $b \sim N(0, \sigma_b^2)$ and $\log b \sim N(0, \sigma_b^2)$.

Denote the observed event time for illness and death as X_{1i} and X_{2i}, respectively. The generation of semicompeting risks data based on illness-death models consists of two steps.

1. In the first step, we simulate $S_{competing}(t) = U_{1i} \sim U(0,1)$, $S_{competing}(t)$ is the joint survival for event 1 and event 2 as competing risk, $S_{competing}(t) = \exp\{-\Lambda_1(t) - \Lambda_2(t)\}$, where Λ_1 and Λ_2 denote the cumulative hazards for illness and death without illness, receptively. T_{1i}^* is the solution for $u_{1i} = S_{competing}(t)$. In this stage, survival times are generated for either illness or death without illness. This is the competing risk stage of semi-competing risk data.
2. In the second stage, we use a Bernoulli experiment to decide which event is assigned at this time T_{1i}^*. we generate another random number $U_{2i} \sim U(0,1)$. If $u_{2i} > \frac{\lambda_1(t_i)}{\lambda_1(t_i)+\lambda_2(t_i)}$, T_{1i}^* is considered as death time. Otherwise, T_{1i}^* is illness time and death time can be generated based on the following conditional probability:

$$S(T_{2i} = t_i + s_i | T_{1i} = t_i) = \exp\{-\Lambda_3(t_i + s_i) - \Lambda_3(t_i)\}$$

where s_i is the additional survival time after illness and Λ_3 is the cumulative hazard for death after illness.

3.6.2 Simulation Results

Data for 1000 replications are generated with a total of $n = 600$ observations for each replication. On average, from each simulated dataset, we observed 283 T_1 events, 285 T_2 events without the precedence of T_1, and 265 T_2 events with the precedence of T_1, respectively. The analyses were conducted using the Cox models, the two stage pseudo likelihood model and the penalized partial likelihood(PPL).

The results are summarized in Table (3.1). The average of estimates, the model based standard error estimates (M.SE) the average values of the estimated standard errors (E.SE), and coverage probabilities (CP) of the 95% intervals based on model based standard error estimates. We can see that all three methods perform well for regression parameters when there was small within subject variance, $\sigma_b^2 = 0.1$. However, as the within subject variance increases, the naive Cox proportional hazard estimation approach, which ignores the within subject correlation, provided very bias estimates. Compare to the naive Cox model, the two stage pseudo likelihood estimation approach provided much more accurate estimates, but the standard error inflated as the variance of frailty increased. The penalized partial likelihood estimation approach was more accurate and more robust in all three estimates. Table (3.2) summarized the simulation results when frailty term was generated from log-normal distribution, but we still assume the frailty term as normal distribution in the estimation. As we can see from the results, the estimates show similar pattern as in Table (3.1). The penalized partial likelihood estimation approach was more accurate and more robust in all three estimates when the model is misspecified.

Figure(3.2) and Figure (3.3) summarized the simulation result in boxplots, where the length of box represented the standard error of the estimates, and in the histogram provided the empirical distribution of the estimates. The red dot line in Figure(3.2) represented the true parameter values from simulation; The very left red box represented Cox naive estimation approach; The middle green box represented penalized partial likelihood estimation approach; The right blue box represented two stage pseudo likelihood estimation approach. As we can observe that, as the variance of frailty increase, the naive Cox estimation approach differs more from the true

value; the two stage pseudo likelihood model performed better than the naive Cox estimation approach, but the result was not stable; the penalized partial likelihood model performed the best and provided the most robust result.

Table 3.1: Results For Comparing Three Estimation Approaches: Based on Normal Frailty Scenario and The True Parameters are $\beta_1 = 1$, $\beta_2 = 1$ and $\beta_3 = 1$, The Data were Simulated from Exponential Distribution

	Cox(Independent)				Two stage Pseudo						Penalized Partial					
	E.ST	M.SE	E.SE	M.CP	E.ST	M.SE	E.SE	M.CP	B.SE	B.CP	E.ST	M.SE	E.SE	M.CP	B.SE	B.CP
								σ_b^2=0.1								
β_1	1.053	0.321	0.378	0.93	1.148	0.239	0.385	0.75	0.387	0.95	1.06	0.376	0.38	0.94	0.381	0.95
β_2	1.006	0.327	0.338	0.94	1.052	0.237	0.331	0.72	0.326	0.94	1.017	0.347	0.34	0.95	0.349	0.96
β_3	0.989	0.359	0.375	0.92	1.05	0.244	0.332	0.76	0.338	0.95	0.987	0.324	0.328	0.93	0.329	0.94
								σ_b^2=0.5								
β_1	0.9	0.31	0.342	0.9	1.024	0.234	0.354	0.75	0.351	0.94	1.034	0.371	0.377	0.92	0.378	0.93
β_2	0.865	0.317	0.335	0.96	0.945	0.272	0.333	0.73	0.337	0.95	1.041	0.379	0.377	0.94	0.374	0.92
β_3	0.732	0.335	0.339	0.84	0.954	0.246	0.346	0.75	0.352	0.95	0.983	0.389	0.395	0.93	0.397	0.95
								σ_b^2=1								
β_1	0.697	0.3	0.328	0.78	0.86	0.241	0.341	0.74	0.354	0.96	1.021	0.328	0.335	0.94	0.337	0.95
β_2	0.679	0.309	0.337	0.77	0.779	0.253	0.335	0.73	0.334	0.95	1.067	0.346	0.351	0.94	0.358	0.95
β_3	0.424	0.321	0.312	0.56	0.792	0.275	0.353	0.73	0.348	0.93	0.957	0.352	0.343	0.96	0.349	0.96

E.ST: Average of 1000 runs estimates compare to true parameter $\beta_i = 1$, $i = 1, 2, 3$.

M.SE: Model based standard error.

E.SE: Empirical standard error.

M.CP: 95 % coverage probability based on M.SE.

B.SE: Bootstrap standard error.

B.CP: 95 % coverage probability based on bootstrap.

Table 3.2: Results For Comparing Three Estimation Approaches: Based on Log-Normal Frailty Scenario and The True Parameters are $\beta_1 = 1$, $\beta_2 = 1$ and $\beta_3 = 1$, The Data were Simulated from Exponential Distribution

	Cox(Independent)				Two stage Pseudo						Penalized Partial					
	E.ST	M.SE	E.SE	M.CP	E.ST	M.SE	E.SE	M.CP	B.SE	B.CP	E.ST	M.SE	E.SE	M.CP	B.SE	B.CP
								σ_b^2=0.1								
β_1	1.052	0.321	0.379	0.93	1.147	0.239	0.385	0.77	0.497	0.95	1.059	0.372	0.381	0.93	0.373	0.95
β_2	1.007	0.327	0.338	0.94	1.052	0.264	0.331	0.83	0.346	0.96	1.018	0.343	0.340	0.94	0.343	0.95
β_3	0.991	0.359	0.378	0.92	1.049	0.224	0.331	0.76	0.358	0.96	0.985	0.322	0.328	0.93	0.323	0.95
								σ_b^2=0.5								
β_1	0.907	0.312	0.337	0.91	1.026	0.226	0.351	0.73	0.371	0.96	1.031	0.372	0.370	0.96	0.376	0.95
β_2	0.871	0.319	0.320	0.94	0.953	0.214	0.328	0.72	0.337	0.95	1.039	0.323	0.367	0.92	0.327	0.95
β_3	0.732	0.340	0.346	0.84	0.957	0.246	0.347	0.76	0.366	0.93	0.975	0.392	0.395	0.95	0.397	0.95
								σ_b^2=1								
β_1	0.756	0.307	0.323	0.84	0.956	0.222	0.347	0.78	0.354	0.94	1.00	0.41	0.43	0.92	0.448	0.96
β_2	0.726	0.315	0.317	0.85	0.880	0.237	0.326	0.83	0.344	0.96	1.04	0.42	0.42	0.95	0.425	0.96
β_3	0.487	0.335	0.325	0.64	0.891	0.226	0.331	0.83	0.323	0.94	0.92	0.48	0.47	0.95	0.480	0.96

E.ST: Average of 1000 runs estimates compare to true parameter $\beta_i = 1$, $i = 1, 2, 3$.

M.SE: Model based standard error.

E.SE: Empirical standard error.

M.CP: 95 % coverage probability based on M.SE.

B.SE: Bootstrap standard error.

B.CP: 95 % coverage probability based on bootstrap.

Figure 3.2: Simulation Results for the Normal Frailty Scenario Presented in Box Vixen Plot and Empirical Distribution Plot.The red dash line represents the true parameter.

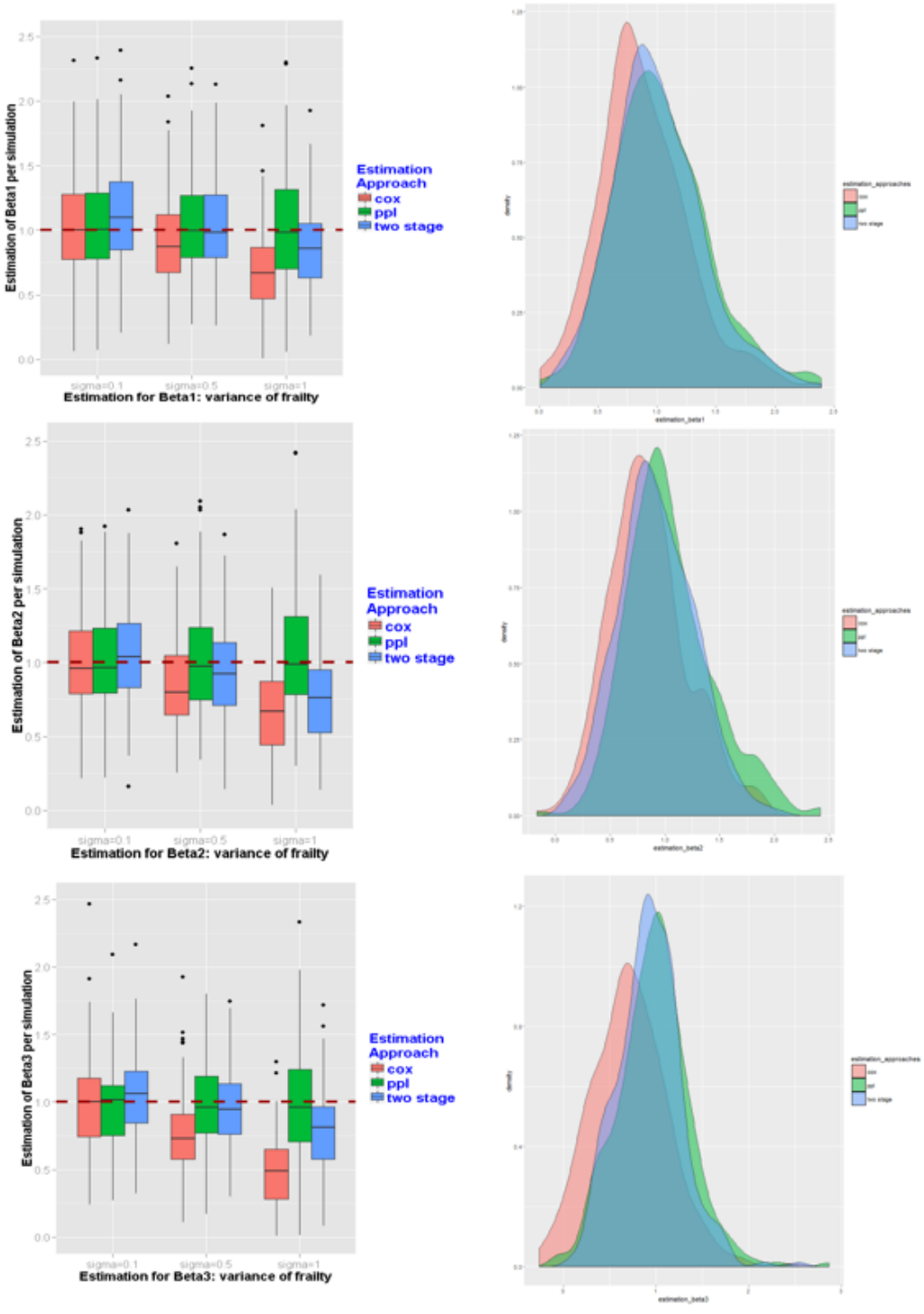
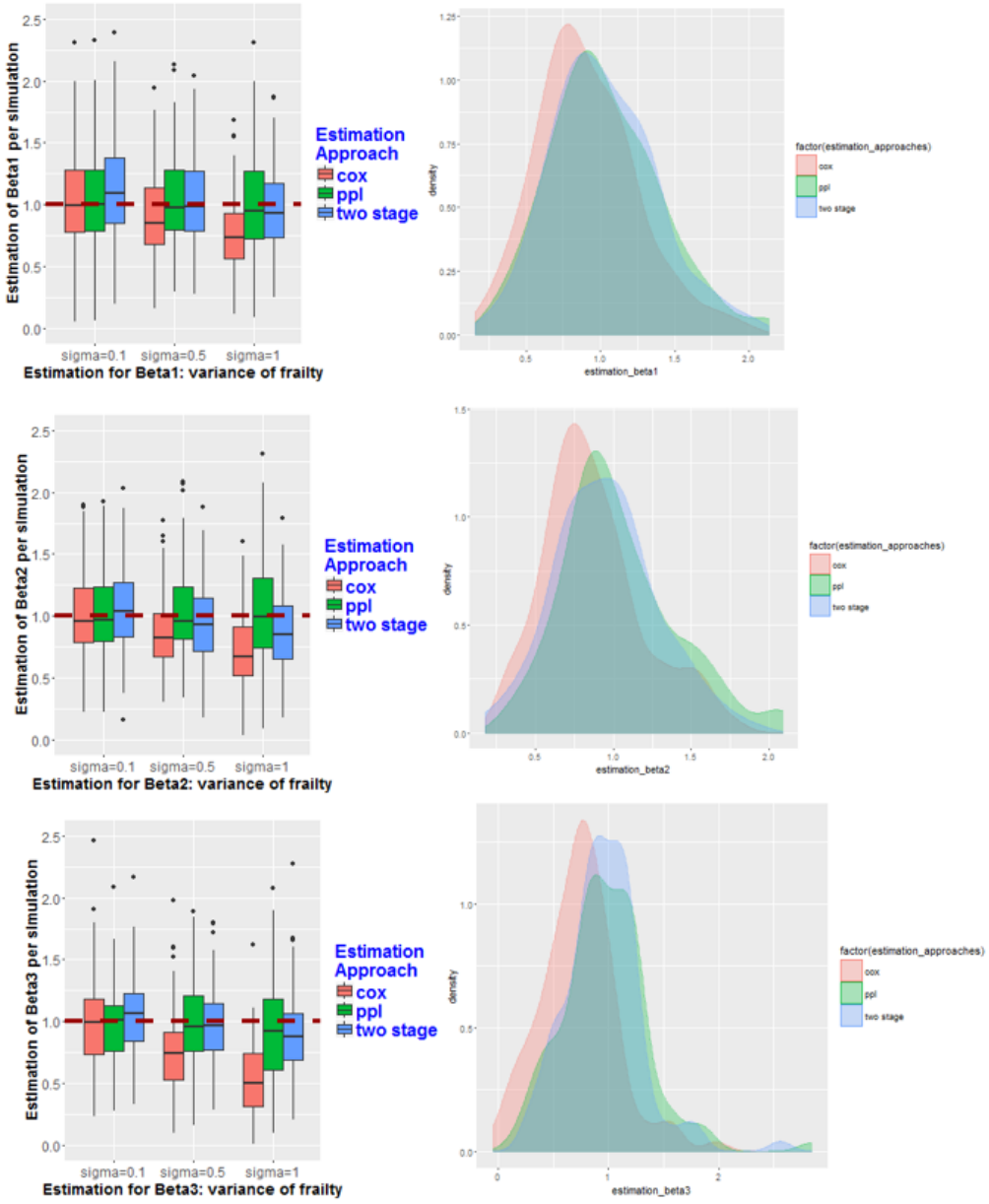


Figure 3.3: Simulation Results for the Log-Normal Frailty Scenario Presented in Box Vixen Plot and Empirical Distribution Plot.The red dash line represents the true parameter.



3.7 Data Application Example

In this section, we demonstrated the two proposed estimation approaches using electronic medical records data.

3.8 Indianapolis-Ibadan Dementia Project (IIDP) Cohort

Electronic medical records (EMR) capture enormous quantities of clinical data including medical diagnosis, laboratory testing, medication dispensing information and they have been increasingly used in many health systems around the country. The availability of EMR data offers an unprecedented research opportunity for monitoring disease development, progression and treatment.

We demonstrated our proposed method by using the electronic medical records from Indianapolis-Ibadan Dementia Project (IIDP) cohort. The Indianapolis-Ibadan Dementia Project is a longitudinal, prospective, community-based epidemiological comparative study of rates and risk factors for dementia and Alzheimer disease in elderly African Americans living in Indianapolis, Indiana and Yoruba in Ibadan, Nigeria. Since 1992, the IIDP has enrolled a cohort of African Americans aged 65 or older and followed the participants until 2011 with cognitive evaluation, clinical diagnosis and collection of risk factor information at regularly scheduled intervals every 2 to 3 years. This rich database was specifically designed to identify incident cases of Alzheimer's disease (AD), over a 20 year period (Gureje et al., 1995).These data include medical diagnoses, clinical findings, diagnostic testing, procedures, and medications. Electronic medical records data are available from all enrolled patients and the information includes diagnosis of medical conditions, laboratory test measures

and medications order and dispensing. The detail of data description can be found in Campbell et al. (2010).

Our interest is to compare the risk of CAD between patients in different gender groups. For patients with an incident CAD event, the date of diagnosis was used as the event time; otherwise, the last outpatient clinic visit prior to December 31, 2010 was used as the right censoring time.

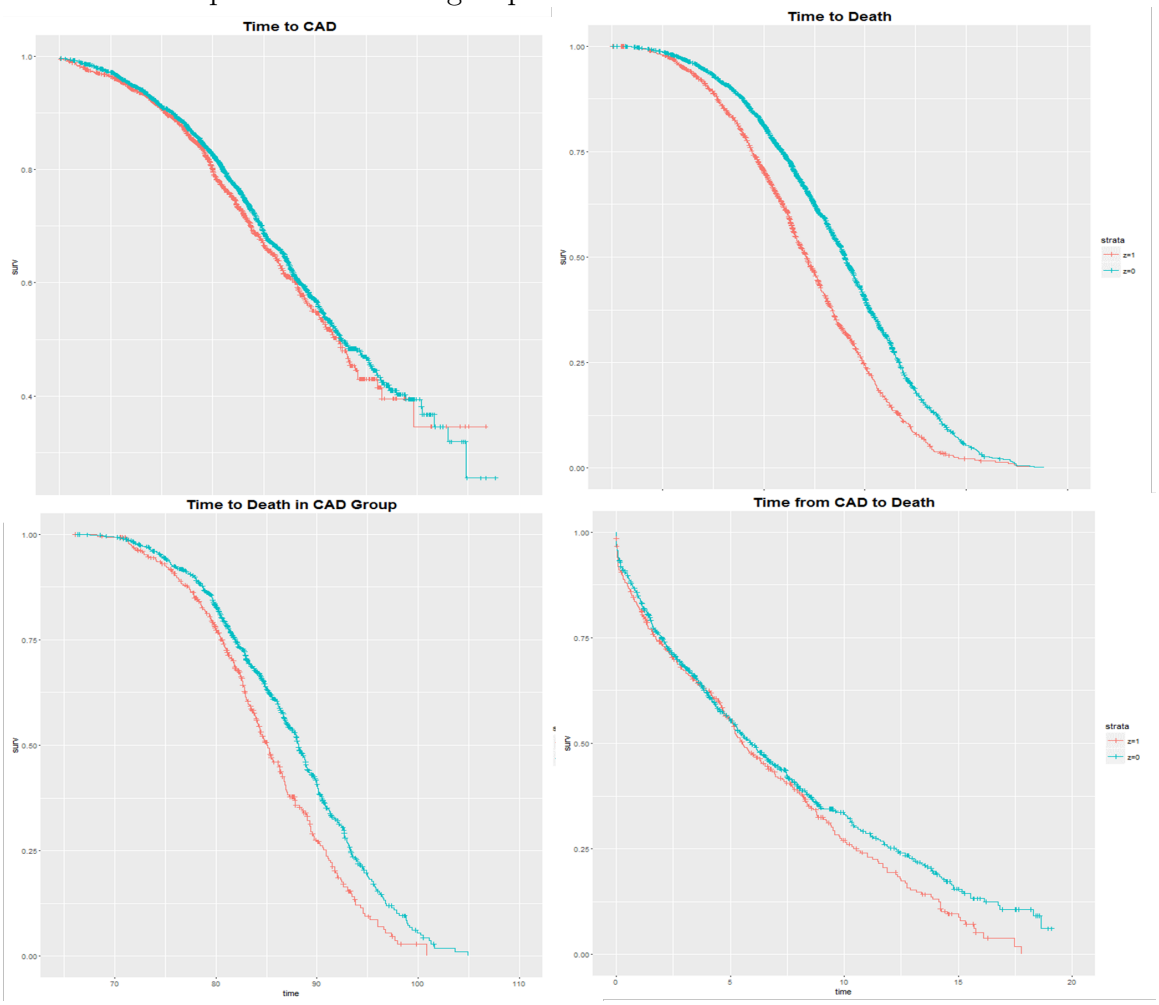
Table 3.3: Median Age at Events (Number of Cases, Incidence) By Gender

Event	Total (N=4105)	Female (N=2666)	Male(N=1439)
CAD	81.76(1280, 31.18%)	82.33(855, 32.07%)	80.65(425, 29.53%)
Death	82.82(2593, 63.17%)	83.34(1568, 58.81%)	82.01(1025, 71.23%)

Table (3.3) summarized the demographic information of the study cohort. In study population, 4105 subjects were free of CAD at baseline enrollment time, within the total cohort, we have 1208 cases of CAD events and 2593 deaths. In CAD cases, 32.07% were female cases and 29.53% were male cases. In death cases, 58.81% were female cases, 71.23% were male cases. The median age of event onset were earlier in the male group, while females experienced later event time.

Figure (3.4) presented the time to event plot by gender, the red line represented male group, the blue line represented female group. From Figure (3.4), we can observe that the female group had later onset than male group in CAD, Death and death conditional CAD.

Figure 3.4: Survival plots of time to CAD, time to death, time to death in the CAD group and time from CAD to death, by gender, the red line represented male group, the blue line represented female group



3.9 Event Specific Hazard and Model Setup

We followed models (3.4,3.5,3.6) and define t_1 as the time of CAD, t_2 as the time of death. The model have the following structure:

$$\lambda_{CAD}(t_1|\mathbf{z}_1,\mathbf{b}) = \lambda_{0,CAD}(t_1)\cdot exp(X_1^T\cdot\beta_1+\mathbf{z}_1^T\cdot\mathbf{b}), t_1>0 \tag{3.34}$$

$$\lambda_{DeathOnly}(t_2|\mathbf{z},\mathbf{b}) = \lambda_{0,DeathOnly}(t_2)\cdot exp(X_2^T\cdot\beta_2+\mathbf{z}_2^T\cdot\mathbf{b}), t_2>0 \tag{3.35}$$

$$\lambda_{DeathAfterCAD}(t_2|t_1,\mathbf{z},\mathbf{b}) = \lambda_{0,DeathAfterCAD}(t_2|t_1)\cdot exp(X_3^T\cdot\beta_3+\mathbf{z}_3^T\cdot\mathbf{b}), t_2>t_1>0 \tag{3.36}$$

The estimation results were shown in Table (3.4). The native Cox proportional hazard estimation approach, two stage pseudo likelihood estimation approach and penalized partial likelihood approaches were used in IIDP data analysis. The naive Cox estimation approach showed higher risk of CAD and death, also higher risk of death conditional on history of CAD experience in male group. However, the two stage pseudo likelihood approach and penalized partial likelihood approach showed lower risk of death conditional on history of CAD in the male group.

Table 3.4: Data Application Result:Estimation of Gender Effect in Time to CAD with Death as a Semi-competing Risk

Parameter	Cox(Naive Independent)	Two Stage Pseudo	Penalized Partial
β_1 **(Health to CAD)**	0.082^* (0.059)	0.033(0.121)	0.067(0.167)
β_2**(Health to Death)**	0.48^*(0.048)	0.597^*(0.126)	0.745^*(0.144)
β_3 **(CAD to Death)**	0.26(0.071)	-0.211^*(0.162)	-0.621^*(0.197)

3.10 Conclusion

We proposed two frailty-based semiparametric models for analyzing survival times with a semi-competing risk. Compared to independent Cox regression, the new model setup takes the correlation between time to event of interest and informative censoring

caused by semi-competing risk into consideration. Simulation studies demonstrate adequate performance for both the two-stage and the penalized partial likelihood methods. The later approach is more stable and robust. Our proposed method can be applied to many studies on aging and clinical trials where deaths of the participants may be related to disease outcomes.

Chapter 4

Frailty-based Multi-event Semiparametric Models for Failure Time Data with Semi-competing Risks

4.1 Abstract

In medical research, multi-event and multi-stage arises when a individual was at risk of multiple disease, or a certain disease progressed in several state. It is crucial to study the inner structure and dependence between multiple diseases or multiple states. In this paper, we propose to use frailty based semparametric model, whereas frailty models introduce random effects to account for unobserved risk factors, possibly shared by multiple diseases or multiple states. For the model estimation, we developed and evaluated three approaches: parametric, two stage semiparametric estimation and penalized partial likelihood approach. Simulation studies, performed by using an innovative method for generating dependent multi-state survival data, show that penalized partial likelihood methods are very competitive to evaluate the effect of covariates.

4.2 Introduction

Multi-event model is formulated that describes the pathway and linkage between the multiple events happened in the same subject; a special case of multi-event model is the multi-state model. Inherited from the feature of multiple event model which describes the dependence structure and development of multiple events. Multi-state

models are commonly used for describing the development or progression stage of failure time for a certain disease.

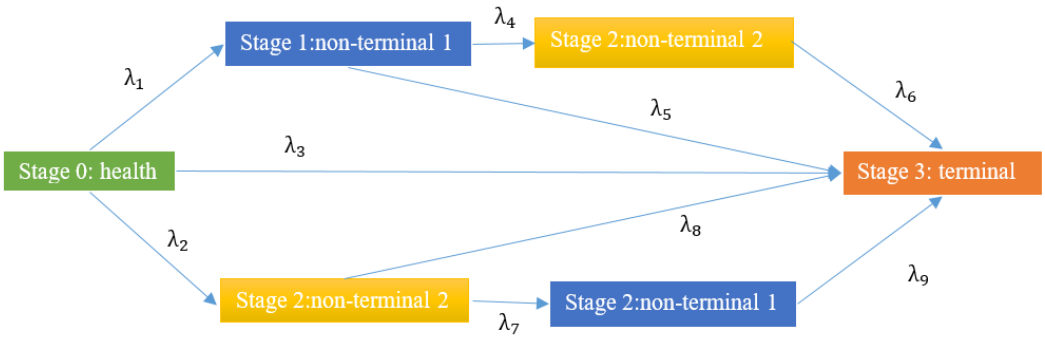
Both multi-event situation and multi-state cases existed widely in medical studies. In medicine, the multiple event situation arises when multiple diseases happened to the same subject; Multi-state can describe conditions like healthy, diseased, diseased with complication and dead. A change of one event to the other event, or a change of state is called a transition. This then corresponds to outbreak of disease, occurrence of complication and death. It important to recognize the difference between an event (like disease, death) and a state (like recurrence of tumor, dead). The multi-event and multi-state structure make it is possible to study the detail of medical history and progression of certain disease. The full statistical model specifies the multi-event and multi-state structure and the form of the hazard function (intensity function) for each possible transition.

In many epidemiological studies of the elderly population, it has been observed that individuals at risk of one chronic condition tend to have increased risk of other medical conditions with a substantial numbers having multiple chronic conditions. Studying the co-occurrence of these conditions may identify common biological pathways linking these disorders and ultimately lead to effective treatment and prevention strategies. Another complication facing the studies in aging is death due to other causes which can be indirectly related to the conditions under study through genetic or environmental exposures related to the individual's susceptibility to both disease and death.

In chapter 3, we focused on illness-death model (with just two states:illnesses and death), which represented in Figure (3.1). In this chapter, we focus on more

general situation, which is a multi-event model for bivariate failure times with a semi-competing risk, represented in Figure 4.1.

Figure 4.1: Two Non-terminal Events and a Terminal Event as Semi-competing Risk



The specific model technique we use to form the model is frailty model approach, whereas frailty models introduce random effects to account for unobserved risk factors, possibly shared by multiple diseases or multiple states. The integration of frailty and multi-event methodology was interesting to control for unobserved heterogeneity in the presence of complex event history structures, particularly appealing in observational study and clinical trials applications.

In the present chapter we propose the incorporation of shared frailties in the transition specific hazard function; then, we develop and evaluate parametric, two stage semiparametric estimation and penalized partial likelihood approaches. Simulation studies, performed by using an innovative method for generating dependent multi-state survival data, show that penalized partial likelihood methods are very competitive to evaluate the effect of covariates.

The following sections were presented in the following order. In section 2, we introduce notation and model setup; in section 3, we review and compare well

established current methodologies; in section 4, we introduce our likelihood and model structure; Section 5 is simulation study; Section 6 is real data application; Section 7 is conclusion and discussion.

4.3 Notation and Setup

Let C be an external censoring variable due to withdrawal of patients or the end of study. T_1 and T_2 were the time to the non-terminal events, for example, disease progression (refer to as illness hereafter), T_3 was the time to terminal event (refer to death hereafter). X_i, where $i = 1, 2, 3$, is the observation at time T_i, where $i = 1, 2, 3$.

$$\delta_1 = I(X_1 = T_1) = I(T_1 < C)$$

$$\delta_2 = I(X_2 = T_2) = I(T_2 < C)$$

$$\delta_3 = I(X_3 = T_3) = I(T_3 < C)$$

where $I(\cdot)$ is the indicator function. Note that T_3 can censor T_1 or T_2 but not vice visa, whereas C can censor T_1, T_2 and T_3. In addition, a vector of covariate x_i is observed. Furthermore, we assume that C is independent of joint distribution of T_1, T_2 and T_3 given Z. Let $\{(T_{1i}, T_{2i}, T_{3i}, C), i = 1, \ldots, n.\}$ be independent and identically distributed (IID) replications of (T_1, T_2, T_3, C). The observed data are IID replications of $(X_1, X_2, X_3, \delta_1, \delta_2, \delta_3)$.

In order to better illustrate the semi-competing risk procedure, we introduce nine more indicators:

$$\delta_4 = I(T_1 < T_2)$$

$$\delta_5 = I(T_2 < T_1)$$

where δ_4 represents if non-terminal event 1 happened before non-terminal event 2; δ_5 represents if non-terminal event 2 happened before non-terminal event 1.

4.4 Review of Current Multi-event and Multi-state Model

In this section, we review several commonly used multi-event and multi-state models. First, we introduce parametric and semiparametric frailty models; Then, we introduce Markov Models.

4.4.1 Parametric and Semiparametric Frailty Model

The inclusion of frailties into multi-state models can provide complex survival models accounting for dependence between grouped subjects as well as between times to events of different types within the same group, some works addressing this problem have begun to appear in recent years in applied statistics, while investigation of theoretical aspects is emerging. Bhattacharyya and Klein (2005), for instance, considered progressive multi-state models with exponential baselines. They introduced frailties correlated within subjects, obtained by summing independent gamma random variables as suggested by Yashin et al. (1995) for correlated frailty models. van Houwelingen and Putter (2011) critically discussed some aspects of more general

multi-state models with dependent frailties within subjects, again. These models account for association between transition intensities of the same subject, while they consider event times of different subjects as independent. So, these models are more in the spirit of univariate frailty models, each subject having a different risk level due to his own unobserved factors.

In such a context the larger the frailty variances, the higher the heterogeneity between subjects and the dependence between event times of different types for each subject. No clustering effect can be accounted for by this approach.

4.4.2 Multi-state Markov Model

One model structure was commonly used in multi-event and multi-state situation is Markov Model, which consider the progression of certain disease as a stochastic process. Review paper on Markov models can be found in Andersen et al. (1985); Cox and Miller (1965); Hoem et al. (1976); Hougaard (1999); Jackson (2016).

In multi-state Markov Model, the state structure is not unique. Choosing the best structure can make the model assumptions more transparent, and simplify some calculations. It is a clear advantage if the model is Markov, because this allows for an intuitive graphical understanding of the model. One property that can be seen in the state structure is whether the process is progressive. This is defined as each state having only a single possible transition into it, and the initial state having no entries. Thus the state at time t determines which states have been visited previously and in which order.

Compared to the frailty model approach, which was more focused on covariate effect, the multi-state Markov model is focused on the transition of one state

to the other state, the specific statistics in Markov model is called the transitional probability. The transitional probability evaluated at time v defined as

$$P_l(v,t) = Pr(X_t = l | X_u, u \in [0,v]) \tag{4.1}$$

Where, $X_t, t \in [0,\infty)$ is a stochastic process, which is a right continuous piecewise constant process, with limits from the left. $X_t = l$, means if the process is in state l at time t.By the word history (or the past) at time t, we mean the information contained in the development of the process over the time interval $[0,t]$. That is, X_s, $0 \leq s \leq t$. $p_l(v,t)$ are conditional on the whole development up to time v. This expression is only considered for $t \geq v$, as it is trivial otherwise. This is the probability of the process X being in state l at time t given the development up to time v. The transition probabilities can be found from the specified transition hazards.

4.5 Model and Likelihood

In this part, we will introduce the definition of path specific hazard with a frailty term. The path for each subject is illustrated by Figure 4.1

4.5.1 Path Specific Hazard with Frailty Setup

The hazard for each translation states are defined as follows:

$$\lambda_i(t) = \lambda_{0i}(t) \cdot exp(x_i^T \cdot \beta_i + z_i^T \cdot b), \tag{4.2}$$

where λ_{0i} is the unspecified baseline hazard; β_i is vectors of regression coefficients associated with each hazard; x_i is a vector of covariates, z_i usually consists of 1 and a subset of covariates from x_i; and b represents random effects that account for possible associations among the three hazards. We assume a normal distribution for the random effects, $b \sim N(0, \Sigma)$. The zero mean constraint is imposed so that the random effects represent deviations from population averages. The covariance matrix Σ is assumed to be unstructured.

For the first state, where three events are competing risks: $i = 1, 2, 3$ represents non-terminal event 1, non-terminal event 2 and terminal event respectively. For the second state where terminal event is the semi-competing risk, $i = 4, 5$ represent from non-terminal event 1 to non-terminal event 2 and from non-terminal event 1 to terminal event, respectively; $i = 6, 7$ represent from non-terminal event 2 to non-terminal event 1 and from non-terminal event 2 to terminal event, respectively. For the third state where only the terminal event can happen, $i = 8, 9$ represent path from non-terminal event 1 to non-terminal event 2 then to terminal event and path from non-terminal event 2 to non-terminal event 1 then to terminal event, respectively. Figure 4.2 listed the detail of hazard definition.

4.5.2 Likelihood

Based on the path specific hazard definition, there are 5 feasible pathes for each subject, Figure (4.3) showed the detail of the event combination and event indicator combination for each feasible path.

Thus, the likelihood can be formed by each feasible path:

- Path 1 : Health $\rightarrow$ Non-terminal event 1 $\rightarrow$ Non-terminal event 2 $\rightarrow$ Terminal event.

$$\lambda_1(t_1)^{\delta_1} \cdot S_1(t_1) \cdot [\lambda_4(t_2)^{\delta_2} \cdot \frac{S_4(t_2)}{S_4(t_1)}]^{\delta_1 \cdot \delta_4} \cdot [\lambda_6(t_3)^{\delta_3} \cdot \frac{S_6(t_3)}{S_6(t_2)}]^{\delta_1 \cdot \delta_2 \cdot \delta_4}$$

- Path 2 : Health $\rightarrow$ Non-terminal event 1 $\rightarrow$ Terminal event.

$$\lambda_1(t_1)^{\delta_1} \cdot S_1(t_1) \cdot [\lambda_5(t_3)^{\delta_3} \cdot \frac{S_5(t_3)}{S_5(t_1)}]^{\delta_1 \cdot \delta_4}$$

- Path 3 : Health $\rightarrow$ Terminal event.

$$\lambda_3(t_3)^{\delta_3} \cdot S_3(t_3)$$

- Path 4 : Health $\rightarrow$ Non-terminal event 2 $\rightarrow$ Terminal event.

$$\lambda_2(t_2)^{\delta_2} \cdot S_2(t_2) \cdot [\lambda_8(t_3)^{\delta_3} \cdot \frac{S_8(t_3)}{S_8(t_2)}]^{\delta_2 \cdot \delta_5}$$

- Path 5: Health $\rightarrow$ Non-terminal event 2 $\rightarrow$ Non-terminal event 1 $\rightarrow$Terminal event.

$$\lambda_2(t_2)^{\delta_2} \cdot S_2(t_2) \cdot [\lambda_7(t_1)^{\delta_1} \cdot \frac{S_7(t_1)}{S_7(t_2)}]^{\delta_2 \cdot \delta_5} \cdot [\lambda_9(t_3)^{\delta_3} \cdot \frac{S_9(t_3)}{S_9(t_1)}]^{\delta_1 \cdot \delta_2 \cdot \delta_5}$$

In the mean time, the model can be built follow in the multi-state model idea. Figure (4.4) showed the detailed structure for the multi-state model procedure.

The likelihood is formed up by three states:

- State 1: At state 1, three events are competing,non-terminal event 1,2 and terminal event. If terminal event happened in this state, the whole procedure will stop in this state. If non-terminal event 1 or 2 happened in this state, the procedure will continue to state 2.

$$\lambda_1(t_1)^{\delta_1} \cdot S_1(t_1) \cdot \lambda_2(t_2)^{\delta_2} \cdot S_2(t_2) \cdot \lambda_3(t_3)^{\delta_3} \cdot S_3(t_3) \tag{4.3}$$

- State 2: At state 2, non-terminal event and terminal event are semi-competing, if non-terminal event 1 happened at state 1, then at state 2, non-terminal event 2 and terminal event are semi-competing; if non-terminal event 2 happened at state 1, then at state 2, non-terminal event 1 and terminal event are semi-competing. If terminal event happend in this state, the whole procedure of illness proceeding will stop there. If non-terminal event 1 or 2 happened in this state, the procedure will continue to state 3.

$$[\lambda_4(t_2)^{\delta_2} \cdot \frac{\lambda_4(t_2)}{\lambda_4(t_1)} \cdot \lambda_5(t_3)^{\delta_3} \cdot \frac{S_5(t_3)}{S_5(t_1)}]^{\delta_1 \cdot \delta_4} \cdot [\lambda_7(t_1)^{\delta_1} \cdot \frac{S_7(t_1)}{S_7(t_2)} \cdot \lambda_8(t_3)^{\delta_3} \cdot \frac{S_8(t_3)}{S_8(t_2)}]^{\delta_2 \cdot \delta_5} \tag{4.4}$$

- State 3: State 3 is the final state, the only possible event at state 3 is terminal event, there are two sources for state 3's terminal event, which depends on the semi-competing result from state 2.

$$[\lambda_6(t_3)^{\delta_3} \cdot \frac{S_6(t_3)}{S_6(t_2)}]^{\delta_1 \cdot \delta_2 \cdot \delta_4} \cdot [\lambda_9(t_3)^{\delta_3} \cdot \frac{S_9(t_3)}{S_9(t_1)}]^{\delta_1 \cdot \delta_2 \cdot \delta_5} \tag{4.5}$$

The two likelihood procedures lead to the same likelihood equation, the final likelihood is the product of each likelihood from each feasible path, also can be viewed

as the product of each likelihood come from each state. Regardless of the random effect, likelihood for each subject is as following.

$$\begin{aligned}
&\lambda_1(t_1)^{\delta_1} \cdot S_1(t_1) \cdot \lambda_2(t_2)^{\delta_2} \cdot S_2(t_2) \cdot \lambda_3(t_3)^{\delta_3} \cdot S_3(t_3) \\
&\cdot [\lambda_4(t_2)^{\delta_2} \cdot \frac{S_4(t_2)}{S_4(t_1)} \cdot \lambda_5(t_3)^{\delta_3} \cdot \frac{S_5(t_3)}{S_5(t_1)}]^{\delta_1 \cdot \delta_4} \cdot \\
&[\lambda_7(t_1)^{\delta_1} \cdot \frac{S_7(t_1)}{S_7(t_2)} \cdot \lambda_8(t_3)^{\delta_3} \cdot \frac{S_8(t_3)}{S_8(t_2)}]^{\delta_2 \cdot \delta_5} \cdot [\lambda_6(t_3)^{\delta_3} \cdot \frac{S_6(t_3)}{S_6(t_2)}]^{\delta_1 \cdot \delta_2 \cdot \delta_4} \\
&\cdot [\lambda_9(t_3)^{\delta_3} \cdot \frac{S_9(t_3)}{S_9(t_1)}]^{\delta_1 \cdot \delta_2 \cdot \delta_5}
\end{aligned} \tag{4.6}$$

Denote $\Lambda(t) = \int_0^t \lambda(u)du$ and $b \sim f_\tau(b)$, where $f_\tau(b)$ is the density function for the frailty term b. And $\vec{\beta} = (\beta_1, \beta_2, \cdots, \beta_9)$. The likelihood function is

$$L(\vec{\beta}) = \int_b \prod_i [\lambda_{01}(t_1) \cdot \exp\left(x_{1i}^T \cdot \beta_1 + z_i^T \cdot \tilde{b}\right)]^{\delta_1} \cdot \exp\left[-\Lambda_{01}(t_1) \cdot \exp\left(x_{1i}^T \cdot \beta_1 + z_i^T \cdot \tilde{b}\right)\right] \tag{4.7}$$

$$\cdot [\lambda_{02}(t_2) \cdot \exp\left(x_{2i}^T \cdot \beta_2 + z_i^T \cdot \tilde{b}\right)]^{\delta_2} \cdot \exp\left[-\Lambda_{02}(t_2) \cdot \exp\left(x_{2i}^T \cdot \beta_2 + z_i^T \cdot \tilde{b}\right)\right]$$

$$\cdot [\lambda_{03}(t_3) \cdot \exp\left(x_{3i}^T \cdot \beta_3 + z_i^T \cdot \tilde{b}\right)]^{\delta_3} \cdot \exp\left[-\Lambda_{03}(t_3) \cdot \exp\left(x_{3i}^T \cdot \beta_3 + z_i^T \cdot \tilde{b}\right)\right]$$

$$\cdot \{[\lambda_{04}(t_2) \cdot \exp\left(x_{4i}^T \cdot \beta_4 + z_i^T \cdot \tilde{b}\right)]^{\delta_2}$$

$$\cdot \exp\left[-(\Lambda_{04}(t_2) - \Lambda_{04}(t_1)) \cdot \exp\left(x_{4i}^T \cdot \beta_4 + z_i^T \cdot \tilde{b}\right)\right]$$

$$\cdot [\lambda_{05}(t_3) \cdot \exp\left(x_{5i}^T \cdot \beta_5 + z_i^T \cdot \tilde{b}\right)]^{\delta_3}$$

$$\cdot \exp\left[-(\Lambda_{05}(t_3) - \Lambda_{05}(t_1)) \cdot \exp\left(x_{5i}^T \cdot \beta_5 + z_i^T \cdot \tilde{b}\right)\right]\}^{\delta_1 \cdot \delta_4}$$

$$\cdot \{[\lambda_{07}(t_1) \cdot \exp\left(x_{7i}^T \cdot \beta_7 + z_i^T \cdot \tilde{b}\right)]^{\delta_1}$$

$$\cdot \exp\left[-(\Lambda_{07}(t_1) - \Lambda_{07}(t_2)) \cdot \exp\left(x_{7i}^T \cdot \beta_7 + z_i^T \cdot \tilde{b}\right)\right]$$

$$\cdot [\lambda_{08}(t_3) \cdot \exp\left(x_{8i}^T \cdot \beta_8 + z_i^T \cdot \tilde{b}\right)]^{\delta_3}$$

$$\cdot \exp\left[-(\Lambda_{08}(t_3) - \Lambda_{08}(t_2)) \cdot \exp\left(x_{8i}^T \cdot \beta_8 + z_i^T \cdot \tilde{b}\right)\right]\}^{\delta_1 \cdot \delta_5}$$

$$\cdot \{[\lambda_{06}(t_3) \cdot \exp\left(x_{6i}^T \cdot \beta_6 + z_i^T \cdot \tilde{b}\right)]^{\delta_3}$$

$$\cdot \exp\left[-(\Lambda_{06}(t_3) - \Lambda_{07}(t_2)) \cdot \exp\left(x_{6i}^T \cdot \beta_6 + z_i^T \cdot \tilde{b}\right)\right]\}^{\delta_1 \cdot \delta_2 \cdot \delta_4}$$

$$\cdot \{[\lambda_{09}(t_3) \cdot \exp\left(x_{9i}^T \cdot \beta_9 + z_i^T \cdot \tilde{b}\right)]^{\delta_3}$$

$$\cdot \exp\left[-(\Lambda_{09}(t_3) - \Lambda_{09}(t_1)) \cdot \exp\left(x_{9i}^T \cdot \beta_9 + z_i^T \cdot \tilde{b}\right)\right]\}^{\delta_1 \cdot \delta_2 \cdot \delta_5}$$

$$\cdot f_\tau(b) db$$

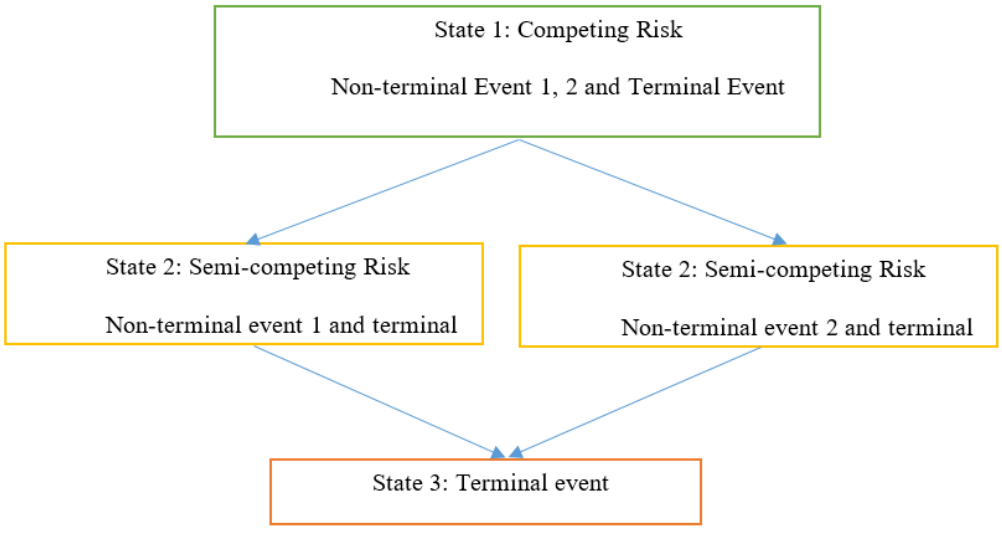
Figure 4.2: Hazard by Each Pathway with Frailty Model Setup in Bivariate Time to Events Data with Semi-competing Risk

Competing risk: from stage 0 (Health)	Non-terminal Event 1	$\lambda_1(t_1) = \lambda_{01}(t_1) \cdot \exp\{x_1^T \cdot \beta_1 + z_1^T \cdot b\}$
	Non-terminal Event 2	$\lambda_2(t_2) = \lambda_{02}(t_2) \cdot \exp\{x_2^T \cdot \beta_2 + z_2^T \cdot b\}$
	Terminal Event	$\lambda_3(t_3) = \lambda_{03}(t_1) \cdot \exp\{x_3^T \cdot \beta_3 + z_3^T \cdot b\}$
Conditional on Non-terminal Event 1	Non-terminal Event 1 to Non-terminal Event 2	$\lambda_4(t_2) = \lambda_{04}(t_2) \cdot \exp\{x_4^T \cdot \beta_4 + z_4^T \cdot b\}$
	Non-terminal Event 1 to Terminal Event	$\lambda_5(t_3) = \lambda_{05}(t_3) \cdot \exp\{x_5^T \cdot \beta_5 + z_5^T \cdot b\}$
	Non-terminal Event 1 to Non-terminal Event 2 to Terminal Event	$\lambda_6(t_3) = \lambda_{06}(t_1) \cdot \exp\{x_6^T \cdot \beta_6 + z_6^T \cdot b\}$
Conditional on Non-terminal Event 2	Non-terminal Event 2 to Non-terminal Event 1	$\lambda_7(t_1) = \lambda_{07}(t_1) \cdot \exp\{x_7^T \cdot \beta_7 + z_7^T \cdot b\}$
	Non-terminal Event 2 to Terminal Event	$\lambda_8(t_3) = \lambda_{08}(t_3) \cdot \exp\{x_8^T \cdot \beta_8 + z_8^T \cdot b\}$
	Non-terminal Event 2 to Non-terminal Event 1 to Terminal Event	$\lambda_9(t_3) = \lambda_{09}(t_3) \cdot \exp\{x_9^T \cdot \beta_9 + z_9^T \cdot b\}$

Figure 4.3: The Summarization of Possible Pathway in Bivariate Time to Events Data with a Semi-competing Risk

Path	Event combination	Event Indicator combination
1	Health→ Non-terminal event1→ Non-terminal event 2→ Terminal event	$\delta_1 = 1, \delta_2 = 1, \delta_3 = 1, \delta_4 = 1, \delta_5 = 0$
2	Health→ Non-terminal event1→ Terminal event	$\delta_1 = 1, \delta_2 = 0, \delta_3 = 1, \delta_4 = 1, \delta_5 = 0$
3	Health→ Terminal event	$\delta_1 = 0, \delta_2 = 0, \delta_3 = 1, \delta_4 = 0, \delta_5 = 0$
4	Health→ Non-terminal event 2→ Terminal event	$\delta_1 = 0, \delta_2 = 1, \delta_3 = 1, \delta_4 = 0, \delta_5 = 1$
5	Health→ Non-terminal event 2→ Non-terminal event 1→ Terminal event	$\delta_1 = 1, \delta_2 = 1, \delta_3 = 1, \delta_4 = 0, \delta_5 = 1$

Figure 4.4: The Multi-state Flow Chart to Composite Joint Likelihood for Bivariate Time to Events with a Semi-competing Risk



4.6 Estimation

4.6.1 Two Stage Pseudo-Likelihood Approach

In the two-stage pseudo likelihood estimation approach, the baseline cumulative hazards $A_{0i}(t)$ where $i = 1 \ldots 9$, are estimated by nonparametric Nelson-Aalen estimates in the first stage. Given the estimated $\hat{A_{0i}}$, the parameter of interest $\beta_1, \ldots, \beta_9$ is estimated in the second stage by maximizing the pseudo likelihood function after plug-in the estimates from the first stage into equation (4.7) .

First Stage: Estimating Baseline Cumulative Hazard

A non-parametric estimtor of cumulative hazard $\Lambda_{0i}(t)$ was first suggested by Wayne Nelson (Nelson, 1969, 1972,add ref) as a graphical tool to obtain engineering information on the form of the survival distribution in reliability studies.

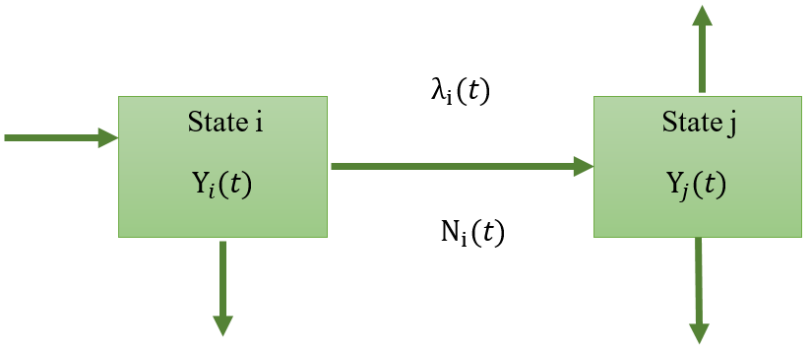


Figure 4.5: The Transition Plot for Component for Nelson-Aalen Estimator in Multistate Model From State i to State j

For time to event T, the counting process $N_i(t)$ represents whether or not the event has happended by or at t for the cause of i:

$$N_i(t) = I(T \leq t)$$

The at risk counting process is $Y(t)$ represents if the subject is at risk or not at time t,

$$Y(t) = I(T \geq t)$$

Note that for uncensored individual, we have

$$Y(t) = 1 - \sum_i N_i(t-)$$

According to the above definition, the Nelson-Aalen estimator for multi-state model can be written as

$$\hat{\Lambda}_i(t) = \int_0^t \frac{dN_i(u)}{Y(u)} du$$

Second Stage: Pseudo Likelihood

By plugging in the Nelson-Aalen estimator from the first stage, we have the pseudo likelihood proportional to the equation in (4.8).

$$
\begin{aligned}
L(\vec{\beta}|\widehat{\underline{\mathbf{\Lambda_0}}}) \propto \int_b \prod_i & \exp\left(x_{1i}^T \cdot \beta_1 + z_i^T \cdot b\right)^{\delta_1} \cdot \exp\left[-\hat{\Lambda}_{01}(t_1) \cdot \exp\left(x_{1i}^T \cdot \beta_1 + z_i^T \cdot b\right)\right] \quad (4.8) \\
& \cdot \exp\left(x_{2i}^T \cdot \beta_2 + z_i^T \cdot b\right)^{\delta_2} \cdot \exp\left[-\hat{\Lambda}_{02}(t_2) \cdot \exp\left(x_{2i}^T \cdot \beta_2 + z_i^T \cdot b\right)\right] \\
& \cdot \exp\left(x_{3i}^T \cdot \beta_3 + z_i^T \cdot b\right)^{\delta_3} \cdot \exp\left[-\hat{\Lambda}_{03}(t_3) \cdot \exp\left(x_{3i}^T \cdot \beta_3 + z_i^T \cdot b\right)\right] \\
& \cdot \{\exp\left(x_{4i}^T \cdot \beta_4 + z_i^T \cdot b\right)^{\delta_2} \cdot \exp\left[-\hat{\Lambda}_{04}(t_2) \cdot \exp\left(x_{4i}^T \cdot \beta_4 + z_i^T \cdot b\right)\right] \\
& \cdot \exp\left(x_{5i}^T \cdot \beta_5 + z_i^T \cdot b\right)^{\delta_3} \cdot \exp\left[-\hat{\Lambda}_{05}(t_3) \cdot \exp\left(x_{5i}^T \cdot \beta_5 + z_i^T \cdot b\right)\right]\}^{\delta_1 \cdot \delta_4} \\
& \cdot \{\exp\left(x_{7i}^T \cdot \beta_7 + z_i^T \cdot b\right)^{\delta_1} \cdot \exp\left[-\hat{\Lambda}_{07}(t_1) \cdot \exp\left(x_{7i}^T \cdot \beta_7 + z_i^T \cdot b\right)\right] \\
& \cdot \exp\left(x_{8i}^T \cdot \beta_8 + z_i^T \cdot b\right)^{\delta_3} \cdot \exp\left[-\hat{\Lambda}_{08}(t_3) \cdot \exp\left(x_{8i}^T \cdot \beta_8 + z_i^T \cdot b\right)\right]\}^{\delta_1 \cdot \delta_5} \\
& \cdot \{\exp\left(x_{6i}^T \cdot \beta_6 + z_i^T \cdot b\right)^{\delta_3} \cdot \exp\left[-\hat{\Lambda}_{06}(t_3) \cdot \exp\left(x_{6i}^T \cdot \beta_6 + z_i^T \cdot b\right)\right]\}^{\delta_1 \cdot \delta_2 \cdot \delta_4} \\
& \cdot \{\exp\left(x_{9i}^T \cdot \beta_9 + z_i^T \cdot b\right)^{\delta_3} \cdot \exp\left[-\hat{\Lambda}_{09}(t_3) \cdot \exp\left(x_{9i}^T \cdot \beta_9 + z_i^T \cdot b\right)\right]\}^{\delta_1 \cdot \delta_2 \cdot \delta_5} \\
& \cdot f_\tau(b) db
\end{aligned}
$$

Where $\widehat{\underline{\mathbf{\Lambda_0}}} = (\Lambda_{01}(t_1), \ldots, \Lambda_{09}(t_3))'$

The parameter of interest can be estimated by maximizing pseudo partial likelihood in (4.8). The standard errors can be estimated using the bootstrap method.

4.6.2 Penalized Partial Likelihood Approach

In this part, we are using penalized partial likelihood following Breslow and Clayton (1993) and Ripatti and Palmgren (2000), by apply the laplace approximation to the log-likelihood.

We assume that frailty term to follow a multivariate normal distribution, under proportional mean model and the general additive fraity model setup, the likelihood for observation data can be written as:

$$
\begin{aligned}
L(\vec{\beta}) =& \frac{1}{D(\tau)^{1/2}} \int_b \prod_i [\lambda_{01}(t_1) \cdot \exp\left(x_{1i}^T \cdot \beta_1 + z_i^T \cdot b\right)]^{\delta_1} \\
& \cdot \exp\left[-\Lambda_{01}(t_1) \cdot \exp\left(x_{1i}^T \cdot \beta_1 + z_i^T \cdot b\right)\right] \\
& \cdot [\lambda_{02}(t_2) \cdot \exp\left(x_{2i}^T \cdot \beta_2 + z_i^T \cdot b\right)]^{\delta_2} \cdot \exp\left[-\Lambda_{02}(t_2) \cdot \exp\left(x_{2i}^T \cdot \beta_2 + z_i^T \cdot b\right)\right] \\
& \cdot [\lambda_{03}(t_3) \cdot \exp\left(x_{3i}^T \cdot \beta_3 + z_i^T \cdot b\right)]^{\delta_3} \cdot \exp\left[-\Lambda_{03}(t_3) \cdot \exp\left(x_{3i}^T \cdot \beta_3 + z_i^T \cdot b\right)\right] \\
& \cdot \{[\lambda_{04}(t_2) \cdot \exp\left(x_{4i}^T \cdot \beta_4 + z_i^T \cdot b\right)]^{\delta_2} \\
& \cdot \exp\left[-(\Lambda_{04}(t_2) - \Lambda_{04}(t_1)) \cdot \exp\left(x_{4i}^T \cdot \beta_4 + z_i^T \cdot b\right)\right] \\
& \cdot [\lambda_{05}(t_3) \cdot \exp\left(x_{5i}^T \cdot \beta_5 + z_i^T \cdot b\right)]^{\delta_3} \\
& \cdot \exp\left[-(\Lambda_{05}(t_3) - \Lambda_{05}(t_1)) \cdot \exp\left(x_{5i}^T \cdot \beta_5 + z_i^T \cdot b\right)\right]\}^{\delta_1 \cdot \delta_4} \\
& \cdot \{[\lambda_{07}(t_1) \cdot \exp\left(x_{7i}^T \cdot \beta_7 + z_i^T \cdot b\right)]^{\delta_1} \\
& \cdot \exp\left[-(\Lambda_{07}(t_1) - \Lambda_{07}(t_2)) \cdot \exp\left(x_{7i}^T \cdot \beta_7 + z_i^T \cdot b\right)\right] \\
& \cdot [\lambda_{08}(t_3) \cdot \exp\left(x_{8i}^T \cdot \beta_8 + z_i^T \cdot b\right)]^{\delta_3} \\
& \cdot \exp\left[-(\Lambda_{08}(t_3) - \Lambda_{08}(t_2)) \cdot \exp\left(x_{8i}^T \cdot \beta_8 + z_i^T \cdot b\right)\right]\}^{\delta_1 \cdot \delta_5} \\
& \cdot \{[\lambda_{06}(t_3) \cdot \exp\left(x_{6i}^T \cdot \beta_6 + z_i^T \cdot b\right)]^{\delta_3} \\
& \cdot \exp\left[-(\Lambda_{06}(t_3) - \Lambda_{07}(t_2)) \cdot \exp\left(x_{6i}^T \cdot \beta_6 + z_i^T \cdot b\right)\right]\}^{\delta_1 \cdot \delta_2 \cdot \delta_4} \\
& \cdot \{[\lambda_{09}(t_3) \cdot \exp\left(x_{9i}^T \cdot \beta_9 + z_i^T \cdot \tilde{b}\right)]^{\delta_3} \\
& \cdot \exp\left[-(\Lambda_{09}(t_3) - \Lambda_{09}(t_1)) \cdot \exp\left(x_{9i}^T \cdot \beta_9 + z_i^T \cdot b\right)\right]\}^{\delta_1 \cdot \delta_2 \cdot \delta_5} \\
& \cdot e^{-\frac{1}{2} b^T D(v)^{-1} b} db
\end{aligned}
\tag{4.9}
$$

Since the intergrated log-likelihood (4.9) does not have a closed form expression. We write (4.9) as $\int \exp(-S(b))db$ and apply the Laplace approximation.

$$
\begin{aligned}
l(\vec{\beta}, \tilde{b}; \nu) \approx \sum_{i=1}^{n} \{ & \delta_{1i}[(x_{1i}^T \cdot \beta_1 + z_i^T \cdot \tilde{b}) - \log \sum_{j \in R_1} \exp(x_{1i}^T \cdot \beta_1 + z_i^T \cdot \tilde{b})] \\
& + \delta_{2i}[(x_{2i}^T \cdot \beta_2 + z_i^T \cdot \tilde{b}) - \log \sum_{j \in R_2} \exp(x_{2i}^T \cdot \beta_2 + z_i^T \cdot \tilde{b})] \\
& + (1 - \delta_{1i}) \cdot (1 - \delta_{2i}) \cdot \delta_{i3}[(x_{3i}^T \cdot \beta_3 + z_i^T \cdot \tilde{b}) - \log \sum_{j \in R_3} \exp(x_{1i}^T \cdot \beta_1 + z_i^T \cdot \tilde{b})] \\
& + \delta_{1i} \cdot \delta_{4i} \cdot \delta_{2i}[(x_{4i}^T \cdot \beta_4 + z_i^T \cdot \tilde{b}) - \log \sum_{j \in R_4} \exp(x_{4i}^T \cdot \beta_4 + z_i^T \cdot \tilde{b})] \\
& + \delta_{1i} \cdot \delta_{4i} \cdot \delta_{3i}[(x_{5i}^T \cdot \beta_5 + z_i^T \cdot \tilde{b}) - \log \sum_{j \in R_5} \exp(x_{5i}^T \cdot \beta_5 + z_i^T \cdot \tilde{b})] \\
& + \delta_{2i} \cdot \delta_{5i} \cdot \delta_{1i}[(x_{7i}^T \cdot \beta_7 + z_i^T \cdot \tilde{b}) - \log \sum_{j \in R_7} \exp(x_{7i}^T \cdot \beta_7 + z_i^T \cdot \tilde{b})] \\
& + \delta_{2i} \cdot \delta_{5i} \cdot \delta_{3i}[(x_{8i}^T \cdot \beta_8 + z_i^T \cdot \tilde{b}) - \log \sum_{j \in R_8} \exp(x_{8i}^T \cdot \beta_8 + z_i^T \cdot \tilde{b})] \\
& + \delta_{1i} \cdot \delta_{2i} \cdot \delta_{4i} \cdot \delta_{3i}[(x_{6i}^T \cdot \beta_6 + z_i^T \cdot \tilde{b}) - \log \sum_{j \in R_6} \exp(x_{6i}^T \cdot \beta_6 + z_i^T \cdot \tilde{b})] \\
& + \delta_{1i} \cdot \delta_{2i} \cdot \delta_{5i} \cdot \delta_{3i}[(x_{9i}^T \cdot \beta_9 + z_i^T \cdot \tilde{b}) - \log \sum_{j \in R_9} \exp(x_{9i}^T \cdot \beta_9 + z_i^T \cdot \tilde{b})]\} \\
& - \frac{1}{2} \tilde{b}^T D(\tau)^{-1} \tilde{b}
\end{aligned}
\tag{4.10}
$$

The above partial likelihood part in above penalized partial likelihood, also can be viewed as multi-state model procedure.

- State 1: At time T_1 three events are competing, illness 1 and 2 and death

$$[\frac{h_{01}(t_{1i})exp(z_i \cdot \beta_1 + \tilde{z}_i \cdot \tilde{b})}{\sum_{j \in R_1(t_{1i})} h_{01}(t_{1i})exp(z_j \cdot \beta_1 + \tilde{z}_j \cdot \tilde{b})}]^{\delta_{1i}} \cdot [\frac{h_{02}(t_{2i})exp(z_i \cdot \beta_2 + \tilde{z}_i \cdot \tilde{b})}{\sum_{j \in R_2(t_{1i})} h_{02}(t_{2i})exp(z_j \cdot \beta_2 + \tilde{z}_j \cdot \tilde{b})}]^{\delta_{2i}}$$
$$\cdot [\frac{h_{03}(t_{2i})exp(z_i \cdot \beta_3 + \tilde{z}_i \cdot \tilde{b})}{\sum_{j \in R_3(t_{1i})} h_{03}(t_{2i})exp(z_j \cdot \beta_3 + \tilde{z}_j \cdot \tilde{b})}]^{\delta_{3i}} \tag{4.11}$$

where R_1, R_2, R_3 are at risk sets, which are defined as:

$$R_1(t_{1i}) = \{j : t_{1j} \geq t_{1i}\} \tag{4.12}$$

$$R_2(t_{2i}) = \{j : t_{2j} \geq t_{2i}\} \tag{4.13}$$

$$R_3(t_{2i}) = \{j : t_{3j} \geq t_{3i}\} \tag{4.14}$$

- State 2: at time T_2 two events are competing, if illness 1 happened at state 1, then at T_2 illness 2 and death are competing; if illness 3 happened at state 1, then at T_2 illness 1 and death are competing. If illness 1 happened before illness 2, then $\delta_4 = 1$; If illness 2 happened before illness 1, then $\delta_5 = 1$.

$$\{[\frac{h_{05}(t_{2i})exp(z_i \cdot \beta_4 + \tilde{z}_i \cdot \tilde{b})}{\sum_{j \in R_4(t_{2i})} h_{04}(t_{2i})exp(z_j \cdot \beta_4 + \tilde{z}_j \cdot \tilde{b})}]^{\delta_{2i}} \tag{4.15}$$
$$\cdot [\frac{h_{05}(t_{3i})exp(z_i \cdot \beta_5 + \tilde{z}_i \cdot \tilde{b})}{\sum_{j \in R_5(t_{3i})} h_{05}(t_{3i})exp(z_j \cdot \beta_5 + \tilde{z}_j \cdot \tilde{b})}]^{\delta_{3i}}\}^{\delta_{1i}\delta_{4i}}$$

$$\{[\frac{h_{06}(t_{1i})exp(z_i \cdot \beta_6 + \tilde{z}_i \cdot \tilde{b})}{\sum_{j \in R_6(t_{1i})} h_{06}(t_{1i})exp(z_j \cdot \beta_6 + \tilde{z}_j \cdot \tilde{b})}]^{\delta_{1i}} \tag{4.16}$$
$$\cdot [\frac{h_{07}(t_{3i})exp(z_i \cdot \beta_7 + \tilde{z}_i \cdot \tilde{b})}{\sum_{j \in R_7(t_{3i})} h_{07}(t_{3i})exp(z_j \cdot \beta_7 + \tilde{z}_j \cdot \tilde{b})}]^{\delta_{3i}}\}^{\delta_{2i}\delta_{5i}}$$

where

$$R_{12}(t_{2i}) = \{j : \delta_{1j} \cdot \delta_{4j} = 1, t_{2j} \geq t_{2i}\} \tag{4.17}$$

$$R_{13}(t_{3i}) = \{j : \delta_{1j} \cdot \delta_{4j} = 1, t_{3j} \geq t_{3i}\} \tag{4.18}$$

$$R_{21}(t_{1i}) = \{j : \delta_{2j} \cdot \delta_{5j} = 1, t_{1j} \geq t_{1i}\} \tag{4.19}$$

$$R_{23}(t_{3i}) = \{j : \delta_{2j} \cdot \delta_{5j} = 1, t_{3j} \geq t_{3i}\} \tag{4.20}$$

- State 3: this is the final state, the only possible event at T_3 is death, there are two sources for state 3 death, one is from illness 1 to 2, the other is from illness 2 to 1

$$\{[\frac{h_{08}(t_{3i})exp(z_i \cdot \beta_8 + \tilde{z}_i \cdot \tilde{b})}{\sum_{j \in R_8(t_{1i})} h_{08}(t_{3i})exp(z_j \cdot \beta_8 + \tilde{z}_j \cdot \tilde{b})}]^{\delta_{3i}}\}^{\delta_{1i}\delta_{4i}\delta_{2i}} \tag{4.21}$$

$$\cdot \{[\frac{h_{09}(t_{3i})exp(z_i \cdot \beta_9 + \tilde{z}_i \cdot \tilde{b})}{\sum_{j \in R_9(t_{3i})} h_{09}(t_{3i})exp(z_j \cdot \beta_9 + \tilde{z}_j \cdot \tilde{b})}]^{\delta_{3i}}\}^{\delta_{2i}\delta_{5i}\delta_{1i}} \tag{4.22}$$

4.7 Simulation Study

We presented here a simulation study to investigate how the incorporation of shared frailties into multi-state models can improve parameter estimation. Simulation studies were conducted to compare the different estimation approaches and evaluate the performance of each estimation method under various scenarios.

4.7.1 Data Preparation

Denote the observed event time for illness 1, illness 2 and death as T_{1i}, T_{2i}, T_{3i}, respectively. The generation of semi-competing risks data based on bivariate time to

events consisted of five steps. Figure (4.6) summarized the procedure for the data simulation.

In the first step, survival times are generated for illness 1, illness 2, on death without illness. This stage is the competing component of semi-competing risks data. The survival function for competing stage can be defined as

$$S_{1\wedge 2\wedge 3} = exp[-\Lambda_1(t) - \Lambda_2(t) - \Lambda_3(t)]$$

Where $\Lambda_1(t), \Lambda_2(t)$, and $\Lambda_3(t)$ denote the cumulative hazards for illness 1, illness 2 and death before illnesses, respectively. $T_1^\star$ is the solution for

$$S_{1\wedge 2\wedge 3} = u_{1i}$$

Where $u_{1i} \sim U(0,1)$.Then, we use a trinominal experiment to decide the cause of failure of T_1^*. Generate another uniform distribution random variable $u_{2i} \sim U(0,1)$. Let

$$\begin{cases} \delta_1 = 1, & \text{if } u_{2i} \leq \frac{\lambda_1(t)}{\lambda_1(t)+\lambda_2(t)+\lambda_3(t)}. \\ \delta_2 = 1, & \text{if } \frac{\lambda_1(t)}{\lambda_1(t)+\lambda_2(t)+\lambda_3(t)} < u_{2i} \leq \frac{\lambda_1(t)+\lambda_2(t)}{\lambda_1(t)+\lambda_2(t)+\lambda_3(t)}. \\ \delta_3 = 1, & \text{if } \frac{\lambda_1(t)+\lambda_2(t)}{\lambda_1(t)+\lambda_2(t)+\lambda_3(t)} < u_{2i} \end{cases} \tag{4.23}$$

In the second step, based on the type of failure from state 1, the data will end in one of the following cases.

1. Case 1: if $\delta_3 = 1$, then death happened before any illnesses, so we have

$$T_1 = T_1^\star, \delta_1 = 0$$

$$T_2 = T_1^\star, \delta_2 = 0$$

$$T_3 = T_1^\star, \delta_3 = I(T_3 < C)$$

where C is universal censoring time.

2. Case 2:

   - If $\delta_1 = 1$, then simulate $T_2^\star$ for two competing risk, which is conditional on the history of illness 1, the illness 2 and death before illness 2 are competing. Generate another uniform random variable $u_{3i} \sim U(0,1)$

$$u_{3i} = \exp(-\frac{\Lambda_4(t)}{\Lambda_4(t_1^\star)} - \frac{\Lambda_5(t)}{\Lambda_5(t_1^\star)}) \tag{4.24}$$

   then use binorminal experiment to decide the cause of failure for $T_2^\star$. Generate another uniform random variable $u_{4i} \sim U(0,1)$

$$\begin{cases} \delta_2 = 1, & \text{if } u_{4i} \leq \frac{\Lambda_4(t)}{\Lambda_4(t)+\Lambda_5(t)}. \\ \delta_3 = 1, & \text{otherwise} \end{cases} \tag{4.25}$$

- If $\delta_3 = 1$, it means death happened before illness 2 but after illness 1, so we have

$$T_1 = T_1^\star, \delta_1 = 1$$

$$T_2 = T_2^\star, \delta_2 = 0$$

$$T_3 = T_2^\star, \delta_3 = I(T_3 < C)$$

- If $\delta_2 = 1$, it means death can only happen after illness 2, generate $T_3^\star$.Generate another uniform random variable $u_{5i} \sim U(0,1)$

$$u_{5i} = exp(-\Lambda_6(t_3^\star) + \Lambda_6(t_2^\star)) \tag{4.26}$$

then, we have

$$T_1 = T_1^\star, \delta_1 = 1$$

$$T_2 = T_2^\star, \delta_2 = 1$$

$$T_3 = T_3^\star, \delta_3 = I(T_3 < C)$$

3. Case 3:

- If $\delta_2 = 1$, then simulate $T_4^\star$ by two competing risks.Generate another uniform random variable $u_{6i} \sim U(0,1)$

$$u_{6i} = exp(-\frac{\Lambda_7(t)}{\Lambda_7(t_1^\star)} - \frac{\Lambda_7(t)}{\Lambda_7(t_1^\star)}) \tag{4.27}$$

then use binorminal experiment to decide the failure type,Generate another uniform random variable $u_{7i} \sim U(0,1)$

$$\begin{cases} \delta_1 = 1, & \text{if } u_{7i} \leq \frac{\Lambda_8(t)}{\Lambda_8(t)+\Lambda_8(t)}. \\ \delta_3 = 1, & \text{otherwise} \end{cases} \tag{4.28}$$

- if $\delta_3 = 1$, it means death happened before illness 1 but after illness 2, so we have

$$T_1 = T_4^\star, \delta_1 = 0$$

$$T_2 = T_1^\star, \delta_2 = 1$$

$$T_3 = T_4^\star, \delta_3 = I(T_3 < C)$$

- if $\delta_1 = 1$, it means death only can happen after illness 1, generate $T_5^\star$. Generate another uniform random variable $u_{8i} \sim U(0,1)$

$$u_{8i} = exp(-\Lambda_9(t_5^\star) + \Lambda_9(t_4^\star)) \tag{4.29}$$

then, we have

$$T_1 = T_4^\star, \delta_1 = 1$$

$$T_2 = T_1^\star, \delta_2 = 1$$

$$T_3 = T_5^\star, \delta_3 = I(T_3 < C)$$

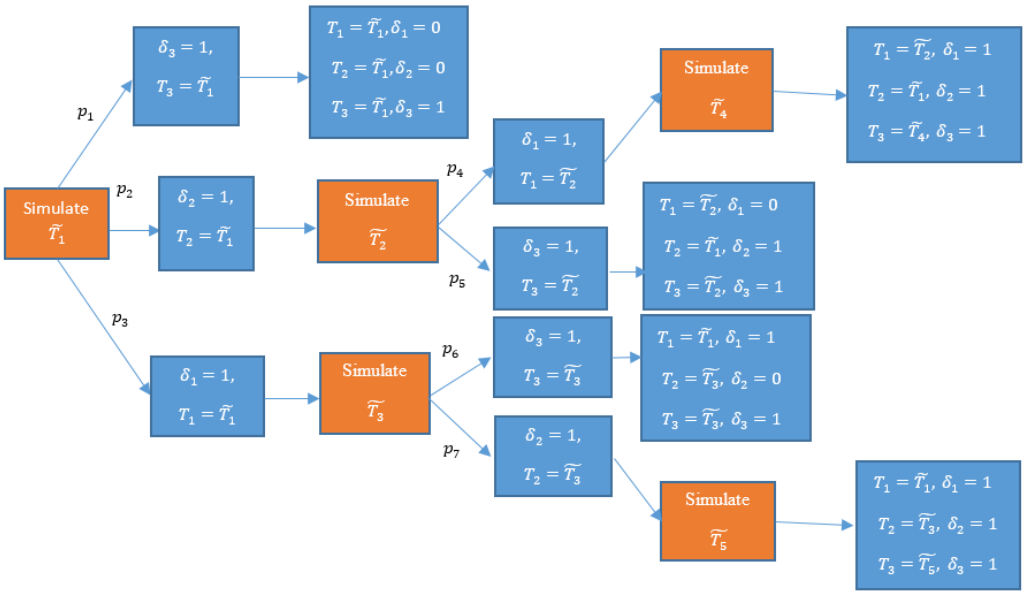


Figure 4.6: Data Preparation Flaw Chart for Simulation Study: Generating Bivariate Time to Events Data with a Semi-competing Risk

4.7.2 Simulation Results

For simulation and parameter estimations, we report in Table (4.1, 4.2), the average relative biase of the estimates(R.Bias), the average model based standard error estimates(M.SE), the empirical standard error estimates(E.SE) and coverage probabilities (M.CP) based on 95% the model based standard error estimates intervals.

We evaluated parameter estimation for the naive Cox model approach, parametric with exponential distribution as baseline hazard, two stage pseudo likelihood and penalized partial likelihood approach. Data for 1000 replications are generated with a total of $n = 100, 200$ observations for each replication. On average, from each simulated dataset, we observed 28.95% T_1 events, 31.87% T_2 events, 34% T_3 events

without T_1 and T_2 event, 18% T_3 event with T_1 only, 16% T_3 event with T_2 event only; 32% of T_3 event with both T_1 and T_2 events, respectively.

The analyses were conducted using the Cox models, the two stage pseudo likelihood model and the penalized partial likelihood(PPL). We can see that the all the methods perform well for regression parameter when there was small within subject variance, $\sigma_b^2 = 0.1$. However, as the within subject variance increase, the naive Cox proportional hazard estimation approach, which ignore the within subject correlation, provided very biased estimates. Compared to naive cox model, the two stage pseudo likelihood model provided much accurate estimator, but the standard error inflated as the variance of frailty increased. Compare to parametric model, the two stage estimate are more biased, since the method introduce more randomness from the first stage. If the baseline model is correctly specified, we expected the parametric model perform better than two stage estimate. On the other hand, when we can't have enough information for the baseline hazard, under model misspecification, we expects the two stage estimates perform better than parametric estimates.

Compare the simulation results summarized in Table (4.2)) in which simulation scenario $n = 200$ has larger sample size than Table (4.1)) scenario $n = 100$. We can see the significant improvement in the penalized partial likelihood estimates due to sample size increase.

Compared to the other three model estimates, the penalized partial likelihood estimator was more accurate and more robust in all the scenario. Thus, we recommend penalized partial likelihood method in applications.

Table 4.1: Bivariate Survival Data with a Semicompeting Risk Simulation Result: σ_b^2=0.1/0.5; Sample size=100

	Cox (Independent)				Parametric				Two Stage				Penalized			
	R.Bias	M.SE	E.SE	M.CP	R.Bias	M.SE	E.SE	M.CP	R.Bias	M.SE	E.SE	M.CP	R.Bias	M.SE	E.SE	M.CP
	σ_b^2=0.1, Sample Size=100															
β_1	-0.016	0.325	0.307	0.98	0.071	0.198	0.206	0.93	0.127	0.298	0.312	0.81	0.025	0.316	0.308	0.96
β_2	-0.094	0.326	0.329	0.96	0.107	0.192	0.185	0.94	0.213	0.305	0.331	0.83	0.106	0.327	0.331	0.96
β_3	-0.056	0.227	0.209	0.98	0.187	0.134	0.139	0.83	0.202	0.231	0.204	0.84	0.070	0.217	0.210	0.96
β_4	-1.025	187.690	2.960	0.90	0.119	0.376	0.379	0.94	1.298	1.470	2.890	0.42	0.191	0.623	0.658	0.96
β_5	-0.374	0.431	0.396	0.92	0.214	0.226	0.220	0.88	0.628	0.361	0.352	0.42	0.060	0.413	0.393	0.94
β_6	-1.195	178.339	2.850	0.90	0.293	0.271	0.275	0.87	1.504	1.881	2.809	0.34	0.105	0.640	0.642	0.99
β_7	-0.562	0.686	0.719	0.90	0.097	0.413	0.420	0.93	0.867	0.727	0.715	0.48	0.040	0.616	0.601	0.95
β_8	-0.385	0.438	0.500	0.86	0.282	0.220	0.225	0.84	0.643	0.454	0.457	0.44	0.065	0.411	0.451	0.92
β_9	-3.170	1470.852	6.555	0.90	0.338	0.319	0.312	0.83	3.508	4.488	6.493	0.86	0.465	17.121	2.177	0.94
	σ_b^2=0.1, Sample Size=100															
β_1	-0.109	0.318	0.318	0.94	0.053	0.226	0.217	0.96	0.042	0.308	0.327	0.98	-0.100	0.317	0.318	0.94
β_2	-0.049	0.315	0.331	0.94	0.089	0.212	0.200	0.94	0.105	0.293	0.330	0.84	-0.034	0.328	0.334	0.94
β_3	-0.085	0.220	0.207	0.92	0.168	0.159	0.164	0.93	0.091	0.230	0.201	0.86	-0.068	0.212	0.209	0.94
β_4	0.364	0.645	0.787	0.94	0.101	0.366	0.381	0.95	0.665	0.469	0.739	0.66	0.635	48.672	3.228	0.95
β_5	0.155	0.411	0.431	0.94	0.194	0.236	0.234	0.95	0.432	0.361	0.390	0.74	-0.073	0.400	0.402	0.96
β_6	0.798	252.134	2.959	0.98	0.272	0.295	0.283	0.96	1.105	0.480	2.927	0.60	-0.086	0.604	0.590	0.96
β_7	0.612	259.908	2.959	0.98	0.080	0.433	0.437	0.94	0.938	0.476	2.940	0.64	-0.053	0.633	0.643	0.96
β_8	0.125	0.409	0.442	0.94	0.263	0.247	0.249	0.94	0.412	0.353	0.400	0.62	-0.076	0.398	0.394	0.94
β_9	1.545	752.755	4.902	0.98	0.317	0.338	0.327	0.94	1.889	0.487	4.845	0.54	0.042	0.656	0.691	0.96

R.Bias:Relative bias

M.SE: Model based standard error.

E.SE: Empirical standard error.

M.CP: 95 % coverage probability based on M.SE.

Table 4.2: Bivariate Survival Data with a Semicompeting Risk Simulation Result: σ_b=0.1/0.5 sample size=200

	Cox (Independent)				Parametric				Two Stage				Penalized			
	R.Bias	M.SE	E.SE	M.CP	R.Bias	M.SE	E.SE	M.CP	R.Bias	M.SE	E.SE	M.CP	R.Bias	M.SE	E.SE	M.CP
	σ_b^2=0.1, sample size=200															
β_1	0.040	0.225	0.244	0.94	0.106	0.185	0.183	0.84	0.170	0.244	0.245	0.83	0.051	0.236	0.245	0.94
β_2	0.051	0.226	0.221	0.98	0.088	0.144	0.143	0.85	0.173	0.218	0.221	0.82	0.062	0.226	0.222	0.96
β_3	-0.008	0.157	0.131	0.98	0.127	0.072	0.073	0.75	0.153	0.128	0.124	0.84	0.009	0.142	0.131	0.96
β_4	0.529	0.445	0.454	0.78	0.099	0.259	0.253	0.95	0.803	0.446	0.449	0.52	0.085	0.462	0.465	0.96
β_5	0.320	0.293	0.352	0.82	0.169	0.165	0.169	0.92	0.602	0.322	0.320	0.56	0.022	0.351	0.353	0.94
β_6	1.051	146.627	2.907	0.68	0.178	0.259	0.256	0.94	1.434	0.276	0.278	0.34	0.030	0.470	0.477	0.92
β_7	0.470	0.433	0.365	0.86	0.137	0.251	0.255	0.93	0.762	0.366	0.367	0.54	0.031	0.353	0.351	0.95
β_8	0.264	0.287	0.272	0.88	0.217	0.126	0.122	0.86	0.554	0.251	0.254	0.53	-0.021	0.257	0.254	0.96
β_9	0.669	0.474	0.461	0.80	0.222	0.176	0.178	0.73	1.068	0.401	0.399	0.35	0.008	0.372	0.369	0.95
	σ_b^2=0.5, sample size=200															
β_1	-0.098	0.218	0.224	0.88	0.105	0.202	0.202	0.95	0.079	0.227	0.228	0.98	0.095	0.251	0.246	0.95
β_2	-0.095	0.218	0.221	0.94	0.088	0.145	0.144	0.95	0.076	0.234	0.224	0.90	0.097	0.247	0.244	0.95
β_3	-0.158	0.152	0.139	0.82	0.125	0.101	0.095	0.86	0.036	0.125	0.129	0.95	0.073	0.155	0.158	0.93
β_4	0.194	0.411	0.438	0.96	0.101	0.254	0.250	0.95	0.518	0.416	0.428	0.48	0.120	0.502	0.498	0.95
β_5	0.057	0.278	0.306	0.94	0.166	0.185	0.187	0.93	0.361	0.276	0.275	0.50	0.083	0.348	0.333	0.96
β_6	0.295	0.434	0.494	0.92	0.170	0.268	0.264	0.93	0.676	0.452	0.443	0.38	0.131	0.502	0.506	0.93
β_7	0.199	0.408	0.354	0.96	0.132	0.262	0.265	0.94	0.528	0.355	0.348	0.46	0.095	0.378	0.371	0.96
β_8	0.051	0.276	0.304	0.94	0.215	0.135	0.132	0.85	0.363	0.294	0.279	0.64	0.061	0.314	0.326	0.94
β_9	0.295	0.435	0.484	0.88	0.222	0.180	0.181	0.85	0.692	0.433	0.404	0.42	0.108	0.422	0.424	0.95

R.Bias:Relative bias

M.SE: Model based standard error.

E.SE: Empirical standard error.

M.CP: 95 % coverage probability based on M.SE.

4.8 Application

To illustrate the penalized partial likelihood estimation approaches in bivariate time to events with a semi-competing risk, we present a data analysis exploring potential gender differences in the association between time to coronary artery disease (CAD) and time to depression using data from the Indianapolis-Ibadan Dementia Project (IIDP). The detail of date description can be found in section 2.5.

For our analysis, the study population consisted of African American participants of the IIDP. All were age 65 or older residing in Indianapolis, Indiana. Recruitment was conducted at two-time points. During the first recruitment in 1992, 2212 African Americans age 65 or older living in Indianapolis were enrolled in the study. In 2001, the project enrolled 1893 additional African American community-dwelling participants 70 years and older. All participants agreed to undergo regular follow-up cognitive assessment and clinical evaluations. Details on the assembling of the original cohort and the enrichment cohort were described elsewhere.(Hall et al., 2009; Hendrie et al., 2001) Electronic medical records from 1992 to December 31, 2014, were retrieved as a re-identified data set to examine cardiovascular diseases and other risk factors. There were 4105 participants enrolled. We restricted our study to the subject who enrolled before 2010, whom have longer medical history recorded in the dataset. we have total 1428 subjects have complete record, within 1428 subjects, we have 79% death incidence, 76% of death events are female, 85% death events are male; The incidence for CAD is 33%, within which 34% is female and 31% is male; The incidence for depression is 17%, within which 20% is female and 12% is male.

Male group has higher incidence rate in death, but lower incidence rate in CAD and depression. The mean and median age of event onset can be found in Table(4.3).

Within 1428 subjects, we have 994 CAD cases happened before depression and death, 85 CAD cases happened after depression and before death; 384 depression cases before CAD and death; 113 depression after CAD; 30 cases of death happened before CAD and depression, 23 cases of death happened after CAD; 15 death happened after depression; 23 death happened after experience CAD at first stage and depression at second stage; 15 cases of death happened after experience depression at first stage and CAD at second stage. The detail of data distribution is showed in Figure (4.7).

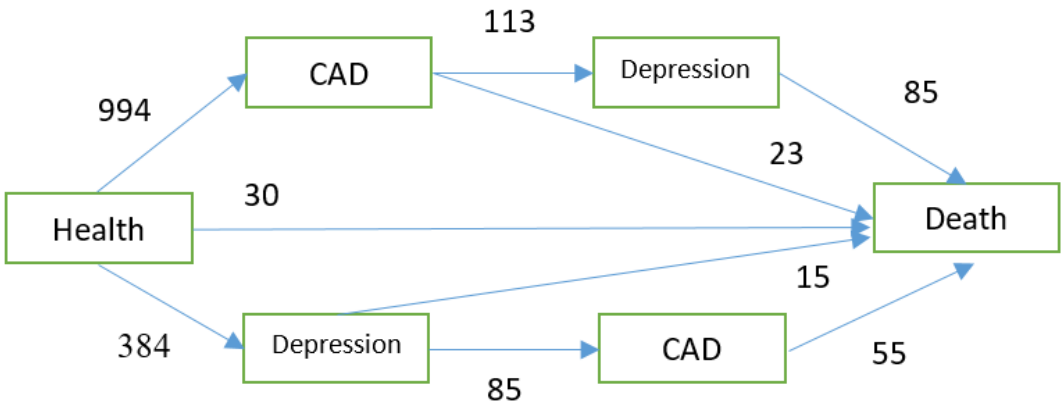


Figure 4.7: IIDP data for CAD and depression bivariate time to event with a semi-compting risk study application in detail

Table(4.4) summarized the results for naive Cox model estimation and penalized partial likelihood estimation. From the results, we observe that male group has higher risk of death in all the scenario using the naive Cox model approach. Compared to Cox proportional hazard estimation approach results in death, penalized partial

likelihood model provided quite similar results, except for death after depression at first stage and CAD at second stage situation.

In the meantime, the male group has higher risk of CAD in all the scenario using naive Cox proportional hazard estimation approach. Compared to Cox model result in CAD, penalized partial likelihood model shows the female group has higher risk of CAD conditional on the depression happened at first.

For the depression, both models give similar results, which shows that males group have reduced risk for depression.

Table 4.3: Baseline Age, Mean and Median Age of Event Onset By Gender

	Total			Female(Gender=0)			Male(Gender=1)		
	N	Mean (Std)	Median	N	Mean (Std)	Median	N	Mean (Std)	Median
Total (Baseage)	1428	74.04(6.73)	72.64	992	74.25(6.75)	72.86	436	73.56(6.68)	72.14
Death	1129(0.79)	84.28(7.53)	83.95	757(0.76)	84.85(7.69)	84.71	372(0.85)	83.12(7.06)	82.76
CAD	473(0.33)	80.07(6.87)	79.97	339(0.34)	80.25(6.96)	80.22	134(0.31)	79.60(6.65)	79.52
Depression	249(0.17)	80.13(7.11)	79.97	195(0.20)	80.13(7.20)	79.97	54(0.12)	80.14(6.81)	80.06

Table 4.4: Application result for gender effect in CAD and Depression with death as semi-competing risk

	Cox (Naive Independent)		Penalized Partial	
	Estimate	Standard Error	Estimate	Standard Error
β_1	0.199^*	0.062	0.020	0.114
β_2	-0.393^*	0.096	-0.353^*	0.121
β_3	0.426^*	0.048	0.345^*	0.099
β_4	-0.264	0.154	-0.403^*	0.124
β_5	0.396^*	0.082	0.255^*	0.124
β_6	0.784^*	0.226	0.327	0.269
β_7	0.152	0.155	-0.784	0.477
β_8	0.221	0.143	0.176	0.291
β_9	0.129	0.314	-0.916	0.793

$*$: $p < 0.05$

4.9 Conclusion and Discussion

We developed flexible frailty based semiparametric model for multi-event and multi-state survival data with a semicompeting risks. Our models can incorporate different covariates into the frailty terms for three different types of hazard functions corresponding to the illness, death without illness, and death after illness. Our methods extended the gamma frailty models by Xu et al. (2010) which used a single frailty term to correlate the events and did not consider covariates for the frailty term.

In observational studies of chronic disease and aging, this model will help address and identify risk factor for the terminal event after the occurrence of the non-terminal event. We used penalized partial likelihood methods for estimation, which provide accurate and robust estimates for inference.

Our models will also work with clustered data (Gray, 1994; Gustafson, 1997). Further they can be extended beyond shared frailty models. For example, Gustafson (1997) described a semicompeting risks model where relapse and death have correlated frailties associated with clusters in addition to the random intercept specific to individual subjects. And Rotolo (2013) thoroughly discussed nested frailty and two level of frailty based parametric model and semiparametric model in multi-state situation. Our model could also be easily extended to such correlated frailty models and multi-level frailty model.

Chapter 5

Conclusion and Discussion

In this thesis we studied several topics related to joint models of time to events data with semi-competing risk. The bivariate and multivariate survival data can arise in practice in difference ways, each study subject may experience several events or when there exists some natural or artificial grouping of subjects which induced dependence among failure times of the same group. Biomedical examples include the sequence of tumor recurrences or infection episodes, the development of physical symptoms or diseases in several organ systems, the onset of a disease among family members, the onset of multiple disease in same subject. Throughout the dissertation,we focus on estimating the association between risk factors and multiple events using model formulation, development of estimation algorithms and asymptotic results.

In chapter 2, we studied the covariate dependent association: cross ratio, between bivariate survival times. The cross ratio is formulated as the ratio of two conditional hazard functions and thus measures the relative hazard of one time component conditional on another time component at some time point and beyond. A question of significant interest in the gender effect to identify risk of coronary artery disease (CAD) and depression. The Indiana Ibadan Dementia Project data provide a unique opportunity for evaluating the gender effect in the risk of chronical disease by assessing the association between age at onset of CAD and age at onset of depression. Formal statistical analysis of this dependence is challenging due to the facts that both

the time to events are subject to right-censoring and that their association depends on age of event happened. Thus, in chapter 2, we consider a covariate dependent cross ratio model estimation the dependence between the two events adjusting for gender by using pseudo partial likelihood method instead of the true likelihood.

In the chapters 3 and 4, we focused on a unified approach to utilize information on electronic medical record (EMR) for series of events. When subjects experienced multiple events, the case can be further complicated by the presence of semi-competing risk. Semi-competing risks data are encountered when there is a terminating event which potentially censors a nonterminating event. We proposed frailty based semiparametric model for univariate event and bivariate events, when there was a semicompeting risk. The concept of frailty model provides a suitable way to introduce random effects in the model to account for unobserved heterogeneity. In its simplest form, a frailty is an unobserved random factor that modifies multiplicatively the hazard function of an individual or a group or cluster of individuals. There has recently been increased attention to semicompeting risk data as distinct from classical competing risks data, in particular, inferences without covariates. In chapters 3 and 4, we incorporate covariates. New penalized partial likelihood estimators are constructed using Laplace approximation of the true likelihood, and the asymptotic property has been demonstrated in simulation studies. The proposed model, associated algorithmic and method were ready to use in statistical estimation and inference.

The current methods can easily take into time-dependent variables account.The dependence between multiple events is introduced through a modulation mechanism that leads itself naturally to incorporation of time-dependent covariates. We will then

have a better chance of understanding the disease progression mechanism and many other aspects of the disease monitoring and treatment.

The current framework may be extended in several directions:

1. Extension to other hazard models.

   In addition to Cox proportional hazard models, there are other important regression models in survival analysis including accelerated failure time models (AFT) (Lawless, 2011; Wei, 1992). The Cox model and its various generalizations are mainly used in medical and biostatistical fields, while the AFT model is primarily applied in reliability theory and industrial experiments. AFT model offers a potentially useful statistical approach that is based upon the survival curve rather than the hazard function. It will be a meaningful extension if our method can be extended into AFT model in the presence of competing risk and semicompeting risk.

2. Interval censoring on the disease outcomes.

   Interval-censored data are often found in medical studies in which subjects are assessed only periodically for the response of interest. The time when the event of interest occurs is not directly observed but is known to take place within some time intervals. For example, in a clinical trial subjects might visit a clinic for assessment at predetermined times.The onset of a condition of interest is known only to have occurred at some time between visits; the exact time of onset is not known. The times of occurrence of these events are said to be interval-censored.

It will be another meaningful extension to consider non-terminal event with interval censoring situation in our semi-competing risk model.

3. Model selection and model diagnostics.

   To identify potential variables and appropriate frailty distribution would be very helpful in failure data analysis. Selection of a proper model as a basis for statistical inference is critical. This is especially so in the analysis of multiple, interrelated events. To develop information criteria for model identification and variable selection would be a meaningful extension. It is also important to develop summary statistic to guide model diagnostics so that deviations from the assumed frailty distribution can be detected.

4. Markov chain Monte Carlo (MCMC) and Bayesian Approach for computation, estimation and prediction.

   The current model presents computational challenge to standard likelihood based approach because it involves high-dimensional integrations. The Bayesian MCMC approach maybe a good option to solve this problem. In the meantime, the Bayesian MCMC can be conveniently implemented in general software package like Stan/WinBUGS. The use of Bayesian methods also makes event prediction very straightforward.

In conclusion, this dissertation developed novel statistical methods for analyzing univariate and bivariate survival times in the presence of semi-competing risk. Our methods are readily applicable to a wide range of studies where multiple time to events are observed in order to achieve unbiased results.

Chapter 6

Appendix

6.1 More Simulation Results for Covariate Dependent Cross Ratio of Bivariate Survival Times

Table 6.1: Clayton Copula Estimation Approach with Data Generated From Clayton Copula with Exponential Distribution as Marginal

T1	T2	EST(θ)	EST(β)	RelativeBias(θ)	RelativeBias(β)	MSE(θ)	MSE(β)	ESE(θ)	ESE(β)	MCP(θ)	MCP(β)
					sample size=100						
0 censor	0 censor	3.06	0.51	0.02	0.02	0.28	0.22	0.42	0.17	0.84	0.98
10% censor	10% censor	2.83	0.53	-0.06	0.06	0.28	0.23	0.43	0.19	0.64	0.98
30% censor	30% censor	2.53	0.51	-0.16	0.02	0.30	0.27	0.43	0.27	0.58	0.96
0 censor	10% censor	2.99	0.50	0.00	-0.01	0.28	0.23	0.45	0.19	0.80	0.96
10% censor	30% censor	2.63	0.51	-0.12	0.03	0.29	0.25	0.45	0.23	0.62	0.96
30% censor	0 censor	2.70	0.52	-0.10	0.04	0.28	0.24	0.43	0.23	0.66	0.92
					sample size=400						
0 censor	0 censor	3.04	0.51	0.01	0.01	0.14	0.11	0.18	0.08	0.88	0.98
10% censor	10% censor	2.84	0.50	-0.05	0.01	0.14	0.12	0.18	0.08	0.76	0.98
30% censor	30% censor	2.52	0.48	-0.16	-0.05	0.15	0.13	0.19	0.12	0.16	0.98
0 censor	10% censor	2.93	0.51	-0.02	0.02	0.14	0.11	0.18	0.09	0.80	0.98
10% censor	30% censor	2.64	0.50	-0.12	0.00	0.14	0.12	0.20	0.10	0.34	0.98
30% censor	0 censor	2.66	0.50	-0.11	-0.01	0.14	0.12	0.17	0.10	0.36	0.98
					sample size=800						
0 censor	0 censor	3.01	0.51	0.00	0.02	0.10	0.08	0.12	0.06	0.84	0.96
10% censor	10% censor	2.82	0.50	-0.06	0.01	0.10	0.08	0.13	0.07	0.52	0.96
30% censor	30% censor	2.51	0.46	-0.16	-0.07	0.10	0.09	0.13	0.07	0.02	0.98
0 censor	10% censor	2.91	0.51	-0.03	0.02	0.10	0.08	0.11	0.06	0.82	0.96
10% censor	30% censor	2.63	0.48	-0.12	-0.04	0.10	0.09	0.13	0.08	0.12	0.98
30% censor	0 censor	2.65	0.51	-0.12	0.01	0.10	0.08	0.14	0.08	0.14	0.98

EST($\cdot$): Average estimated value

RelativeBias($\cdot$): Relative bias=(EST($\cdot$)-TRUE($\cdot$))/TRUE($\cdot$)

MSE($\cdot$): Model based standard error.

ESE($\cdot$): Empirical standard error.

MCP($\cdot$): Coverage probability based on model based standard error.

Table 6.2: Two-stage Semiparametric Estimation Approach with Data Generated From Clayton Copula with Exponential Distribution Marginal

T1	T2	EST(θ)	EST(β)	RelativeBias(θ)	RelativeBias(β)	MSE(θ)	MSE(β)	ESE(θ)	ESE(β)	MCP(θ)	MCP(β)
sample size=100											
0 censor	0 censor	2.95	0.49	-0.02	-0.02	0.28	0.22	0.60	0.24	0.64	0.96
10% censor	10% censor	3.07	0.52	0.02	0.04	0.28	0.23	0.70	0.26	0.70	0.94
30% censor	30% censor	3.86	0.55	0.29	0.09	0.30	0.27	1.64	0.70	0.58	0.90
0 censor	10% censor	3.17	0.46	0.06	-0.08	0.28	0.23	0.80	0.31	0.60	0.88
10% censor	30% censor	4.24	0.50	0.41	0.00	0.29	0.25	2.60	0.63	0.52	0.86
30% censor	0 censor	2.64	0.28	-0.12	-0.44	0.28	0.24	0.42	0.25	0.76	0.92
sample size=400											
0 censor	0 censor	3.03	0.49	0.01	-0.02	0.14	0.11	0.22	0.08	0.74	0.98
10% censor	10% censor	3.18	0.50	0.06	0.00	0.14	0.12	0.27	0.10	0.62	0.98
30% censor	30% censor	3.58	0.50	0.19	0.00	0.15	0.13	0.54	0.21	0.40	0.92
0 censor	10% censor	3.16	0.47	0.05	-0.05	0.14	0.11	0.28	0.11	0.72	1.00
10% censor	30% censor	3.89	0.43	0.30	-0.14	0.14	0.12	0.64	0.28	0.18	0.74
30% censor	0 censor	2.54	0.29	-0.15	-0.42	0.14	0.12	0.19	0.12	0.16	0.66
sample size=800											
0 censor	0 censor	2.92	0.54	-0.03	0.08	0.10	0.08	0.33	0.21	0.76	0.96
10% censor	10% censor	3.06	0.54	0.02	0.08	0.10	0.08	0.37	0.21	0.62	0.98
30% censor	30% censor	3.42	0.52	0.14	0.05	0.10	0.09	0.49	0.22	0.30	1.00
0 censor	10% censor	3.06	0.51	0.02	0.01	0.10	0.08	0.36	0.22	0.60	0.92
10% censor	30% censor	3.63	0.42	0.21	-0.16	0.10	0.09	0.56	0.25	0.12	0.76
30% censor	0 censor	2.53	0.32	-0.16	-0.36	0.10	0.08	0.24	0.25	0.06	0.16

EST($\cdot$): Average estimated value

RelativeBias($\cdot$): Relative bias=(EST($\cdot$)-TRUE($\cdot$))/TRUE($\cdot$)

MSE($\cdot$): Model based standard error.

ESE($\cdot$): Empirical standard error.

MCP($\cdot$): Coverage probability based on model based standard error.

Table 6.3: Pseudo Partial Likelihood Estimation Approach with Data Generated From Clayton Copula with Exponential Distribution as Marginal

T1	T2	EST(θ)	EST(β)	RelativeBias(θ)	RelativeBias(β)	MSE(θ)	MSE(β)	ESE(θ)	ESE(β)	MCP(θ)	MCP(β)
sample size=100											
0 censor	0 censor	3.19	0.48	0.06	-0.05	0.26	0.21	0.27	0.24	0.68	0.96
10% censor	10% censor	3.16	0.50	0.05	0.01	0.29	0.23	0.26	0.25	0.70	0.98
30% censor	30% censor	3.21	0.50	0.07	-0.01	0.38	0.30	0.34	0.30	0.70	0.93
0 censor	10% censor	3.24	0.46	0.08	-0.09	0.28	0.22	0.28	0.25	0.66	0.93
10% censor	30% censor	3.25	0.48	0.08	-0.04	0.34	0.27	0.35	0.30	0.68	0.96
30% censor	0 censor	3.18	0.50	0.06	0.00	0.31	0.25	0.26	0.25	0.70	0.95
sample size=400											
0 censor	0 censor	3.09	0.51	0.03	0.02	0.12	0.09	0.10	0.11	0.71	0.98
10% censor	10% censor	3.08	0.51	0.03	0.03	0.13	0.10	0.10	0.11	0.71	0.98
30% censor	30% censor	3.04	0.52	0.01	0.04	0.16	0.13	0.11	0.12	0.72	0.94
0 censor	10% censor	3.09	0.51	0.03	0.02	0.12	0.10	0.10	0.12	0.71	0.94
10% censor	30% censor	3.08	0.51	0.03	0.02	0.15	0.12	0.10	0.11	0.71	0.98
30% censor	0 censor	3.05	0.51	0.02	0.03	0.14	0.11	0.11	0.12	0.71	0.96
sample size=800											
0 censor	0 censor	3.03	0.50	0.01	0.00	0.08	0.06	0.06	0.07	0.70	0.92
10% censor	10% censor	3.02	0.50	0.01	-0.01	0.09	0.07	0.06	0.07	0.70	0.98
30% censor	30% censor	3.02	0.48	0.01	-0.05	0.11	0.09	0.07	0.07	0.68	0.92
0 censor	10% censor	3.03	0.50	0.01	-0.01	0.08	0.07	0.06	0.07	0.70	0.94
10% censor	30% censor	3.04	0.48	0.01	-0.04	0.10	0.08	0.07	0.07	0.68	0.94
30% censor	0 censor	3.03	0.49	0.01	-0.02	0.09	0.08	0.06	0.07	0.69	0.92

EST($\cdot$): Average estimated value

RelativeBias($\cdot$): Relative bias=(EST($\cdot$)-TRUE($\cdot$))/TRUE($\cdot$)

MSE($\cdot$): Model based standard error.

ESE($\cdot$): Empirical standard error.

MCP($\cdot$): Coverage probability based on model based standard error.

BIBLIOGRAPHY

Alexopoulos, G. S., B. S. Meyers, R. C. Young, S. Campbell, D. Silbersweig, and M. Charlson (1997). 'vascular depression'hypothesis. *Archives of general psychiatry 54*(10), 915–922.

Andersen, P. K., Ø. Borgan, N. L. Hjort, E. Arjas, J. Stene, and O. Aalen (1985). Counting process models for life history data: A review [with discussion and reply]. *Scandinavian Journal of Statistics*, 97–158.

Bandeen-Roche, K. and K.-Y. Liang (2002). Modelling multivariate failure time associations in the presence of a competing risk. *Biometrika 89*(2), 299–314.

Bandeen-Roche, K. and J. Ning (2008). Nonparametric estimation of bivariate failure time associations in the presence of a competing risk. *Biometrika*.

Bates, D. M. and S. DebRoy (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis 91*(1), 1–17.

Bhattacharyya, M. and J. P. Klein (2005). A random effects model for multistate survival analysis with application to bone marrow transplants. *Mathematical biosciences 194*(1), 37–48.

Box-Steffensmeier, J. M., S. De Boef, and K. A. Joyce (2007). Event dependence and heterogeneity in duration models: the conditional frailty model. *Political Analysis 15*(3), 237–256.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association 88*(421), 9–25.

Callahan, C. M., F. D. Wolinsky, T. E. Stump, N. A. Nienaber, S. L. Hui, and W. M. Tierney (1998). Mortality, symptoms, and functional impairment in late-life depression. *Journal of general internal medicine 13*(11), 746–752.

Campbell, N., M. Boustani, K. Lane, S. Gao, H. Hendrie, B. Khan, J. Murrell, F. Unverzagt, A. Hake, V. Smith-Gamble, et al. (2010). Use of anticholinergics and the risk of cognitive impairment in an african american population. *Neurology 75*(2), 152–159.

Cheng, Y. and J. P. Fine (2008). Nonparametric estimation of cause-specific cross hazard ratio with bivariate competing risks data. *Biometrika 95*(1), 233–240.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika 65*(1), 141–151.

Cortinas Abrahantes, J. and T. Burzykowski (2005). A version of the em algorithm for proportional hazard model with random effects.

Cox, D. and H. Miller (1965). The theory of stochastic processes, methuen & co. *Ltd, London, UK*.

Cox, D. R. and D. V. Hinkley (1979). *Theoretical statistics*. CRC Press.

Cox, D. R. and D. Oakes (1984). *Analysis of survival data*, Volume 21. CRC Press.

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

de Groot, J. C., F.-E. de Leeuw, M. Oudkerk, A. Hofman, J. Jolles, and M. M. Breteler (2000). Cerebral white matter lesions and depressive symptoms in elderly adults. *Archives of general psychiatry 57*(11), 1071–1076.

de Jonge, P. and A. M. Roest (2012). Depression and cardiovascular disease: the end of simple models. *The British Journal of Psychiatry 201*(5), 337–338.

Ding, A. A., G. Shi, W. Wang, and J.-J. HSIEH (2009). Marginal regression analysis for semi-competing risks data under dependent censoring. *Scandinavian Journal of Statistics 36*(3), 481–500.

Diva, U., D. K. Dey, and S. Banerjee (2008). Parametric models for spatially correlated survival data for individuals with multiple cancers. *Statistics in medicine 27*(12), 2127–2144.

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

Gao, S., H. C. Hendrie, K. S. Hall, and S. Hui (1998). The relationships between age, sex, and the incidence of dementia and alzheimer disease: a meta-analysis. *Archives of general psychiatry 55*(9), 809–815.

Gorfine, M. and L. Hsu (2011). Frailty-based competing risks model for multivariate survival data. *Biometrics 67*(2), 415–426.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association 87*(420), 942–951.

Gray, R. J. (1994). A bayesian analysis of institutional effects in a multicenter cancer clinical trial. *Biometrics*, 244–253.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique*, 245–259.

Gureje, O., C. A. Rodenberg, and O. Baiyewu (1995). Prevalence of alzheimer's disease and dementia in two communities: Nigerian africans and african americans. *Am J Psychiatry 152*, 1485–1492.

Gustafson, P. (1995). A bayesian analysis of bivariate survival data from a multicentre cancer clinical trial. *Statistics in medicine 14*(23), 2523–2535.

Gustafson, P. (1997). Large hierarchical bayesian analysis of multivariate survival data. *Biometrics*, 230–242.

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

Han, B., M. Yu, J. J. Dignam, and P. J. Rathouz (2014). Bayesian approach for flexible modeling of semicompeting risks data. *Statistics in medicine 33*(29), 5111–5125.

Hawkins, M. A., C. M. Callahan, T. E. Stump, and J. C. Stewart (2014). Depressive symptom clusters as predictors of incident coronary artery disease events: A 15-year prospective study of older adults. *Psychosomatic medicine 76*(1), 38.

He, W. and J. F. Lawless (2003). Flexible maximum likelihood methods for bivariate proportional hazards models. *Biometrics 59*(4), 837–848.

Hendrie, H. C., A. Ogunniyi, K. S. Hall, O. Baiyewu, F. W. Unverzagt, O. Gureje, S. Gao, R. M. Evans, A. Ogunseyinde, A. Adeyinka, et al. (2001). Incidence of dementia and alzheimer disease in 2 communities: Yoruba residing in ibadan, nigeria, and african americans residing in indianapolis, indiana. *Jama 285*(6), 739–747.

Hoem, J. M., N. Keiding, H. Kulokari, B. Natvig, O. Barndorff-Nielsen, and J. Hilden (1976). The statistical theory of demographic rates: A review of current developments [with discussion and reply]. *Scandinavian Journal of Statistics*, 169–185.

Hougaard, P. (1999). Multi-state models: a review. *Lifetime data analysis 5*(3), 239–264.

Hougaard, P. (2012). *Analysis of multivariate survival data*. Springer Science & Business Media.

Hsu, L. and Z. Moodie (2007). Analysis of dependence in multivariate failure-time data. *Statistical Advances in the Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics*, 221–243.

Hu, T. (2011). *Time-dependent cross ratio estimation for bivariate failure times*. Ph. D. thesis, University of Michigan.

Hu, T., B. Nan, X. Lin, and J. M. Robins (2011). Time-dependent cross ratio estimation for bivariate failure times. *Biometrika 98*(2), 341–354.

Hyttinen, V., J. Kaprio, L. Kinnunen, M. Koskenvuo, and J. Tuomilehto (2003). Genetic liability of type 1 diabetes and the onset age among 22,650 young finnish twin pairs a nationwide follow-up study. *Diabetes 52*(4), 1052–1055.

Jackson, C. (2016). Multi-state modelling with r: the msm package.

Kalbfleisch, J. D. and R. L. Prentice (2002). Relative risk (cox) regression models. *The Statistical Analysis of Failure Time Data, Second Edition*, 95–147.

Karaca-Mandic, P. and K. Train (2003). Standard error correction in two-stage estimation with nested samples. *The Econometrics Journal 6*(2), 401–407.

Kendall, M. G. (1948). Rank correlation methods.

Klein, J. P. (1992). Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics*, 795–806.

Lakhal, L., L.-P. Rivest, and B. Abdous (2008). Estimating survival and association in a semicompeting risks model. *Biometrics 64*(1), 180–188.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data*, Volume 362. John Wiley & Sons.

Lawless, J. F. and Y. E. Yilmaz (2011). Comparison of semiparametric maximum likelihood estimation and two-stage semiparametric estimation in copula models. *Computational Statistics & Data Analysis 55*(7), 2446–2455.

Li, Y. and X. Lin (2012). Semiparametric normal transformation models for spatially correlated survival data. *Journal of the American Statistical Association*.

Li, Y., R. L. Prentice, and X. Lin (2008). Semiparametric maximum likelihood estimation in normal transformation models for bivariate survival data. *Biometrika 95*(4), 947–960.

Liu, L., R. A. Wolfe, and X. Huang (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics 60*(3), 747–756.

McGilchrist, C. and C. Aisbett (1991). Regression with frailty in survival analysis. *Biometrics*, 461–466.

Möller-Leimkühler, A. M. (2010). Higher comorbidity of depression and cardiovascular disease in women: a biopsychosocial perspective. *The world journal of biological psychiatry 11*(8), 922–933.

Nan, B., X. Lin, L. D. Lisabeth, and S. D. Harlow (2006). Piecewise constant cross-ratio estimation for association of age at a marker event and age at menopause. *Journal of the American Statistical Association 101*(473), 65–77.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Ning, J. and K. Bandeen-Roche (2014). Estimation of time-dependent association for bivariate failure times in the presence of a competing risk. *Biometrics 70*(1), 10–20.

Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 414–422.

Oakes, D. (1986). Semiparametric inference in a model for association in bivanate survival data. *Biometrika 73*(2), 353–361.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association 84*(406), 487–493.

Othus, M. and Y. Li (2010). A gaussian copula model for multivariate survival data. *Statistics in biosciences 2*(2), 154–179.

Peng, L. and J. P. Fine (2007). Regression modeling of semicompeting risks data. *Biometrics 63*(1), 96–108.

Perperoglou, A. (2014). Cox models with dynamic ridge penalties on time-varying effects of the covariates. *Statistics in medicine 33*(1), 170–180.

Putter, H., M. Fiocco, R. Geskus, et al. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine 26*(11), 2389.

Ripatti, S. and J. Palmgren (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics 56*(4), 1016–1022.

Rondeau, V., Y. Mazroui, and J. R. Gonzalez (2012). frailtypack: an r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software 47*(4), 1–28.

Rotolo, F. (2013). Frailty multi-state models for the analysis of survival data from multicenter clinical trials.

Rotolo, F. and C. Legrand (2012). Frailty multi-state models based on maximum penalized partial likelihood. In *46TH SCIENTIFIC MEETING OF THE ITALIAN STATISTICAL SOCIETY*.

Rotolo, F., C. Legrand, and I. Van Keilegom (2013). A simulation procedure based on copulas to generate clustered multi-state survival data. *Computer methods and programs in biomedicine 109*(3), 305–312.

Shih, J. H. and P. S. Albert (2010). Modeling familial association of ages at onset of disease in the presence of competing risk. *Biometrics 66*(4), 1012–1023.

Shih, J. H. and T. A. Louis (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 1384–1399.

Therneau, T. M. and P. M. Grambsch (2000). *Modeling survival data: extending the Cox model*. Springer Science & Business Media.

Therneau, T. M., P. M. Grambsch, and V. S. Pankratz (2003). Penalized survival models and frailty. *Journal of computational and graphical statistics 12*(1), 156–175.

Thomsen, S., D. Duffy, K. Kyvik, A. Skytthe, and V. Backer (2011). Relationship between type 1 diabetes and atopic diseases in a twin population. *Allergy 66*(5), 645–647.

Unverzagt, F. W., S. Gao, O. Baiyewu, A. O. Ogunniyi, O. Gureje, A. Perkins, C. Emsley, J. Dickens, R. Evans, B. Musick, et al. (2001). Prevalence of cognitive impairment data from the indianapolis study of health and aging. *Neurology 57*(9), 1655–1662.

van Houwelingen, H. and H. Putter (2011). *Dynamic prediction in clinical survival analysis*. CRC Press.

Vaupel, J. W., K. G. Manton, and E. Stallard (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography 16*(3), 439–454.

Wang, W. (2003). Estimating the association parameter for copula models under dependent censoring. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(1), 257–273.

Wei, L. (1992). The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine 11*(14-15), 1871–1879.

Wienke, A. (2010). *Frailty models in survival analysis*. CRC Press.

Wright, L., W. Simpson, R. J. Van Lieshout, and M. Steiner (2014). Depression and cardiovascular disease in women: is there a common immunological basis? a theoretical synthesis. *Therapeutic advances in cardiovascular disease 8*(2), 56–69.

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

Xue, Q.-L., L. P. Fried, T. A. Glass, A. Laffan, and P. H. Chaves (2008). Life-space constriction, development of frailty, and the competing risk of mortality the women's health and aging study i. *American journal of epidemiology 167*(2), 240–248.

Yashin, A. I., J. W. Vaupel, and I. A. Iachine (1995). Correlated individual frailty: an advantageous approach to survival analysis of bivariate data. *Mathematical Population Studies 5*(2), 145–159.

Zeng, D. and D. Lin (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(4), 507–564.

Zheng, M. and J. P. Klein (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika 82*(1), 127–138.

CURRICULUM VITAE

Ran Liao

EDUCATION

- Ph.D. in Biostatistics, Indiana University, Indianapolis, IN, 2017 (minor in Epidemiology)
- M.S. in Applied Mathematics, Beijing University of Aeronautics and Astronautics, Beijing, China, 2011
- B.S. in Applied mathematics, Xiamen University, Xiamen, China, 2008 (Dual B.S. in Economics)

WORKING EXPERIENCE

- Research Assistant, Indiana University, Indiana, U.S.A. Jan.2014-Mar.2016
- Teaching Assistant, Indiana University, Indiana, U.S.A. Aug.2012-Dec.2013
- Teaching Assistant, Beijing University of Aeronautics and Astronautics, China Aug.2008-Feb.2011

HONORS, AWARDS AND FELLOWSHIPS

- The 39th annual Midwest Biopharmaceutical Statistics Workshop, Best Poster Award, 2016
- The 39th annual Midwest Biopharmaceutical Statistics Workshop, Honor of Student Grant, 2016

- Indiana University School of Medicine, Public Health Graduate School Fellowship, 2011
- Beijing University of Aeronautics and Astronautics, Honor of Science and Technological Innovation Star, 2011
- National Graduate Mathematical Modeling Contest, 2nd Place, Beijing,China, 2010
- National undergraduate Mathematical Modeling Contest, 2nd Place, Beijing, China. 2008
- Xiamen University Outstanding Student scholarship, Xiamen, Fujian, China, 2004-2008

SELECT PUBLICATIONS

- **Liao, R.**, Hu, T., Gao, S., (2016) Covariate Dependent Cross Ratio Estimation between Bivariate Survival Time , *In submission to Statistical Method in Medical Research*
- **Liao, R.**, Gao, S., (2016) Frailty Models of Time to Event Data with Semi-competing Risk, *In progress*
- Liqin Su, Yinlong Jin, Frederick W. Unverzagt,Yibin Cheng, Ann M. Hake, Feng Ma, Jingyi Liu, Chen Chen, Jianchao Bian, Ping Li, **Liao, R.**, Sujuan Gao (2016) Nail Selenium Level and Diabetes in Rural Elderly Chinese, *Submitted to International Journal of Epidemiology*
- **Liao, R.**, Chen, D. (2010). Learning Rate of Support Vector Regression for Functional Data, *2010 Second International Conference on Future Computer and Communication.* Volume I (338-341), IEEE

- Xing, T., **Liao, R.** (2010). Multi-category Support Vector Machine Algorithm for Quality Analysis in Universities' Students Enrollment. *Journal of Harbin Institute of Technology (Nature Science),* 4(11): 1828-1832