Analysis of Time-to-Event Data with Multi-State Models and Causal Inference Methods

by

© **Yongho Lim**

A thesis submitted to the School of Gradate Studies in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Department of Mathematics and Statistics

Memorial University

May 2024

St. John's, Newfoundland and Labrador, Canada

Abstract

Analyses of disease-free survival data for certain cancer types indicate that cohorts of patients treated for cancer consist of individuals who are susceptible to experience cancer related events and individuals who are cured. Cured individuals do not experience any cancer related event, and eventually die due to other causes. Individuals who are not cured may die after experiencing cancer recurrence or without experiencing any recurrence. Cure status is a partially latent variable and is only known if a disease related event, cancer recurrence or cancer death, is observed. Causes of some observed deaths may be masked. To model disease progression events, which are cancer recurrence and cancer death, we consider a multi-state model including partially latent cured and not cured states. We describe our modeling approach and discuss an inference method incorporating masked causes of deaths. Our method allows us to identify factors associated with the risk of experiencing a disease related event and with timing of disease events after the treatment of cancer.

It is of interest to make inference on direct exposure effects on time-to-event outcomes in many studies. Traditional survival analysis methods may not reveal direct exposure effects on time-to-event outcomes when there are indirect exposure effects through intermediate variables which are confounded by some unmeasured factors. We propose a mediation analysis method to make inference about direct exposure effects on time-to-event outcomes under additive hazards model using estimating equations methodology. We examine properties of the proposed method and compare them with traditional survival analysis methods and the existing two-stage mediation analysis method which uses additive hazards model. The results show that our method provides valid inference about controlled direct exposure effects on time-to-event outcomes by successfully removing indirect effects through intermediate variables. It is robust against measured and unmeasured confounding of indirect effects.

To Yuna

Acknowledgements

First and above all, I would like to thank God for giving me this opportunity.

I would like to express my deepest appreciation to my supervisors, Dr. Yildiz Yilmaz and Dr. Candemir Cigsar, for giving me the opportunity to work with them over the past few years. Their support, encouragement, patience and guidance have been invaluable throughout this research work. I would like to appreciate members of my Ph.D. supervisory committee Dr. J Concepcion Loredo-Osti and Dr. Alwell Oyet for their invaluable guidance and support.

I wish to thank members of my thesis examining committee Dr. Xuewen Lu, Dr. Zhaozhi Fan and Dr. Yanqing Yi for their insightful comments.

I sincerely acknowledge the financial support provided by the School of Graduate Studies and the Department of Mathematics & Statistics, Memorial University of Newfoundland in the forms of Graduate Fellowship, Teaching Assistants, and Graduate Assistants.

It is my great pleasure to thank friends who encouraged and helped me during my Ph.D. program.

Finally, I wish to give special thank to my beautiful wife Yuna Oh and my daughters Raina Lim and Riley Lim for their support and love. I would not be able to complete this thesis without constant support, patience and love from my family.

Statement of contribution

Dr. Yildiz Yilmaz and Dr. Candemir Cigsar proposed the research questions that were investigated throughout this thesis. The overall study was jointly designed by Dr. Yildiz Yilmaz, Dr. Candemir Cigsar and Yongho Lim. The algorithms were implemented, the simulation studies were conducted and the manuscript was drafted by Yongho Lim. Dr. Yildiz Yilmaz and Dr. Candemir Cigsar jointly supervised the study and contributed to the final manuscript.

Table of contents

List of tables

List of figures

Chapter 1

Introduction

Multi-state modeling of multivariate survival times is widely used when subjects undergo a number of events in a given time period (Cook and Lawless, 2018). Multi-state modeling allows to model multiple time-to-events. It includes sequential time-to-events, competing risks, and illness-death modeling. Applications of multi-state models can be found in many studies such as cancer prognostic studies (Lawless and Yilmaz, 2011; Beesley and Taylor, 2018). Patients with cancer may experience one or more events after cancer treatment which includes cancer recurrence and death.

One may encounter challenges when modeling cancer progression events : (i) time from cancer diagnosis to cancer recurrence, time from recurrence to death and time from diagnosis to death could be subject to censoring, (ii) there could be dependence between sequential event times, (iii) there could be multiple causes of deaths, and a cause of death could compete with other causes, (iv) causes of deaths for some individuals may not be available when there are competing causes, (v) some individuals may not experience any cancer related event even though followup times are long enough. This indicates that the population is a mixture of individuals who are susceptible to a disease related event, which is recurrence or death due to cancer, and who are

long-term disease free survivors (cured).

The objective of the study is to introduce a multi-state model for cancer prognostic studies and to develop an estimation method to analyze the multi-state time-to-disease related events data. To model disease progression events, we consider a multi-state model with partially latent cured and not cured states. We describe our modeling approach and discuss an inference method incorporating masked causes of deaths. Our method allows us to identify factors associated with the risk of experiencing a disease related event and with timing of disease events after the treatment of cancer.

In this chapter, we give a brief review of notation terminology and statistical methods for univariate and multivariate survival time analyses. We focus on standard multi-state modeling without cured population in Chapter 1. In Chapter 2, the model is extended to include a cured fraction, cause specific deaths and masked causes of deaths.

1.1 Univariate Survival Data Analysis

In this section, we give a brief review of statistical modeling of univariate survival data. We let T be a continuous nonnegative time-to-event. The distribution function of T is defined by

$$F(t) = \Pr(T \leq t), \tag{1.1}$$

and the survival function of T is defined by

$$S(t) = \Pr(T > t). \tag{1.2}$$

We let Δt be an infinitesimal positive valued real number. The hazard function of random variable T is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(T \in [t, t + \Delta t) | T \geq t)}{\Delta t}, \qquad t > 0. \tag{1.3}$$

The function $h(t)\Delta t$ is an approximate probability of an event occurrence over $[t, t + \Delta t)$.

Suppose the time-to-event T_i for subject i is subject to right-censoring and the censoring time C_i and the time-to-event T_i are independent for $i = 1, 2, \ldots, n$. Let $t_i = \min(T_i, C_i)$ be the observed time-to-event and $\delta_i = I(T_i \leq C_i)$ be the event indicator, where $I(\cdot)$ is the indicator function.

Then, the likelihood function for the observed data $\{(t_i, \delta_i), i = 1, 2, \ldots, n\}$ is

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{(1-\delta_i)}, \tag{1.4}$$

where $f(t) = \partial F(t)/\partial t$.

1.2 Sequential Time-to-Events

Sequential events refer to a series of events that occur one after the other in order. Sequential times are the times between a specified series of sequentially observed events. They are also called gap times or sojourn times between sequential events. For example, cancer diagnosis, cancer recurrence and death can be considered as sequential events when there is no death before cancer recurrence.

Suppose T_1 and T_2 denote sequential time-to-events where T_2 cannot be observed unless T_1 is observed. T_1 and T_2 may not be independent. The joint distribution of

T_1 and T_2 is defined by

$$F(t_1, t_2) = \Pr(T_1 \leq t_1, T_2 \leq t_2), \tag{1.5}$$

and the joint survival function of T_1 and T_2 is defined by

$$S(t_1, t_2) = \Pr(T_1 > t_1, T_2 > t_2). \tag{1.6}$$

The joint survival function in (1.6) can be expressed as

$$S(t_1, t_2) = 1 - F_1(t_1) - F_2(t_2) + F(t_1, t_2), \tag{1.7}$$

where $F_1(t_1) = F(t_1, \infty)$ and $F_2(t_2) = F(\infty, t_2)$ are the marginal distribution functions of T_1 and T_2, respectively.

Let T_{1i}, T_{2i} be sequential time-to-events which are subject to right-censoring and C_i be the independent right-censoring time for individual i, $i = 1, 2, \ldots, n$. We let $t_{1i} = \min(T_{1i}, C_i)$ and $t_{2i} = \min(T_{2i}, C_i - t_{1i})$ be the observed sequential gap times and $\delta_{1i} = I(T_{1i} = t_{1i})$ and $\delta_{2i} = I(T_{2i} = t_{2i})$ be the event indicators. The observed data consists of $\{(t_{1i}, t_{2i}, \delta_{1i}, \delta_{2i}), i = 1, 2, \ldots, n\}$. The second time-to-event t_{2i} and its event indicator δ_{2i} only exist if $\delta_{1i} = 1$. To obtain the likelihood function of the observed data, we consider the likelihood contributions for subjects (i) who experienced both the first and the second event ($\delta_{1i} = 1$, $\delta_{2i} = 1$), (ii) who experienced the first event and then censored before experiencing the second event ($\delta_{1i} = 1$, $\delta_{2i} = 0$), (iii) who were censored before experiencing the first event ($\delta_{1i} = 0$). The likelihood of the observed data becomes (Lawless, 2003)

$$L = \prod_{i=1}^{n} \left[f(t_{1i}, t_{2i}) \right]^{\delta_{1i}\,\delta_{2i}} \left[-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{1i}} \right]^{\delta_{1i}(1-\delta_{2i})} \left[S_1(t_{1i}) \right]^{(1-\delta_{1i})}, \tag{1.8}$$

where $f(t_1, t_2) = \partial^2 F(t_1, t_2)/\partial t_1 \partial t_2$ is the joint probability density function of T_1 and T_2 and $S_1(t_1) = S(t_1, 0)$ is the marginal survival function of T_1.

There are three main ways to model the dependence between sequential event times. These are conditional modeling, random effects modeling, and marginal modeling approaches. Putter et al. (2006) used a semi-parametric Cox proportional hazard model where the gap times depend on the previous states by using "clock reset" approach which is to set time t to 0 at each state. Their hazard function of the second gap time was modeled conditionally dependent on the previous gap time. Fine et al. (2001) considered random effects modeling using a gamma frailty to model the dependence between successive gap times. A random effects model can be used to model the association of successive event times. In random effects modeling, successive event times are dependent through random components with a specified distribution (Cook and Lawless, 2007, Chapter 4). The marginal distributions in random effects models are not always in simple forms. He and Lawless (2003) considered marginal modeling approach in which they used a copula function to model the joint distribution of sequential event times.

We review the marginal approach to model the joint distribution by using copula functions. Copula models have attractive features when modeling the dependence structure between time-to-events. Copula modeling allows to consider different marginal distributions for time-to-events. The dependence structure in copula models does not depend on the marginal distributions of time-to-events. Therefore, marginal distributions can be modeled separately considering the features of each time-to-event. As a result, the interpretation of the dependence structure and the marginal distribution remain relatively simpler.

The joint distribution of sequential times T_1 and T_2 can be formed with a copula

function as

$$F(t_1, t_2) = \Pr(T_1 \leq t_1, T_2 \leq t_2) = C(F_1(t_1), F_2(t_2)), \tag{1.9}$$

where $C(.,.)$ is a copula function, $F_1(t_1)$ is the distribution function of T_1 and $F_2(t_2)$ is the distribution function of T_2. Based on Sklar's theorem, there exists a unique copula to construct the joint distribution of T_1 and T_2 if $F_1(t_1)$ and $F_2(t_2)$ are continuous (Nelsen, 2006, page 21).

We can re-write the likelihood function for sequential time-to-events in (1.8) by using a copula function as follows (He and Lawless, 2003; Lawless and Yilmaz, 2011)

$$\begin{aligned} L = \prod_{i=1}^{n} \left[\frac{\partial C(F_1(t_{1i}), F_2(t_{2i}))}{\partial t_{1i} \partial t_{2i}}\right]^{\delta_{1i}\,\delta_{2i}} \left[\frac{\partial F_1(t_{1i})}{\partial t_{1i}} - \frac{\partial C(F_1(t_{1i}), F_2(t_{2i}))}{\partial t_{1i}}\right]^{\delta_{1i}(1-\delta_{2i})} \\ \times \left[S_1(t_{1i})\right]^{(1-\delta_{1i})}. \end{aligned} \tag{1.10}$$

Fully parametric one-stage estimation using maximum likelihood estimation method or two-stage parametric estimation using pseudomaximum likelihood estimation method can be considered. A variety of copula families can be used with parametric estimation methods (Joe, 2014). In addition to parametric estimation, He and Lawless (2003) considered piecewise constant estimation to model the marginal hazard functions of the gap times while the joint distribution was modeled by a copula function. Later, Lawless and Yilmaz (2011) considered a semiparametric estimation method to estimate the joint distribution of sequential gap times considering a parametric copula function while the marginal distributions of the gap times were treated nonparametrically.

1.3 Competing Risks Model

Competing risks modeling is considered in many settings (Cook and Lawless, 2018, Chapter 3) where there are multiple possible outcomes or causes of an event that are of interest. For example, competing risks problem arises when we consider possible transitions from the healthy state to death state with different causes as shown in Figure 1.1. Causes of death compete with each other and death due to a cause prevents death due to the other causes.

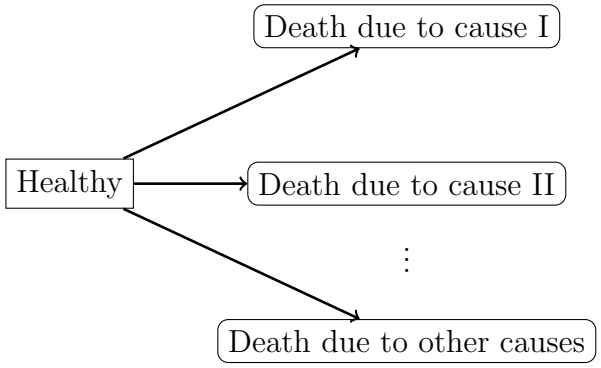


Figure 1.1: A competing risks structure

Prentice et al. (1978) stressed out the use of cause-specific hazards in competing risks modeling. Along with cause-specific hazards, cumulative incidence functions take an important role in the competing risks modeling (Kalbfleisch and Prentice, 2002). Nonparametric estimation in competing risks has been discussed by Aalen (1978) using the Nelson-Aalen estimation of the cumulative hazard function and the Kaplan-Meier estimation of the survival function. Because of the violation of the independence assumption in censoring, Kaplan-Meier estimator of overall survival function might be biased when cumulative incidence function is estimated (Putter et al., 2007). Bryant and Dignam (2004) suggested a semiparametric cumulative incidence estimator which considered Kaplan-Meier estimator of a survival function for a particular cause rather than the overall survival function.

Larson and Dinse (1985) considered decomposition of cumulative incidence function of time to failure to conditional subdistribution given the type of failure and the marginal probability of the type of failure. Nicolaie et al. (2010) considered a similar decomposition methodology to Larson and Dinse (1985) where they used decomposition of subdistribution of time to failure into distribution of time to failure and conditional probability of the type of failure given a failure occurred.

For parametric estimation of competing risks model, Larson and Dinse (1985) considered conditional subdistribution (cumulative incidence) functions and utilized parameterized marginal probability of cause. Jeong and Fine (2006) discussed a model with direct parametric assumption in cumulative incidence function using Gompertz distribution whose asymptote maybe less than one in certain settings. Fine and Gray (1999) proposed another approach using the subdistribution hazard which is the instantaneous risk of having a particular cause given that the subject has not experienced this particular cause.

An important function in competing risks is the cause specific hazard function. Suppose we have K distinct types of events. When an event occurs, it could be one of the K causes. We let M denote the type of events where $M = 1, 2, \ldots, K$. We denote the cause specific hazard function for type k as

$$\lambda_k(t) = \lim_{\Delta t \to 0} \frac{\Pr(T \in [t, t + \Delta t), M = k | T \geq t)}{\Delta t}, \qquad \text{for } k = 1, 2, \ldots, K. \tag{1.11}$$

We denote the marginal intensity function of T and the marginal cumulative intensity function of T as

$$\lambda(t) = \sum_{k=1}^{K} \lambda_k(t), \tag{1.12}$$

and

$$\Lambda(t) = \sum_{k=1}^{K} \Lambda_k(t), \tag{1.13}$$

respectively, where $\Lambda_k(t) = \int_0^t \lambda_k(s)ds$. We denote the cumulative incidence function for type k as

$$F_k(t) = \Pr(T \leq t, M = k) = \int_0^t \lambda_k(u)S(u)du, \tag{1.14}$$

where $S(t) = \exp(-\Lambda(t))$ is the marginal survival function of T.

Let T_i and C_i be the time-to-event and the independent right-censoring time for the ith individual, respectively. We have the observed data as $\{(t_i, \delta_{1i}, \delta_{2i}, \ldots, \delta_{Ki},), i = 1, 2, \ldots, n\}$ where $t_i = \min(T_i, C_i)$ and $\delta_{ki} = I[\text{Cause } k \text{ occurs for the } i\text{th individual}]$, $k = 1, 2, \ldots, K$. The likelihood function for the observed data $\{(t_i, \delta_{1i}, \delta_{2i}, \ldots, \delta_{Ki},), i = 1, 2, \ldots, n\}$ is proportional to

$$L = \prod_{i=1}^{n} \left[\prod_{k=1}^{K} \left(\frac{\partial F_k(t_i)}{\partial t_i} \right)^{\delta_{ki}} \right] S(t_i)^{\prod_{k=1}^{K}(1-\delta_{ki})}. \tag{1.15}$$

1.4 Illness-Death Model

Illness-death model or semi-competing risks model can be used to model cancer progression events. Patients who were diagnosed for cancer and became disease free may have a cancer recurrence or may die without experiencing a recurrence. Patients who had recurrence may die during followup time period. In illness-death model, there are terminal and non-terminal events where a terminal event terminates the process while a non-terminal event is an event that does not terminate the process. For example, cancer recurrence refers to a non-terminal event and death refers to a terminal event in the illness-death model. Figure 1.2 describes an illness death model structure where there are healthy, cancer recurrence and death states.

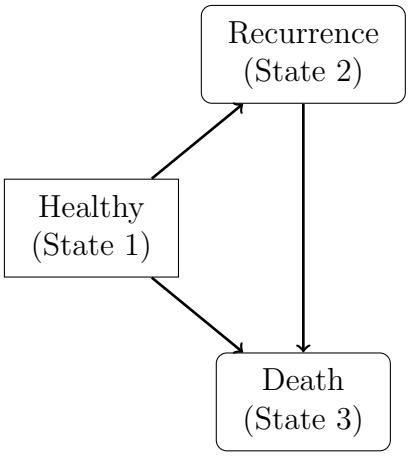


Figure 1.2: An illness-death model structure

Illness-death model was first introduced by Fix and Neyman (1951). Markov illness-death model was discussed in detail in Andersen et al. (1993). Semi-Markov illness death model was discussed by Voelkel and Crowley (1984) considering hazards for a particular state depending on the previous state. In semi-Markov illness death model, the dependence between sequentially observed event times in semi-competing risks needs to be taken into account.

Let T_{1i} and T_{2i} be the time-to-first event and time from recurrence to death, respectively for individual i, $i = 1, 2, \ldots, n$. Time-to-first event is the time until cancer recurrence or cancer death without experiencing cancer recurrence. Let C_i be independent right-censoring time for individual i, $i = 1, 2, \ldots, n$. We let $t_{1i} = \min(T_{1i}, C_i)$ and $t_{2i} = \min(T_{2i}, C_i - t_{1i})$ for $i = 1, 2, \ldots, n$. We let $\delta_{jki} = I[\text{transition } j \to k \text{ occurs for individual } i]$ be event indicators with $j = 1$ and $k = 2, 3$ or $j = 2$ and $k = 3$ for $i = 1, 2, \ldots, n$ and M denote the progression state at t_1 for $M = 2, 3$.

We denote the joint distribution function of time to recurrence and time from recurrence to death by $F_{12,23}(t_1, t_2) = \Pr(T_1 \leq t_1, T_2 \leq t_2, M = 2)$. The survival

function of time to recurrence and time from recurrence to death is denoted by

$$S_{12,23}(t_1, t_2) = \Pr(T_1 > t_1, T_2 > t_2, M = 2). \tag{1.16}$$

We denote $F_{13}(t_1) = Pr(T_1 \leq t_1, M = 3)$ as the cumulative incidence function of T_1 for subjects who died without experiencing recurrence.

The observed data is $\{(t_{1i}, t_{2i}, \delta_{12i}, \delta_{13i}, \delta_{23i}, i = 1, 2, \ldots, n\}$. The second gap time t_{2i} and its event indicator δ_{23i} only exist if $\delta_{12i} = 1$. Due to the nature of competing risks, if $\delta_{12i} = 1$, then $\delta_{13i} = 0$; and if $\delta_{13i} = 1$, then $\delta_{12i} = 0$. To obtain the likelihood function of the observed data, we consider the likelihood contributions for subjects (i) who have censored time-to-first event ($\delta_{12i} = 0$, $\delta_{13i} = 0$), (ii) who have experienced cancer recurrence and died ($\delta_{12i} = 1$, $\delta_{23i} = 1$), (iii) who have experienced cancer recurrence and not died before the end of followup ($\delta_{12i} = 1$, $\delta_{23i} = 0$), and (iv) who died without experiencing recurrence ($\delta_{13i} = 1$).

Then, the likelihood function for the observed data $\{(t_{1i}, t_{2i}, \delta_{12i}, \delta_{13i}, \delta_{23i}), i = 1, 2, \ldots, n\}$ is proportional to

$$\begin{aligned} L = \prod_{i=1}^{n} \left(f_{12,23}(t_{1i}, t_{2i})\right)^{\delta_{12i}\,\delta_{23i}} \left(-\frac{\partial S_{12,23}(t_{1i}, t_{2i})}{\partial t_{1i}}\right)^{\delta_{12i}(1-\delta_{23i})} \\ \times \left(\frac{\partial F_{13}(t_{1i})}{\partial t_{1i}}\right)^{\delta_{13i}} \left(S_1(t_{1i})\right)^{(1-\delta_{12i})(1-\delta_{13i})}, \end{aligned} \tag{1.17}$$

where $f_{12,23}(t_1, t_2) = \partial^2 F_{12,23}(t_1, t_2)/\partial t_1 \partial t_2$ is the joint probability density function of time to recurrence and time from recurrence to cancer death and $S_1(t_1) = \Pr(T_1 > t_1)$ is the marginal survival function of T_1.

Parametric maximum likelihood estimation under illness-death model is well listed in Andersen et al. (1993, Chapter 6). Royston and Parmar (2002) proposed a parametric estimation method under illness-death model using cubic spline function to

model the baseline log cumulative hazard function. Xu et al. (2010) considered an illness–death model with nonparametric maximum likelihood estimation method.

1.5 Multi-State Modeling Terminology

We let T be life time. We consider a multi-state process with K states. We let $N_{jk}(t)$ be the number of event occurrences in transition from state j to k over a time interval $(0, t]$, where $t > 0$ and $j \neq k = 1, 2, \ldots, K$. Then, $\{\mathbf{N}(t), t \geq 0\}$ is called a counting process where $\mathbf{N}(t) = (N_{jk}(t), j \neq k = 1, 2, \ldots, K)^T$. The history at time t, $\mathcal{H}(t) = \{\mathbf{N}(u), 0 \leq u \leq t\}$, includes all the information about the counting process $\{\mathbf{N}(t), t \geq 0\}$ from time 0 to time t.

We let Δt be an infinitesimal positive valued real number and $\Delta N_{jk}(t)$ be an increment in transition from state j to k over a small interval $[t, t + \Delta t)$ which gives the number of event occurrences in transition from state j to k over the time interval $[t, t + \Delta t)$. We define the intensity function for the transition from state j to k as

$$\lambda_{jk}(t|\mathcal{H}(t^-)) = \lim_{\Delta t \to 0} \frac{P\{\Delta N_{jk}(t) = 1|\mathcal{H}(t^-)\}}{\Delta t} \qquad \text{for } j \neq k, \tag{1.18}$$

where $\Delta t > 0$. We assume that two or more events cannot occur simultaneously at the same instant.

Using the multi-state model terminology, we can also define the intensity function. We let $Z(t)$ denote the state occupied at time t. Then, $\{\mathbf{Z}(t), t \geq 0\}$ is associated with a stochastic process. The history for the process is denoted by $\mathcal{H}(t) = \{\mathbf{Z}(u), 0 \leq u \leq t\}$. We define $P_{jk}(s, t) = \Pr(Z(t) = k|Z(s) = j, \mathcal{H}(s^-))$ for $s < t$ be the probability of being in state k at time t given that the process was in state j at time s and $\mathcal{H}(s^-)$ denotes the history of events and covariate information up to and right before time

s. The multi-state process is characterized through transition intensities

$$\lambda_{jk}(t|\mathcal{H}(t^-)) = \lim_{\Delta t \to 0} \frac{P_{jk}(t, t+\Delta t)}{\Delta t} \qquad \text{for } j \neq k = 1, 2, \ldots, K. \tag{1.19}$$

We assume that all the individuals are observed from time 0 and independent right-censoring time is C_i for $i = 1, 2, \ldots, n$. We define $Y_{ij}(t) = I(Z_i(t) = j)$ as the indicator that an individual i is in state j at time t. The likelihood function is given by

$$L = \prod_{j \neq k} L_{jk}, \tag{1.20}$$

with

$$L_{jk} = \prod_{i=1}^{n} \left\{ \prod_{t_{ir} \in D_{ijk}} \lambda_{jk}(t_{ir}) \exp\left(- \int_0^\infty \bar{Y}_{ij}(u) \lambda_{jk}(u) du \right) \right\}, \tag{1.21}$$

where $\bar{Y}_{ij}(u) = I(0 \leq u \leq C_i)\, Y_{ij}(u)$ and D_{ijk} is the set of $j \to k$ transition times observed over the interval $[0, C_i]$.

1.6 Likelihood Inference

We let $y_1, y_2, \ldots, y_n$ be a random sample which are from a distribution function $f(y; \boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$ is a vector of unknown parameters for $\boldsymbol{\theta} \in \Omega$. The likelihood function of observed data $\{y_1, y_2, \ldots, y_n\}$ becomes

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}). \tag{1.22}$$

Assuming the required regularity conditions (Cox and Hinkley, 1979) are satisfied, we let $U(\boldsymbol{\theta}) = (U_1(\boldsymbol{\theta}), U_2(\boldsymbol{\theta}), \ldots, U_p(\boldsymbol{\theta}))^T$ be a score vector where

$$U_j(\boldsymbol{\theta}) = \frac{\partial \log L(\boldsymbol{\theta})}{\partial \theta_j}, \quad j = 1, 2, \ldots, p \tag{1.23}$$

and maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ can be obtained by solving

$$U(\boldsymbol{\theta}) = \mathbf{0}. \tag{1.24}$$

Assuming that the model is correctly specified and the required regularity conditions are satisfied, $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$ and

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N_p(\mathbf{0}, \mathcal{J}_1{}^{-1}(\boldsymbol{\theta})), \tag{1.25}$$

where $\mathcal{J}_1(\boldsymbol{\theta})$ is the Fisher information matrix with entries

$$\mathcal{J}_{1,jk}(\boldsymbol{\theta}) = \frac{1}{n} E\left(-\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial\theta_j \partial\theta_k}\right) \qquad j,k = 1,2,\ldots,p. \tag{1.26}$$

The observed information matrix $\mathcal{I}(\hat{\boldsymbol{\theta}})$ is a consistent estimator of $\mathcal{J}(\boldsymbol{\theta}) = n\mathcal{J}_1(\boldsymbol{\theta})$ where the entries of $\mathcal{I}(\boldsymbol{\theta})$ are

$$\mathcal{I}_{jk}(\boldsymbol{\theta}) = -\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial\theta_j \partial\theta_k} \qquad j,k = 1,2,\ldots,p. \tag{1.27}$$

1.7 Competing Risks with Masked Causes

In classic competing risk model, causes of failures are all assumed to be known. However, it is not always possible to know causes of all failures.

Competing risks modeling with masked causes has been discussed by many authors. Dinse (1986) considered nonparametric maximum likelihood estimation when there are missing causes of failures. Two-stage data collection was considered by Flehinger et al. (1998, 2002) to obtain information on some masked causes. They considered maximum likelihood estimation for cause specific hazards functions using

an EM algorithm. Estimation under a piecewise-constant hazards competing risks model with two-stage data collection was proposed by Craiu and Duchesne (2004). Basu and Tiwari (2010) considered Bayesian estimation method under mixture-cure model with competing risks to tackle masked causes problem.

Suppose there are $K-1$ competing risk events with $k \geq 2$ and G masking groups with $G \geq 1$. Since masking groups can be subsets of all causes, we can define masking groups as

$$\mathcal{G} = \{g_1, g_2, \ldots, g_G; g_q \subset \{2, 3, \ldots, K\}, q = 1, 2, \ldots, G\}, \tag{1.28}$$

where masking groups g_q, for $q = 1, 2, \ldots, G$ contains 2 or more causes that are subsets of $\{2, 3, \ldots, K\}$. The ith individual has complete data $(t_i, \gamma_{ig_1}, \gamma_{ig_2}, \ldots, \gamma_{ig_G}, \delta_{12i}, \delta_{13i}, \ldots, \delta_{1Ki})$ where γ_{ig_q}, for $q = 1, 2, \ldots, G$, is an indicator that the cause of failure of ith individual is masked to group g_q for $q = 1, 2, \ldots, G$ in the first stage data. If any γ_{ig_q} for $q = 1, 2, \ldots, G$ is 1 for ith individual, then $\delta_{12i}, \delta_{13i}, \ldots, \delta_{1Ki}$ are missing. The masked probability becomes (Craiu and Duchesne, 2004)

$$P_{g_q|k}(t) = \Pr(\text{cause masked to group } g_q \text{ in the first stage } |T = t, M = k), \tag{1.29}$$

for $q = 1, 2, \ldots, G$ and $k = 2, 3, \ldots, K$. Also, the diagnostic probability is defined by

$$\pi_{k|g_q}(t) = \Pr(\text{actual cause of failure is } k|\text{cause masked to group } g_q \text{ and failed at time } t). \tag{1.30}$$

The diagnostic probability can be obtained by using Bayes' rule

$$\pi_{k|g}(t) = \frac{\lambda_{1k}(t)\, P_{g|k}(t)}{\sum_{l \in g} \lambda_{1l}(t) P_{g|l}(t)}. \tag{1.31}$$

Let $\mathcal{G}_k^*$ be the set of masked groups that contains cause k. Then, the likelihood function under the complete data is

$$\begin{aligned} L = & \prod_{i=1}^{n} \prod_{k=2}^{K} \left(\frac{\partial F_{1k}(t_i)}{\partial t_i} \right)^{\delta_{1ki}} S(t_i)^{\prod_{k=2}^{K}(1-\delta_{1ki})} \\ & \times \prod_{g \in \mathcal{G}_k^*} P_{g|k}(t)^{\delta_{1ki}\gamma_{ig}} \prod_{g \in \mathcal{G}_k^*} (1 - P_{g|k}(t))^{\delta_{1ki}(1-\sum_{g \in \mathcal{G}_k^*} \gamma_{ig})}. \end{aligned} \tag{1.32}$$

Note that δ_{1ki} follows multinomial distribution with size 1 and probabilities $\pi_{k|g}(t_i)$, for $k \in g$. Therefore,

$$E\left[\delta_{1ki}|\text{Obs}\right] = \pi_{k|g}(t_i) = \frac{\lambda_{1k}(t)\, P_{g_{q|k}}(t)}{\sum_{l \in g} \lambda_{1l}(t) P_{g|l}(t)}. \tag{1.33}$$

Therefore, the E step consists of substituting δ_{1ki} by $E\left[\delta_{1ki}|\text{Obs}\right]$ in the natural logarithm of the likelihood given in (1.32) and the conditional expectation is maximized in the M step.

1.7.1 EM algorithm

Expectation-Maximization (EM) algorithm is useful in finding maximum likelihood estimates when some of the data are missing. Dempster et al. (1977) popularized the EM algorithm. It has been extensively used in various studies. The details can also be found in McLachlan and Krishnan (2007).

We let $\boldsymbol{Y}$ be the random vector corresponding to observed data $\boldsymbol{y}$ with pdf $f(\boldsymbol{y}|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the vector of unknown parameters. We let $\boldsymbol{X}$ be the random vector corresponding to the complete data $\boldsymbol{x}$ with pdf $f_c(\boldsymbol{x}|\boldsymbol{\theta})$. We only observe incomplete data $\boldsymbol{y}$ instead of the complete data $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{z})$ where $\boldsymbol{Z}$ is the random vector corresponding to the missing data $\boldsymbol{z}$ whose pdf is $k(\boldsymbol{z}|\boldsymbol{\theta})$.

The complete likelihood function is defined as

$$L_c(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{z}) = f_c(\boldsymbol{x}|\boldsymbol{\theta}), \tag{1.34}$$

and the observed likelihood function is $L(\boldsymbol{\theta}|\boldsymbol{y})$. Our goal is to maximize $L(\boldsymbol{\theta}|\boldsymbol{y})$. We let $Q(\boldsymbol{\theta}|\boldsymbol{\theta_0},\boldsymbol{y})$ be

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta_0},\boldsymbol{y}) = E_{\boldsymbol{\theta_0}}[logL_c(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{z})|\boldsymbol{\theta_0},\boldsymbol{y}], \tag{1.35}$$

where $\boldsymbol{\theta_0}$ is given value of $\boldsymbol{\theta}$.

We maximize $L(\boldsymbol{\theta}|\boldsymbol{y})$ by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta_0},\boldsymbol{y})$. Since the conditional distribution of $\boldsymbol{z}$ given $\boldsymbol{Y} = \boldsymbol{y}$ is

$$k(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}) = \frac{f_c(\boldsymbol{z},\boldsymbol{y}|\boldsymbol{\theta})}{f(\boldsymbol{y}|\boldsymbol{\theta})}, \tag{1.36}$$

we can write $L(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta}) = f_c(\boldsymbol{z},\boldsymbol{y}|\boldsymbol{\theta})/k(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta})$. Thus, we obtain

$$\begin{aligned}
\log L(\boldsymbol{\theta}|\boldsymbol{y}) &= \int \log L(\boldsymbol{\theta}|\boldsymbol{y}) k(\boldsymbol{z}|\boldsymbol{\theta_0},\boldsymbol{y}) d\boldsymbol{z} \\
&= \int \left[\log f_c(\boldsymbol{z},\boldsymbol{y}|\boldsymbol{\theta}) - \log k(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{y})\right] k(\boldsymbol{z}|\boldsymbol{\theta_0},\boldsymbol{y}) d\boldsymbol{z} \\
&= \int \left[\log L_c(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{z}) - \log k(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{y})\right] k(\boldsymbol{z}|\boldsymbol{\theta_0},\boldsymbol{y}) d\boldsymbol{z} \\
&= E_{\boldsymbol{\theta_0}}[logL_c(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{z})|\boldsymbol{\theta_0},\boldsymbol{y}] - E_{\boldsymbol{\theta_0}}[\log k(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{y})|\boldsymbol{\theta_0},\boldsymbol{y}] \\
&= Q(\boldsymbol{\theta}|\boldsymbol{\theta_0},\boldsymbol{y}) - E_{\boldsymbol{\theta_0}}[\log k(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{y})|\boldsymbol{\theta_0},\boldsymbol{y}]
\end{aligned} \tag{1.37}$$

Therefore, the EM algorithm consists of finding the expectation in (1.35) and maximizing it. We let $\boldsymbol{\theta}^{(0)}$ be some initial value for $\boldsymbol{\theta}$. The E-step and M-step are repeated to find $\boldsymbol{\theta}^{(l+1)}$ until the difference $L(\boldsymbol{\theta}^{(l+1)}) - L(\boldsymbol{\theta}^{(l)})$ meets the convergence criterion.

1. E-step (Expectation Step)

Calculate

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}, \boldsymbol{y}) = E_{\boldsymbol{\theta_0}}[logL_c(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z})|\boldsymbol{\theta}^{(l)}, \boldsymbol{y}]. \tag{1.38}$$

2. M-step (Maximization Step)

   Find $\boldsymbol{\theta}^{(l+1)}$ by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}, \boldsymbol{y})$

$$\boldsymbol{\theta}^{(l+1)} = Argmax\, Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}, \boldsymbol{y}). \tag{1.39}$$

1.7.2 Estimation of Variance of $\hat{\boldsymbol{\theta}}$ in EM Algorithm

When the EM algorithm is used, a special attention is needed to obtain the standard error estimates of $\hat{\boldsymbol{\theta}}$ (Xu et al., 2014). There are several methods to calculate standard error estimates of estimators obtained by EM algorithm. In this study, we focus on the supplemented EM (SEM) algorithm and the nonparametric bootstrap.

SEM algorithm

SEM is useful to calculate standard error estimates with extra variability due to EM procedure (Meng and Rubin, 1991; Xu et al., 2014). EM algorithm uses the relation

$$\boldsymbol{\theta}^{(l+1)} = M(\boldsymbol{\theta}^{(l)}), \tag{1.40}$$

where $\boldsymbol{\theta}^{(l)}$ is the estimate of unknown parameter $\boldsymbol{\theta}$ from the lth iteration and $M(\boldsymbol{\theta}^{(l)})$ is the value of $\boldsymbol{\theta}^{(l)}$ that maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}, \boldsymbol{y})$. Then, the variance estimates can be obtained by (Meng and Rubin, 1991)

$$V = I_{oc}^{-1} + I_{oc}^{-1}\, DM(I - DM)^{-1}, \tag{1.41}$$

where I_{oc} is a complete information matrix obtained from

$$I_{oc} = E\left(-\frac{\partial^2 log f(\boldsymbol{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\middle|\boldsymbol{y},\boldsymbol{\theta}\right)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \tag{1.42}$$

and DM has entries

$$r_{ij} = \left(\frac{\partial M_j(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i}\right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \tag{1.43}$$

where r_{ij} can be estimated below. We let $\tilde{\boldsymbol{\theta}}_i^{(l)}$ to be

$$\tilde{\boldsymbol{\theta}}^{(l)}(i) = (\theta_1^*, \theta_2^*, \ldots, \theta_{i-1}^*, \theta_i^{(l)}, \theta_{i+1}^*, \ldots, \theta_p^*), \tag{1.44}$$

where θ_i^* is the maximum likelihood estimator of θ_i for $i = 1, 2, \ldots, p$. (1.44) means that only the ith component is in the lth process of the algorithm while other components are maximum likelihood estimators of θ_i. Meng and Rubin (1991) showed that r_{ij} can be estimated by the following procedure

1. Obtain θ_i^* for $i = 1, 2, \ldots, p$ where θ_i^* is the maximum likelihood estimator of θ_i using the EM algorithm.

2. Obtain $\boldsymbol{\theta}_i^{(l+1)}(i)$ by treating $\tilde{\boldsymbol{\theta}}^{(l)}(i)$ in (1.44) as current iteration and running one more EM iteration.

3. Obtain the ratio
$$r_{ij}^{(l)} = \frac{\theta_j^{(l+1)}(i) - \theta_j^*}{\theta_i^{(l)} - \theta_i^*}. \tag{1.45}$$

4. Obtain r_{ij} until $r_{ij}^{(l)}$ is stable.

Bootstrap Procedure for EM Algorithm

Nonparametric bootstrap procedure can be used to obtain standard error estimates of estimators. The procedure is as follows:

(i) Obtain a random sample $\{(\tilde{t}_{1r}, \tilde{t}_{2r}, \tilde{\delta}_{12r}, \tilde{\delta}_{13r}, \tilde{\delta}_{23r}, \tilde{\gamma}_{g_m r}), r = 1, 2, \ldots, n\}$ with replacement from the data $\{(t_{1i}, t_{2i}, \delta_{12i}, \delta_{13i}, \delta_{23i}, \gamma_{g_m i}), i = 1, 2, \ldots, n\}$.

(ii) Using the E step and the M step, obtain $\hat{\boldsymbol{\theta}}$ which is the estimate of unknown parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ using the data $\{(\tilde{t}_{1r}, \tilde{t}_{2r}, \tilde{\delta}_{12r}, \tilde{\delta}_{13r}, \tilde{\delta}_{23r}, \tilde{\gamma}_{g_m r}), r = 1, 2, \ldots, n\}$.

(iii) Repeat the steps (i), (ii) B times and obtain estimates $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \ldots, \hat{\boldsymbol{\theta}}_B$.

(iv) Bootstrap variance-covariance matrix is

$$\frac{1}{B}\sum_{r}^{B}(\hat{\boldsymbol{\theta}}_r - \overline{\hat{\boldsymbol{\theta}}})(\hat{\boldsymbol{\theta}}_r - \overline{\hat{\boldsymbol{\theta}}})^T, \tag{1.46}$$

where $\overline{\hat{\boldsymbol{\theta}}}$ is the mean of $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \ldots, \hat{\boldsymbol{\theta}}_B$.

1.8 Outline of Research

In this chapter, we introduced the notation and some models used for univariate and multivariate survival data analysis. In some types of cancer, such as breast, leukemia, and colorectal cancer, a proportion of patients may not experience any cancer-related events after treatment in a long followup time. These patients are considered as statistically cured. Cured individuals do not experience any cancer-related events, and eventually die due to other causes unrelated to cancer. On the other hand, individuals who are not cured may die due to cancer either after experiencing cancer recurrence or

without experiencing cancer recurrence. Cure status is a partially latent variable and is only known if a disease-related event, cancer recurrence or cancer death, is observed. Moreover, cause of death for some individuals may not be immediately observed and may be masked. To model disease progression events, we consider a multi-state model including partially latent cured and not cured states. Our modeling approach provides a new method to model curable cancer progression data on time to cancer recurrence and cancer death in the presence of masked causes of deaths. We study our modeling approach and discuss an inference method incorporating masked causes of deaths in Chapter 2.

In Chapter 3, we study the second topic which is about making inference on direct exposure effects on time-to-event outcomes. Traditional survival analysis methods may not reveal direct exposure effects on time-to-event variables when there are indirect exposure effects through intermediate variables which are confounded by some unmeasured factors. We propose a mediation analysis method to make inference about direct exposure effects on time-to-event outcomes under additive hazards model using estimating equations methodology. We follow the sequential G-estimation idea but consider a one-stage estimation solving two sets of unbiased estimating equations simultaneously. The direct exposure effect is measured using the adjusted time-to-event outcomes obtained by removing indirect exposure effect. We examine the properties of the proposed method and compare them with traditional survival analysis methods and the existing two-stage causal inference method which uses additive hazards model. In Chapter 3, we introduce the notation and some existing modeling approaches and estimation methods for regression analysis with time-to-event data and mediation analysis. We propose our mediation analysis method to infer the controlled direct exposure effect on time-to-event outcome in Chapter 4. Our one stage estimating equation approach gives a closed-form estimator for the direct exposure effect and

allows to use the robust Huber-White sandwich estimator of the standard error of direct exposure effect estimator.

In Chapter 5, we summarize our study and discuss future research.

Chapter 2

Statistical Inference in Multi-State Semi-Markov Models with a Cured Fraction

2.1 Introduction

There have been significant advances in treatment options for certain cancer types and stages. Successful treatments lead to improvements in survival and disease-free life of cancer patients. In many observational studies with long followup time, substantial proportions of individuals receiving cancer treatment do not experience any cancer related events, cancer recurrence or death due to cancer. These individuals are possibly not susceptible to any cancer events after receiving their cancer treatment and they eventually die due to other reasons rather than their diagnosed cancer. We call treated individuals who are not susceptible to any cancer events as "cured". Individuals who are susceptible to cancer related events may experience cancer recurrence and then die or may die due to cancer without experiencing recurrence. Cancer

death without recurrence might be observed as a result of an adverse effect of cancer treatment (Dillekås et al., 2019), and in this study it is considered as a death due to cancer.

Figure 2.1 illustrates a multi-state model for cancer progression events. The treatment state (state 1) represents patients who have been treated for their primary cancer. It includes both "cured" and "not cured" states. Being cured or not is partially latent. Patients who had cancer recurrence or died due to cancer are known to be not cured. The possible transitions for non-cured patients are from cancer treatment (state 1) to cancer recurrence (state 2) and then from recurrence (state 2) to cancer death (state 3) or from treatment to cancer death (state 3) without experiencing recurrence. There is an illness-death model for non-cured patients. Cancer death without recurrence precludes cancer recurrence for non-cured patients who received cancer treatment. Patients who have cancer recurrence may die due to cancer. Cured individuals do not experience any disease related event in a long term followup and eventually die due to other causes. Cure status is unknown if no cancer related event is observed, and there can only be empirical evidence for being cured if no cancer related event is observed during a long followup time (Lambert et al., 2010).

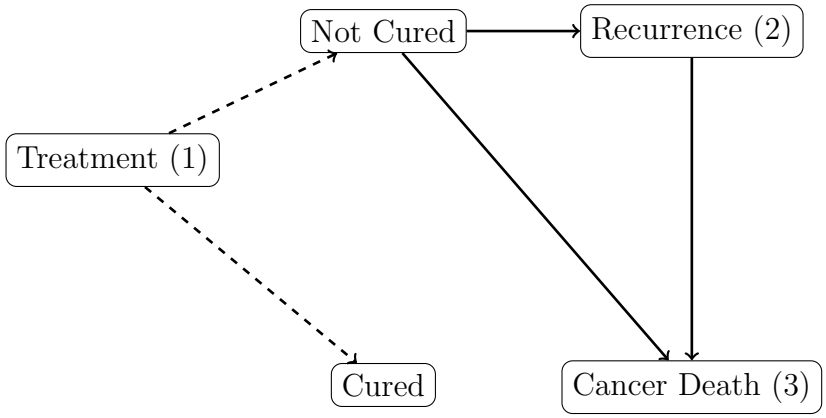


Figure 2.1: Multi-state model with partially latent cured and not cured states

The objective of our study is to model the risk of experiencing a cancer related

event and distributions of time to cancer recurrence, time from recurrence to cancer death and time to cancer death without experiencing cancer recurrence.

When a study cohort consists of a mixture of long-term cancer free survivors and non-cured patients, mixture cure modeling is helpful to identify factors associated with the risk and timing of a cancer progression event (Yilmaz et al., 2013). Mixture cure models have been used in various multiple modes of failure settings. It was used in a multi-state modeling setting with a cured fraction in Conlon et al. (2014) and Beesley and Taylor (2018) with an all-cause mortality state where causes of deaths are ignored. Conlon et al. (2014) considered a semi-Markov multi-state cure model. They assumed fully parametric models to model the probability of being cured and transition intensities. They used conditional modeling approach to incorporate the effect of time from treatment to recurrence on time from recurrence to death to fulfill the semi-Markov assumption and proposed a Bayesian estimation method. Beesley and Taylor (2018) also considered a multi-state cure model with an all-cause mortality state but had the Markov assumption for successive time-to-events. They conducted maximum likelihood estimation using a Monte Carlo EM algorithm.

In this study, we consider the multi-state cure model in Figure 2.1 with a cause specific death state to suit the objective of the study described above. We use mixture-cure model to incorporate cured fraction. For non-cured patients, the successive time-to-events, time to recurrence and time from recurrence to cancer death, can be associated for a given patient. Thus, we work under the semi-Markov assumption. We consider the marginal modeling approach using copula functions to model the dependence between time to recurrence and time from recurrence to cancer death. Copula modeling allows to model marginal distributions of time-to-events separately from the dependence structure (Nelsen, 2006; Joe, 2014). Therefore, the marginal distributions can be selected based on modeling needs and can be combined using

a copula function to obtain the joint distribution of the sequential gap times. The marginal modeling allows the effect of covariates to be easily interpreted compared to the conditional modeling (Cook and Lawless, 2007, Chapter 4). We obtain the likelihood function under this semi-Markov multi-state model including partially latent cured and not cured states and discuss the likelihood and pseudo-likelihood based inference methods in Section 2.2.

Cure status is a partially latent variable and is only known if a disease related event, cancer recurrence or cancer death, is observed. In addition to unknown cure status, the true causes of some observed deaths may be masked. A death can be due to the diagnosed cancer or due to other causes. When there is no cured proportion, under a competing risks model with masked causes of failure, Craiu and Duchesne (2004) proposed a design and an EM algorithm to estimate cause specific hazard functions. They considered a two-stage data collection design where the true causes of some masked causes of failures are obtained in the second stage. In our setting, we use the empirical evidence on cure status for the masked causes of deaths which occurred after the last observed cancer related event time. We propose an EM algorithm in Section 2.3 to fit the multi-state cure model while incorporating the empirical evidence on cure status.

The remainder of this Chapter is organized as follows. In Section 2.2, we introduce the semi-Markov multi-state cure model with cause-specific cancer death state and discusses the likelihood and pseudo-likelihood methods to fit the model. In Section 2.3, we propose the EM algorithm to fit the model in the presence of masked causes of deaths. We describe the simulation studies performed to investigate the properties of the estimation methods and presents the results in Section 2.4. In Section 2.5, we provide an application of a real data using the proposed estimation methods.

2.2 Modeling and Estimation

Suppose T_{1i} denotes the first disease progression event time, T_{2i} denotes time from recurrence to cancer death for patients who experience cancer recurrence, and C_i denotes the right censoring time for individual i, $i = 1, 2, \ldots, n$. Let $t_{1i} = \min(T_{1i}, C_i)$ and $t_{2i} = \min(T_{2i}, C_i - t_{1i})$ be the observed time-to-events and $\delta_{jki} = I[\text{transition } j \to k \text{ occurs for individual } i]$ be the event indicator from state $j = 1$ to state $k = 2$ or 3 or from state $j = 2$ to state $k = 3$ for $i = 1, 2, \ldots, n$. Suppose $M = k$ denotes the first disease progression state k $(k = 2, 3)$.

When there are no masked causes of deaths, the observed data under the model in Figure 2.1 is $\{(t_{1i}, t_{2i}, \delta_{12i}, \delta_{13i}, \delta_{23i}), i = 1, 2, \ldots, n\}$. The second gap time t_{2i} and its event indicator δ_{23i} only exist if $\delta_{12i} = 1$. Due to the nature of how competing risks occur, if $\delta_{12i} = 1$, then $\delta_{13i} = 0$ and if $\delta_{13i} = 1$, then $\delta_{12i} = 0$. To obtain the likelihood function of the observed data, we consider the likelihood contributions for subjects (i) who have censored time-to-first event ($\delta_{12i} = 0$, $\delta_{13i} = 0$), (ii) who have experienced cancer recurrence and died due to cancer ($\delta_{12i} = 1$, $\delta_{23i} = 1$), (iii) who have experienced cancer recurrence and not died due to cancer before the end of followup ($\delta_{12i} = 1$, $\delta_{23i} = 0$), and (iv) who died due to cancer without experiencing recurrence ($\delta_{13i} = 1$).

Since there is a substantial proportion of long-term cancer-free survivors after their cancer treatment, we use a mixture cure model to model the marginal survival function of the first disease progression event time T_1. The first disease progression event time is the time until cancer recurrence or death due to cancer. Thus, the survival function of T_1 can be written in the form of

$$S_1(t_1) = \Pr(T_1 > t_1) = pS_{10}(t_1) + 1 - p \tag{2.1}$$

for $t_1 > 0$, where $1 - p$ denotes probability of being cured and $S_{10}(t_1) = \Pr(T_1 > t_1|\text{Not cured})$ denotes the survival function of T_1 for subjects susceptible to a disease progression event. The likelihood contribution for subjects i who have censored time-to-first event ($\delta_{12i} = 0$, $\delta_{13i} = 0$) is $S_1(t_{1i})$.

We denote the joint distribution function of time to cancer recurrence and time from recurrence to cancer death by $F_{12,23}(t_1, t_2) = \Pr(T_1 \leq t_1, T_2 \leq t_2, M = 2)$. We consider a copula function to model the conditional joint distribution function of T_1 and T_2 for subjects experiencing cancer recurrence as follows

$$\Pr(T_1 \leq t_1, T_2 \leq t_2 \,|\, M = 2) = C(F_{1|2}(t_1), F_2(t_2)), \tag{2.2}$$

where $C(.,.)$ is a bivariate copula function, $F_{1|2}(t_1) = \Pr(T_1 \leq t_1 \,|\, M = 2)$ is the conditional distribution function of T_1 for subjects experiencing recurrence and $F_2(t_2) = \Pr(T_2 \leq t_2 \,|\, M = 2)$ is the distribution function of T_2 for subjects who have experienced recurrence. A bivariate copula $C(u_1, u_2)$ is a distribution function on the unit square having uniform marginal distributions (Nelsen, 2006). Popular copula models in survival analysis include some Archimedean copulas (Genest and Rivest, 1993) such as the Clayton family (Clayton, 1978) and the Gumbel-Hougaard family (Gumbel, 1960).

The likelihood contribution for subjects i who have experienced cancer recurrence and died due to cancer ($\delta_{12i} = 1$, $\delta_{23i} = 1$) is

$$f_{12,23}(t_{1i}, t_{2i}) = \frac{\partial^2 F_{12,23}(t_{1i}, t_{2i})}{\partial t_{1i} \partial t_{2i}}.$$

The likelihood contribution for subjects i who have experienced cancer recurrence and

not died due to cancer before the end of followup ($\delta_{12i} = 1$, $\delta_{23i} = 0$) is

$$-\frac{\partial S_{12,23}(t_{1i}, t_{2i})}{\partial t_{1i}},$$

where $S_{12,23}(t_1, t_2) = \Pr(T_1 > t_1, T_2 > t_2, M = 2)$. The likelihood contribution for subjects i who died due to cancer without experiencing recurrence ($\delta_{13i} = 1$) is

$$\frac{\partial F_{13}(t_{1i})}{\partial t_{1i}},$$

where $F_{13}(t_1) = \Pr(T_1 \leq t_1, M = 3)$ is the cumulative incidence function of T_1 for subjects who died due to cancer without experiencing recurrence.

Then, the likelihood function for the observed data $\{(t_{1i}, t_{2i}, \delta_{12i}, \delta_{13i}, \delta_{23i}), i = 1, 2, \ldots, n\}$ becomes proportional to

$$\begin{aligned} L = \prod_{i=1}^{n} [f_{12,23}(t_{1i}, t_{2i})]^{\delta_{12i}\,\delta_{23i}} \left[-\frac{\partial S_{12,23}(t_{1i}, t_{2i})}{\partial t_{1i}}\right]^{\delta_{12i}(1-\delta_{23i})} \\ \times \left[\frac{\partial F_{13}(t_{1i})}{\partial t_{1i}}\right]^{\delta_{13i}} [S_1(t_{1i})]^{(1-\delta_{12i})(1-\delta_{13i})}. \end{aligned} \tag{2.3}$$

We can rewrite the likelihood function in (2.3) by expressing its components in terms of the non-cure probability p, probability of experiencing the state k as the first disease progression event $\Pr(M = k) = p\,\pi_{1k} = F_{1k}(\infty)$ for $k = 2, 3$, where $F_{1k}(t_1) = \Pr(T_1 \leq t_1, M = k)$, $\pi_{1k} = \Pr(M = k|\text{Not cured})$ subject to $\pi_{12} + \pi_{13} = 1$, and the copula model in (2.2). The survival function of T_1 in (2.1) becomes $1 - p\,\pi_{12}\,F_{1|2}(t_1) - p\,\pi_{13}\,F_{1|3}(t_1)$. The joint distribution function of time to recurrence and time from recurrence to cancer death becomes $F_{12,23}(t_1, t_2) = p\,\pi_{12}\,C(F_{1|2}(t_1), F_2(t_2))$. The joint survival function of time to recurrence and time from recurrence to cancer death, $S_{12,23}(t_1, t_2)$, can be written as $S_{12,23}(t_1, t_2) = p\,\pi_{12}\,\Pr(T_1 > t_1, T_2 > t_2 \,|\, M = 2)$

where $\Pr(T_1 > t_1, T_2 > t_2 \,|\, M = 2) = 1 - F_{1|2}(t_1) - F_2(t_2) + \Pr(T_1 \leq t_1, T_2 \leq t_2 \,|\, M = 2)$. Therefore, $-\frac{\partial S_{12,23}(t_1,t_2)}{\partial t_1}$ in (2.3) is equal to $p\,\pi_{12}\left[\frac{\partial F_{1|2}(t_1)}{\partial t_1} - \frac{\partial C(F_{1|2}(t_1),F_2(t_2))}{\partial t_1}\right]$. The likelihood function in (2.3) can be rewritten in terms of the copula model $C(F_{1|2}(t_1), F_2(t_2))$ and π_{1k}, $k = 2, 3$ as

$$
\begin{aligned}
L = \prod_{i=1}^{n} (p\,\pi_{12})^{\delta_{12i}} & \left(\frac{\partial^2 C(F_{1|2}(t_{1i}), F_2(t_{2i}))}{\partial t_{1i} \partial t_{2i}}\right)^{\delta_{12i}\,\delta_{23i}} \\
& \times \left(\frac{\partial F_{1|2}(t_{1i})}{\partial t_{1i}} - \frac{\partial C(F_{1|2}(t_{1i}), F_2(t_{2i}))}{\partial t_{1i}}\right)^{\delta_{12i}(1-\delta_{23i})} \\
& \times \left(p\,\pi_{13}\frac{\partial F_{1|3}(t_{1i})}{\partial t_{1i}}\right)^{\delta_{13i}} \left(1 - p\,\pi_{12}\,F_{1|2}(t_{1i}) - p\,\pi_{13}\,F_{1|3}(t_{1i})\right)^{(1-\delta_{12i})(1-\delta_{13i})}.
\end{aligned}
\tag{2.4}
$$

Maximum likelihood estimators of the parametric models are obtained by maximizing the likelihood function in (2.4). A nonlinear optimization algorithm could be used to obtain the maximum likelihood estimates. We used `nlm` function in `R` for this purpose.

2.2.1 Two-Stage Pseudo-Likelihood Estimation Method

An alternative estimation method is a two-stage pseudo-likelihood estimation method. Lawless and Yilmaz (2011) proposed a two-stage estimation method to estimate the bivariate copula model of sequentially observed time-to-events with a cured fraction for the first event. They estimated the distribution of the first gap time in the first stage and the distribution of the second gap time and the dependence parameter of the copula model in the second stage. The two-stage estimation method can also be applied to the current multi-state model.

In the first stage, the likelihood function for the observed data $\{(t_{1i}, \delta_{12i}, \delta_{13i}), i =$

$1, 2, \ldots, n\}$,

$$L_1 = \prod_{i=1}^{n} \left(p\,\pi_{12} \frac{\partial F_{1|2}(t_{1i})}{\partial t_{1i}} \right)^{\delta_{12i}} \left(p\,\pi_{13} \frac{\partial F_{1|3}(t_{1i})}{\partial t_{1i}} \right)^{\delta_{13i}} \times \left(1 - p\,\pi_{12}\,F_{1|2}(t_{1i}) - p\,\pi_{13}\,F_{1|3}(t_{1i})\right)^{(1-\delta_{12i})(1-\delta_{13i})}, \tag{2.5}$$

is maximized to obtain estimates of p, π_{12}, $\pi_{13} = 1 - \pi_{12}$, $F_{1|2}(.)$ and $F_{1|3}(.)$.

In the second stage, we estimate parameters for the second gap time distribution and the copula parameter(s) by maximizing the following pseudo-likelihood function where $F_{1|2}(.)$ is replaced with its estimate $\hat{F}_{1|2}(.)$ obtained in the first stage:

$$L_2 = \prod_{i=1}^{n} \left(\frac{\partial^2 C(\hat{F}_{1|2}(t_{1i}), F_2(t_{2i}))}{\partial \hat{F}_{1|2}(t_{1i}) \partial t_{2i}} \right)^{\delta_{12i}\,\delta_{23i}} \left(1 - \frac{\partial C(\hat{F}_{1|2}(t_{1i}), F_2(t_{2i}))}{\partial \hat{F}_{1|2}(t_{1i})} \right)^{\delta_{12i}(1-\delta_{23i})}. \tag{2.6}$$

When p, π_{12} and parameters in $F_{1|2}(.)$ and $F_{1|3}(.)$ are fixed, the likelihood function in (2.4) is proportional to L_2.

The two-stage estimation procedure is computationally more efficient since fewer parameters are simultaneously estimated at each stage. The standard errors of the estimators of parameters can be estimated through a nonparametric bootstrap.

2.3 Estimation in the Presence of Masked Causes of Deaths

In cancer prognosis data, it is common to have missing causes for some observed deaths. Individuals who die without experiencing cancer recurrence may die due to cancer or due to other causes. If the cause of death is unknown for an individual who did not experience any cancer recurrence, we add a masked cause of death in

addition to the latent cure status. Thus, individuals with masked causes of deaths have unknown cancer death indicator, δ_{13}, and unknown cure status. Because our interest is to only make an inference on time to cancer related events, when constructing our multi-state model, we consider the cancer progression events in Figure 2.1 with only the cancer death state as a death state. However, the data generation process is as in Figure 2.2 with an additional death due to other causes. There are three possibilities if an individual has a masked cause of death. The individual can be (i) uncured and died due to cancer, (ii) uncured and died due to other causes, or (iii) cured and died due to other causes.

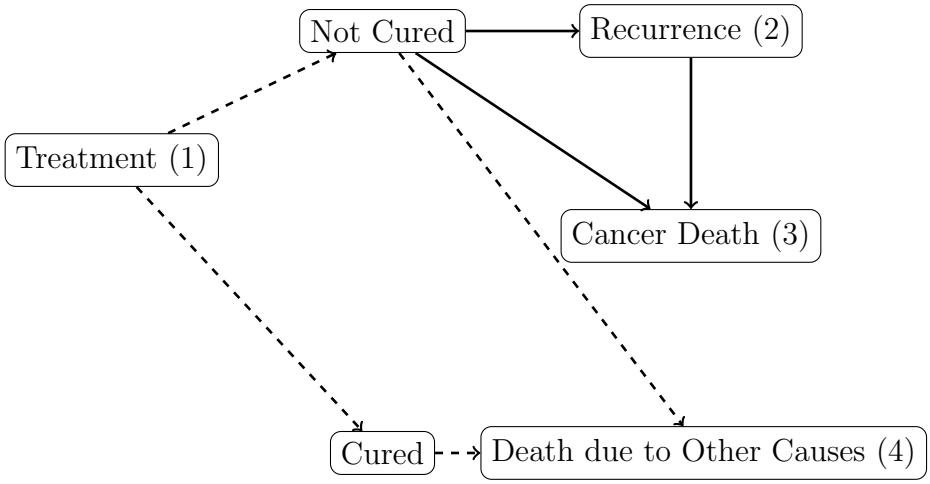


Figure 2.2: Multi-state model structure with death due to other causes state

In order to study with masked causes in a competing risks setting without any cure fraction, a two-stage data collection design was considered by Flehinger et al. (1998) and Craiu and Duchesne (2004). In their two-stage data collection, first a regular data collection is executed with possibly masked causes of deaths, and in the second stage, the true causes for some of the masked causes of failure are determined through further data collection. In our study, instead of collecting additional data, we use empirical evidence obtained from the distribution of T_1 to determine true causes of certain masked causes of deaths. Cured individuals are expected not to experience

any cancer related event and they are expected to eventually die due to other causes. When there is a cured proportion, if the follow-up time is long enough, we expect the empirical distribution of T_1 to level off beyond some value $\tau_{\max} < C_{\max}$ where $C_{\max}$ is the largest followup time for T_1. Thus, if a death occurs after the last observed cancer related event time $\tau_{\max}$, we have empirical evidence to assume that the individual is cured and death is due to other causes than cancer.

2.3.1 Maximum Likelihood Estimation via EM Algorithm

We denote the transition from the "Treatment" state to the "Cancer Death" state $(1 \rightarrow 3)$ as transition III and the transition from the "Treatment" state to the "Death due to Other Causes" state $(1 \rightarrow 4)$ as transition IV. The transition IV can occur for both cured and not cured individuals, but the transition III can only occur for not cured individuals. The masked cause of death would belong to the group $g_m = g_{III} \cup g_{IV} = \{III, IV\}$, where $g_{III} = \{III\}$ and $g_{IV} = \{IV\}$. We let $\gamma_{g_m i} = I[\text{cause of death masked to group } g_m \text{ for individual } i]$. Thus, $\gamma_{g_m i} = 1$ if ith individual died but δ_{13i} is unknown and $\gamma_{g_m i} = 0$ if δ_{13i} is known. We define a death due to any cause indicator $\delta_{di} = I[\text{death occurs for individual } i]$. Then, the complete data is $\{(t_{1i}, t_{2i}, \delta_{12i}, \delta_{13i}, \delta_{23i}, \delta_{di}, \gamma_{g_m i}), i = 1, 2, \ldots, n\}$.

The likelihood function based on the model in Figure 2.1 for the complete data

becomes

$$
\begin{aligned}
L_c(\boldsymbol{\theta}) = & \prod_{i=1}^{n} (p\,\pi_{12})^{\delta_{12i}} \left(\frac{\partial^2 C(F_{1|2}(t_{1i}), F_2(t_{2i}))}{\partial t_{1i} \partial t_{2i}} \right)^{\delta_{12i}\,\delta_{23i}} \\
& \times \left(\frac{\partial F_{1|2}(t_{1i})}{\partial t_{1i}} - \frac{\partial C(F_{1|2}(t_{1i}), F_2(t_{2i}))}{\partial t_{1i}} \right)^{\delta_{12i}(1-\delta_{23i})} \\
& \times (p\pi_{13})^{\delta_{13i}} \left(\frac{\partial F_{1|3}(t_{1i})}{\partial t_{1i}} \right)^{\delta_{13i}} \\
& \times [1 - p\,\pi_{12}\,F_{1|2}(t_{1i}) - p\pi_{13}\,F_{1|3}(t_{1i})]^{(1-\delta_{12i})(1-\delta_{13i})} \\
& \times (P_{g_m|g_{III}})^{\delta_{13i}\gamma_{g_m i}} (1 - P_{g_m|g_{III}})^{\delta_{13i}(1-\gamma_{g_m i})} \\
& \times (P_{g_m|g_{IV}})^{(1-\delta_{12i})(1-\delta_{13i})\delta_{di}\gamma_{g_m i}} (1 - P_{g_m|g_{IV}})^{(1-\delta_{12i})(1-\delta_{13i})\delta_{di}(1-\gamma_{g_m i})},
\end{aligned}
\tag{2.7}
$$

where $P_{g_m|g_j} = \Pr(\text{cause of death masked to group } g_m | \text{cause is actually in group } j)$ for $j = III, IV$ and $\boldsymbol{\theta}$ is the vector of unknown parameters. We assume that δ_{13i} is missing at random (i.e., $P_{g_m|g_j}$ does not depend on δ_{13i} for $j = III, IV$) (Craiu and Duchesne, 2004).

We define the cause-specific hazard function for cause $M = k$, $k = 2, 3$ conditional on not being cured as

$$
\lambda_{1k}(t_1) = \lim_{\Delta t \to 0} \frac{\Pr(T_1 \in [t_1, t_1 + \Delta t), M = k \mid T_1 \geq t_1, \text{Not cured})}{\Delta t}, \tag{2.8}
$$

and the cause-specific hazard function for cause $M = 4$ as

$$
\lambda_{14}(t_1) = \lim_{\Delta t \to 0} \frac{\Pr(T_1 \in [t_1, t_1 + \Delta t), M = 4 \mid T_1 \geq t_1)}{\Delta t}. \tag{2.9}
$$

Using the Baye's rule, we obtain the diagnostic probability $\pi_{g_{III}|g_m}(t_1)$ that actually died due to cancer at time t_1 given that it is masked to group g_m as

$$
\pi_{g_{III}|g_m}(t_1) = \frac{p\,\lambda_{13}(t_1)\,P_{g_m|g_{III}}}{p\,\lambda_{13}(t_1)\,P_{g_m|g_{III}} + \lambda_{14}(t_1)\,P_{g_m|g_{IV}}}. \tag{2.10}
$$

The conditional expectation of the logarithm of (2.7) given the observed data is

$$
\begin{aligned}
E[l_c(\boldsymbol{\theta})|\text{Obs}] = & \sum_{i=1}^{n} \delta_{12i}\ \log(p\,\pi_{12}) + \sum_{i=1}^{n} \delta_{12i}\,\delta_{23i} \log\left(\frac{\partial^2 C(F_{1|2}(t_{1i}), F_2(t_{2i}))}{\partial t_{1i}\partial t_{2i}}\right) \\
& + \sum_{i=1}^{n} \delta_{12i}(1-\delta_{23i}) \log\left(\frac{\partial F_{1|2}(t_{1i})}{\partial t_{1i}} - \frac{\partial C(F_{1|2}(t_{1i}), F_2(t_{2i}))}{\partial t_{1i}}\right) \\
& + \log(p\,\pi_{13}) \sum_{i=1}^{n} E[\delta_{13i}|\text{Obs}] + \sum_{i=1}^{n} E[\delta_{13i}|\text{Obs}] \log\left(\frac{\partial F_{1|3}(t_{1i})}{\partial t_{1i}}\right) \\
& + \sum_{i=1}^{n} (1-\delta_{12i})(1-E[\delta_{13i}|\text{Obs}]) \log\left[1 - p\,\pi_{12}\,F_{1|2}(t_{1i}) - p\,\pi_{13}\,F_{1|3}(t_{1i})\right] \\
& + \log(P_{g_m|g_{III}}) \sum_{i=1}^{n} \gamma_{g_m i} E[\delta_{13i}|\text{Obs}] + \log(1-P_{g_m|g_{III}}) \sum_{i=1}^{n} (1-\gamma_{g_m i}) E[\delta_{13i}|\text{Obs}] \\
& + \log(P_{g_m|g_{IV}}) \sum_{i=1}^{n} \delta_{di}\gamma_{g_m i}(1-\delta_{12i})(1-E[\delta_{13i}|\text{Obs}]) \\
& + \log(1-P_{g_m|g_{IV}}) \sum_{i=1}^{n} \delta_{di}(1-\gamma_{g_m i})(1-\delta_{12i})(1-E[\delta_{13i}|\text{Obs}]),
\end{aligned}
\tag{2.11}
$$

where $l_c(\boldsymbol{\theta}) = \log L_c(\boldsymbol{\theta})$ and

$$
E[\delta_{13i}|\text{Obs}] = \begin{cases}
1 & \text{if the cause of death for individual } i \text{ is known to be cancer,} \\
0 & \text{if the cause of death for individual } i \text{ is known to be other causes or} \\
 & \text{no death is observed for individual } i, \\
\pi_{g_{III}|g_m}(t_{1i}) & \text{if the death is observed but the cause of death} \\
 & \text{for individual } i \text{ is masked in } g_m.
\end{cases}
\tag{2.12}
$$

We use the following expectation (E) and maximization (M) steps to obtain the maximum likelihood estimates of the parametric models:

E-step : Compute

$$
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}) = E[l_c(\boldsymbol{\theta})|\text{Obs}, \boldsymbol{\theta}^{(l)}], \tag{2.13}
$$

M-step : Find

$$\boldsymbol{\theta}^{(l+1)} = Argmax\, Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}),$$

where $\lambda_{14}(t_1)$ in (2.10) is estimated nonparametrically as follows using the data without masked causes.

The cause specific cumulative hazard function for cause $M = 4$, $\Lambda_{14}(t_1)$, is first estimated by the Nelson-Aalen type estimator

$$\hat{\Lambda}_{14}(t) = \sum_{i=1}^{n} \frac{(1-\gamma_{g_{mi}})I(t_{1i} < t, M_i = 4)}{\sum_{l=1}^{n} I(t_{1l} \geq t_{1i})}. \tag{2.14}$$

Then, we obtain $\hat{\Lambda}_{14}(t^*_{IV(j)})$ for $t^*_{IV(1)} < t^*_{IV(2)} < \cdots < t^*_{IV(n_{IV})}$ which are distinct observed t_{1i}'s with $M_i = 4$ and n_{IV} is the total number of t_{1i}'s with $M_i = 4$. For $t_1 \in (t^*_{IV(j-1)}, t^*_{IV(j)}]$, we approximate $\lambda_{14}(t_1)$ by

$$\hat{\lambda}_{14}(t_{mj}) = \frac{\hat{\Lambda}_{14}(t^*_{IV(j)}) - \hat{\Lambda}_{14}(t^*_{IV(j-1)})}{\Delta_j}, \tag{2.15}$$

where $t_{mj} = 1/2(t^*_{IV(j)} + t^*_{IV(j-1)})$ and $\Delta_j = t^*_{IV(j)} - t^*_{IV(j-1)}$.

We use a general purpose optimization software, specifically the function `nlm` in `R`, to compute the M-step.

2.3.2 Standard Error Estimation

We apply the supplemented EM algorithm (SEM) and nonparametric bootstrap to obtain the standard error estimates of the obtained parameter estimators with the EM algorithm in Section 2.3.1.

SEM algorithm

The SEM algorithm can be used to obtain standard error estimates of the EM estimators. It takes the extra variability due to EM procedure into account (Meng and Rubin, 1991; Xu et al., 2014). EM algorithm uses the relation of a mapping $\boldsymbol{\theta} \rightarrow M(\boldsymbol{\theta})$ for $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$ such that

$$\boldsymbol{\theta}^{(l+1)} = M(\boldsymbol{\theta}^{(l)}), \tag{2.16}$$

where $\boldsymbol{\theta}^{(l)}$ is the estimate of unknown parameter vector $\boldsymbol{\theta}$ from the lth iteration. Then, the variance covariance matrix of the EM estimator $\hat{\boldsymbol{\theta}}$ can be estimated by (Meng and Rubin, 1991)

$$\widehat{Var}(\hat{\boldsymbol{\theta}}) = I_{oc}^{-1} + I_{oc}^{-1}\, DM(I - DM)^{-1}, \tag{2.17}$$

where I_{oc} is the complete observed information matrix, I is the identity matrix and DM is the Jacobian matrix for $M(\boldsymbol{\theta})$ evaluated at $\hat{\boldsymbol{\theta}}$ with entries

$$r_{ij} = \left(\frac{\partial M_j(\boldsymbol{\theta})}{\partial \theta_i} \right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \quad i = 1, 2, \ldots, p, \quad j = 1, 2, \ldots, p. \tag{2.18}$$

We let $\tilde{\boldsymbol{\theta}}^{(l)}(i)$ to be

$$\tilde{\boldsymbol{\theta}}^{(l)}(i) = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_{i-1}, \theta_i^{(l)}, \hat{\theta}_{i+1}, \ldots, \hat{\theta}_p), \tag{2.19}$$

where $\hat{\theta}_j$ is the maximum likelihood estimator of θ_j for $j \neq i$ and $i, j = 1, 2, \ldots, p$. Equation (2.19) means that only the ith component is in the lth iteration of the algorithm while other components are maximum likelihood estimators of θ_j. Meng and Rubin (1991) showed that r_{ij} can be approximated by the following procedure

1. Obtain $\hat{\theta}_j$ for $j = 1, 2, \ldots, p$ where $\hat{\theta}_j$ is the maximum likelihood estimator of θ_j

using the EM algorithm.

2. For $i = 1, 2, \ldots, p$, obtain $\theta_j^{(l+1)}(i)$ by treating $\tilde{\boldsymbol{\theta}}^{(l)}(i)$ in (2.19) as the lth iteration and running one more maximization step.

3. Obtain the ratio
$$r_{ij}^{(l)} = \frac{\theta_j^{(l+1)}(i) - \hat{\theta}_j}{\theta_i^{(l)} - \hat{\theta}_i}. \tag{2.20}$$

4. Obtain r_{ij} until $r_{ij}^{(l+1)} - r_{ij}^{(l)} < \epsilon$ for some small positive ϵ value.

Bootstrap Procedure

Nonparametric bootstrap procedure can also be used to obtain the standard error estimates of the EM estimators as follows:

(i) Obtain a random sample $\{(\tilde{t}_{1r}, \tilde{t}_{2r}, \tilde{\delta}_{12r}, \tilde{\delta}_{13r}, \tilde{\delta}_{23r}, \tilde{\delta}_{dr}, \tilde{\gamma}_{g_m r}), r = 1, 2, \ldots, n\}$ with replacement from the observed data $\{(t_{1i}, t_{2i}, \delta_{12i}, \delta_{13i}, \delta_{23i}, \delta_{di}, \gamma_{g_m i}), i = 1, 2, \ldots, n\}$.

(ii) Using the E and M steps in Section 2.3.2, obtain the estimate $\hat{\boldsymbol{\theta}}$ of unknown parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ from $\{(\tilde{t}_{1r}, \tilde{t}_{2r}, \tilde{\delta}_{12r}, \tilde{\delta}_{13r}, \tilde{\delta}_{23r}, \tilde{\gamma}_{g_m r}, \tilde{\delta}_{dr}), r = 1, 2, \ldots, n\}$.

(iii) Repeat the steps (i) and (ii) B times and obtain the estimates $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \ldots, \hat{\boldsymbol{\theta}}_B$.

(iv) Variance-covariance matrix is estimated by
$$\frac{1}{B} \sum_{b=1}^{B} (\hat{\boldsymbol{\theta}}_b - \overline{\hat{\boldsymbol{\theta}}})(\hat{\boldsymbol{\theta}}_b - \overline{\hat{\boldsymbol{\theta}}})^T, \tag{2.21}$$

where $\overline{\hat{\boldsymbol{\theta}}}$ is the mean vector of $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \ldots, \hat{\boldsymbol{\theta}}_B$.

2.4 Simulation Study

A Monte Carlo simulation study was conducted to study the finite sample properties of the proposed estimation methods. The algorithm to generate data from the Figure 2.1 model in the absence of masked causes of deaths is described in Section 2.4.1 and the algorithm to generate data from the Figure 2.2 model in the presence of masked causes of deaths is described in Section 2.4.2. Properties of the proposed estimation methods were assessed when there are no masked causes of deaths in Section 2.4.3 and when there are masked causes of deaths in Section 2.4.4.

2.4.1 Data Generation Algorithm in the Absence of Masked Causes of Deaths

In this section, we describe the data generation algorithm from the Figure 2.1 model when there are no masked causes of deaths. We generated $1,000$ random samples of $\{(t_{1i}, t_{2i}, \delta_{12i}, \delta_{13i}, \delta_{23i}), i = 1, 2, \ldots, n\}$ from the multi-state model in Figure 2.1 for each of size $n = 200$ and $n = 400$. We generated cure status for each individual i from Bernoulli distribution with cure probability $1-p$. We set $p = 0.70$. For uncured individuals, we generated T_{1i} from $F_{10}(t_{1i}) = \Pr(T_{1i} \leq t_{1i}|\text{Not Cured} = 1 - S_{10}(t_{1i}) = 1 - \exp\left[-\int_0^{t_{1i}} (\lambda_{12}(u) + \lambda_{13}(u))\, du\right]$. We assumed log-logistic distribution to model the cause specific hazard functions (2.8) for disease progression events $M = k$, $k = 2, 3$ conditional on not being cured:

$$\lambda_{1k}(t_1) = \frac{\left(\frac{\beta_{1k}}{\alpha_{1k}}\right)\left(\frac{t_1}{\alpha_{1k}}\right)^{\beta_{1k}-1}}{1 + \left(\frac{t_1}{\alpha_{1k}}\right)^{\beta_{1k}}}, \qquad t_1 > 0, \qquad k = 2, 3. \tag{2.22}$$

The log-logistic distribution is widely used for modeling time-to-events when the rate of the event increases initially and decreases later. The log-logistic distribution is,

for example, a good fit for the distribution of time-to-first disease progression events, cancer recurrence or cancer death, in colon cancer data analyzed in Section 2.5. We set $\alpha_{12} = 2.0$, $\beta_{12} = 4.0$ and $\alpha_{13} = 3.5$, $\beta_{13} = 3.0$ to have more observed cancer recurrences than cancer deaths without experiencing recurrence. For uncured individuals, we set $M_i = 2$ with probability $\frac{\lambda_{12}(T_{1i})}{\lambda_{12}(T_{1i})+\lambda_{13}(T_{1i})}$, and set $M_i = 3$ otherwise. For an uncured individual i, if $M_i = 2$, we generated T_{2i} from $C(F_{1|2}(T_{1i}), F_2(T_{2i}))$ in (2.2), the conditional joint distribution of T_1 and T_2 for patients who experience cancer recurrence. We considered the Clayton copula function (Clayton, 1978)

$$C_\phi(u_1, u_2) = (u_1^{-\phi} + u_2^{-\phi} - 1)^{-1/\phi}, \qquad \phi > 0, \tag{2.23}$$

where $u_1 = F_{1|2}(t_1), u_2 = F_2(t_2)$ and ϕ is the dependence parameter. The Clayton copula is one of the widely used copula families to model bivariate time-to-event data. The dependence measure Kendall's τ can be written in terms of the Clayton copula parameter ϕ as $\tau = \phi/(\phi + 2)$. U_1 and U_2 are positively associated when $\phi > 0$ and the dependence increases as the value of the parameter ϕ increases. We considered two levels of dependence, moderate and strong dependence between the sequential gap times with Kendall's $\tau = 0.3$ ($\phi = 0.857$) and $\tau = 0.7$ ($\phi = 4.667$), respectively. The marginal distribution of T_2 for subjects who have experienced recurrence was assumed as the Weibull distribution with

$$F_2(t_2) = 1 - \exp\left[-\left(\frac{t_2}{\alpha_{23}}\right)^{\beta_{23}}\right], \qquad t_2 > 0. \tag{2.24}$$

The Weibull distribution is one of the most widely used time-to-event distributions since it is fairly flexible. The Weibull hazard function can be monotone increasing, decreasing or constant based on its shape parameter (β_{23}) value. We set $\alpha_{23} = 2.5$ and $\beta_{23} = 1.5$. Since $\beta_{23} > 1$, the rate of death after experiencing cancer recurrence

increases over time. We generated censoring times $\{C_i, i = 1, 2, \ldots, n\}$ from Uniform $(0, 10)$. For cured individuals, we generated time-to-death due to other causes from $F_{14|\text{Cured}}(t_1) = \Pr(T_{1i} \leq t_{1i}, M_i = 4|\text{Cured})$ and set the censoring time C_i as the time to death due to other causes if the censoring time is greater than the time to death due to other causes. Time to death due to other causes is considered as a censoring time when the estimation is performed using the likelihood function in (2.4) and the two-stage pseudo-likelihood estimation method described in Section 2.2.1. We considered the Weibull model for $F_{14|\text{Cured}}(t_1)$ in the form of

$$F_{14|\text{Cured}}(t_1) = 1 - \exp\left[-\left(\frac{t_1}{\alpha_{14|\text{Cured}}}\right)^{\beta_{14|\text{Cured}}}\right], \qquad t_1 > 0. \tag{2.25}$$

We set $\alpha_{14|\text{Cured}} = 7.0$ and $\beta_{14|\text{Cured}} = 4.0$. We obtained $t_{1i} = \min(T_{1i}, C_i)$, $\delta_{12i} = I[T_{1i} = t_{1i}, M_i = 2]$, $\delta_{13i} = I[T_{1i} = t_{1i}, M_i = 3]$. If $M_i = 2$, we obtained $t_{2i} = \min(T_{2i}, C_i - t_{1i})$ and $\delta_{23i} = I[T_{2i} = t_{2i}, M_i = 2]$.

In Section 2.4.2, we discuss the simulation results when the data is generated as described in this section from the Figure 2.1 model and when the Figure 2.1 model is fitted using the maximum likelihood and the two-stage pseudo-likelihood estimation methods in the absence of masked causes of deaths.

2.4.2 Data Generation Algorithm in the Presence of Masked Causes of Deaths

In this section, we describe the data generation algorithm from the Figure 2.2 model when there are masked causes of deaths.

Our multi-state model in Section 2.3.1 is based on the cancer progression events in Figure 2.1. However, the data generation process follows the multi-state model in Figure 2.2 with the additional "Death Due to Other Causes" state. We first explain

the data generation process. We then explain how the complete data is obtained by using the empirical evidence to assess if an individual is cured.

We generated 1000 random samples of $\{(t_{1i}, t_{2i}, \delta_{12i}, \delta_{13i}, \delta_{23i}, \delta_{14i}, \delta_{di}, \gamma_{g_m i}), i = 1, 2, \ldots, n\}$ from the multi-state model in Figure 2.2 with masked causes of deaths for each of size $n = 400$. In Figure 2.2 individuals who died due to other causes can be cured or not cured. Thus, the cumulative incidence function of T_1 for subjects who died due to other causes is $F_{14}(t_1) = \Pr(T_1 \leq t_1, M = 4) = (1-p)F_{14|\text{Cured}}(t_1) + pF_{14|\text{Not Cured}}(t_1)$, where $F_{14|\text{Cured}}(t_1) = \Pr(T_1 \leq t_1, M = 4|\text{Cured})$ and $F_{14|\text{Not Cured}}(t_1) = \Pr(T_1 \leq t_1, M = 4|\text{Not Cured})$. When an individual is not cured, the first disease progression event can be any of $M = 2, 3, 4$. Therefore, time to first event for not cured is generated from $\Pr(T_{1i} \leq t_{1i}|\text{Not Cured}) = 1 - S_{10}(t_{1i}) = 1 - \exp\left[-\int_0^{t_{1i}} \left(\lambda_{12}(u) + \lambda_{13}(u) + \lambda_{14|\text{Not Cured}}(u)\right) du\right]$, where $\lambda_{14|\text{Not Cured}}(t_1)$ is the cause specific intensity function for cause $M = 4$ conditional on not being cured.

We generated cure status for each individual i from Bernoulli distribution with cure probability $1 - p$. We set $p = 0.70$. For uncured individuals, we generated T_{1i} from $F_{10}(t_{1i}) = \Pr(T_{1i} \leq t_{1i}|\text{Not Cured}) = 1 - S_{10}(t_{1i}) = 1 - \exp\left[-\int_0^{t_{1i}} \left(\lambda_{12}(u) + \lambda_{13}(u) + \lambda_{14|\text{Not Cured}}(u)\right) du\right]$. We assumed log-logistic distribution (2.22) to model the cause specific hazard functions for disease progression events $M = k$, $k = 2, 3$ conditional on not being cured. We set $\alpha_{12} = 2.0$, $\beta_{12} = 4.0$ and $\alpha_{13} = 3.5$, $\beta_{13} = 3.0$ to have more observed cancer recurrences than cancer deaths without experiencing recurrence. We considered uniform distribution to model the cause specific hazard for cause $M = 4$ conditional on not being cured:

$$\lambda_{14|\text{Not Cured}}(t_1) = \frac{1}{\beta_{14|\text{Not Cured}} - t_1}, \qquad \alpha_{14|\text{Not Cured}} < t_1 < \beta_{14|\text{Not Cured}}. \tag{2.26}$$

We set $\alpha_{14|\text{Not Cured}} = 1.3$ and $\beta_{14|\text{Not Cured}} = 6.0$. For uncured individuals, we set

$M_i = 2$ with probability $\frac{\lambda_{12}(T_{1i})}{\lambda_{12}(T_{1i})+\lambda_{13}(T_{1i})+\lambda_{14|\text{Not Cured}}(T_{1i})}$, set $M_i = 3$ with probability $\frac{\lambda_{13}(T_{1i})}{\lambda_{12}(T_{1i})+\lambda_{13}(T_{1i})+\lambda_{14|\text{Not Cured}}(T_{1i})}$, and set $M_i = 4$ otherwise. For individuals with $M_i = 2$, we generated T_{2i} from the copula function $C(F_{1|2}(T_{1i}), F_2(T_{2i}))$ in (2.2). We considered the Clayton copula function in (2.23) with $\phi = 0.857$ which gives the Kendall's $\tau = \phi/(\phi + 2) = 0.3$. The marginal distribution of T_2 for subjects who have experienced recurrence was considered as Weibull distribution with $F_2(t_2) = 1 - \exp\left[-\left(\frac{t_2}{\alpha_{23}}\right)^{\beta_{23}}\right]$. We set $\alpha_{23} = 2.5$ and $\beta_{23} = 1.5$. For cured individuals, we set $M_i = 4$ and generated time-to-death due to other causes, T_{1i}, from the Weibull distribution (2.25) with $\alpha_{14|\text{Cured}} = 7.0$ and $\beta_{14|\text{Cured}} = 4.0$. We generated censoring times $\{C_i, i = 1, 2, \ldots, n\}$ from Uniform $(0, 15)$. We obtained $t_{1i} = \min(T_{1i}, C_i)$, $\delta_{12i} = I[T_{1i} = t_{1i}, M_i = 2]$, $\delta_{13i} = I[T_{1i} = t_{1i}, M_i = 3]$, $\delta_{14i} = I[T_{1i} = t_{1i}, M_i = 4]$. If $M_i = 2$, we obtained $t_{2i} = \min(T_{2i}, C_i - t_{1i})$, $\delta_{23i} = I[T_{2i} = t_{2i}, M_i = 2]$.

We then obtained the complete data $\{(t_{1i}, t_{2i}, \delta_{12i}, \delta_{13i}, \delta_{23i}, \delta_{di}, \gamma_{g_m i}), i = 1, 2, \ldots, n\}$ with masked causes of deaths using the generated data $\{(t_{1i}, t_{2i}, \delta_{12i}, \delta_{13i}, \delta_{23i}, \delta_{14i}), i = 1, 2, \ldots, n\}$. We obtained $\delta_{di} = \delta_{13i} + \delta_{14i}$ for each individual. For subjects with $M_i = 3$, we generated V_{1i} from Bernoulli distribution with probability $P_{g_m|g_{III}}$. For subjects with $M_i = 4$, we generated V_{2i} from Bernoulli distribution with probability $P_{g_m|g_{IV}}$. If $V_{1i} = 1$ or $V_{2i} = 1$, then δ_{13i} and δ_{14i} are masked and we set $\gamma_{g_m i} = 1$. Otherwise, $\gamma_{g_m i} = 0$. We considered two cases: lower percentages of masked causes of deaths with probabilities $P_{g_m|g_{III}} = 0.2, P_{g_m|g_{IV}} = 0.1$ and higher percentages of masked causes of deaths with probabilities $P_{g_m|g_{III}} = 0.4, P_{g_m|g_{IV}} = 0.3$. We obtained the last observed disease related event time $\tau_{\max} = \max\{(1-\gamma_{g_m i})[\delta_{12i}(1-\delta_{13i}) + (1-\delta_{12i})\delta_{13i}]t_{1i}, i = 1, 2, \ldots, n\}$. For subjects with $\gamma_{g_m i} = 1$ and $t_{1i} > \tau_{\max}$, we assumed $\delta_{13i} = 0$ and $\delta_{14i} = 1$. Thus, we obtained complete data $\{(t_{1i}, t_{2i}, \delta_{12i}, \delta_{13i}, \delta_{23i}, \delta_{di}, \gamma_{g_m i}), i = 1, 2, \ldots, n\}$.

In Section 2.4.4, we discuss the simulation results when the data is generated as

described in this section from the Figure 2.2 model but when the Figure 2.1 model is fitted.

2.4.3 Simulation Results in the Absence of Masked Causes of Deaths

Table 2.1 gives the empirical means and standard deviations and mean standard errors of maximum likelihood estimates and two-stage pseudo-likelihood estimates over $1,000$ replications when the data is generated as described in Section 2.4.1 from the Figure 2.1 model and when the Figure 2.1 model is fitted in the absence of masked causes of deaths. It shows that the mean point estimates using both the maximum likelihood estimation and the two-stage pseudo-likelihood estimation are close to the true values compared to their standard deviations. Maximum likelihood estimation gives slightly more efficient estimators than the two-stage pseudo-likelihood estimation under both moderate and strong levels of dependence between sequential gap times. When there is heavy censoring and only a few number of events is observed for a transition (for transition $1 \to 3$ when $n = 200$), standard errors of the two-stage pseudo-likelihood estimators are overestimated through the nonparametric bootstrap. As the sample size increases to $n = 400$, both methods yield more efficient estimators. Maximum likelihood estimation method is computationally more intensive than the two-stage pseudo-likelihood estimation method.

	True	Maximum Likelihood Estimation			Two-Stage pseudo-likelihood Estimation		
	Value	$Mean(Est)$	$SD(Est)$	$\overline{\widehat{SE}}(Est)$	$Mean(Est)$	$SD(Est)$	$\overline{\widehat{SE}}_{boot}(Est)$
				Kendall's $\tau = 0.3, n = 200$			
α_{12}	2.000	2.021	0.093	0.090	2.020	0.094	0.092
β_{12}	4.000	3.973	0.366	0.355	3.988	0.370	0.377
α_{13}	3.500	3.581	0.438	0.426	3.578	0.437	0.481
β_{13}	3.000	3.076	0.504	0.494	3.079	0.506	0.526
α_{23}	2.500	2.507	0.217	0.209	2.508	0.217	0.214
β_{23}	1.500	1.546	0.150	0.145	1.546	0.149	0.151
ϕ	0.857	0.837	0.229	0.228	0.831	0.227	0.243
p	0.700	0.709	0.038	0.038	0.709	0.038	0.038
				Kendall's $\tau = 0.3, n = 400$			
α_{12}	2.000	2.016	0.064	0.063	2.014	0.064	0.065
β_{12}	4.000	3.961	0.251	0.249	3.975	0.256	0.262
α_{13}	3.500	3.563	0.300	0.292	3.559	0.299	0.295
β_{13}	3.000	3.012	0.348	0.342	3.015	0.349	0.348
α_{23}	2.500	2.501	0.148	0.148	2.502	0.148	0.151
β_{23}	1.500	1.531	0.100	0.101	1.533	0.100	0.103
ϕ	0.857	0.837	0.162	0.160	0.833	0.160	0.166
p	0.700	0.709	0.027	0.027	0.708	0.027	0.027
				Kendall's $\tau = 0.7, n = 200$			
α_{12}	2.000	2.022	0.085	0.084	2.013	0.092	0.091
β_{12}	4.000	3.952	0.338	0.334	4.001	0.362	0.378
α_{13}	3.500	3.618	0.506	0.440	3.605	0.509	0.512
β_{13}	3.000	3.023	0.490	0.488	3.032	0.496	0.518
α_{23}	2.500	2.482	0.200	0.195	2.471	0.206	0.208
β_{23}	1.500	1.570	0.139	0.140	1.583	0.142	0.147
ϕ	4.667	4.653	0.778	0.770	4.600	0.775	0.817
p	0.700	0.707	0.039	0.038	0.706	0.039	0.038
				Kendall's $\tau = 0.7, n = 400$			
α_{12}	2.000	2.029	0.059	0.059	2.017	0.064	0.064
β_{12}	4.000	3.942	0.244	0.235	3.989	0.262	0.262
α_{13}	3.500	3.591	0.302	0.296	3.574	0.301	0.313
β_{13}	3.000	3.000	0.353	0.342	3.011	0.355	0.351
α_{23}	2.500	2.500	0.143	0.138	2.483	0.148	0.146
β_{23}	1.500	1.565	0.097	0.098	1.577	0.100	0.102
ϕ	4.667	4.519	0.530	0.529	4.480	0.528	0.543
p	0.700	0.707	0.028	0.027	0.706	0.028	0.027

Table 2.1: Monte-Carlo simulation study results in the absence of masked causes of deaths. Simulation study was conducted using $1,000$ replications with sample size $n = 200$ and $n = 400$ under moderate dependence ($\tau = 0.3$) and strong dependence ($\tau = 0.7$) between the sequential gap times. Censoring rates are approximately 55% for $1 \to 2$ transition, 87% for $1 \to 3$ transition and 68% for $2 \to 3$ transition. Est refers to estimate of the corresponding parameter, $Mean(Est)$ refers to the mean of the estimates, $SD(Est)$ refers to the standard deviation of the estimates, $\overline{\widehat{SE}}(Est)$ refers to the average standard error estimates and $\overline{\widehat{SE}}_{boot}(Est)$ refers to average standard error estimates obtained by nonparametric bootstrap with $1,000$ bootstrap samples over $1,000$ replications.

2.4.4 Simulation Results in the Presence of Masked Causes of Deaths

In this section the data was generated as described in Section 2.4.2 from the Figure 2.2 model but the Figure 2.1 model was fitted. We considered three different scenarios: (a) Masked causes of deaths are present and the estimation method in Section 2.3 is used, (b) causes of deaths are all assumed to be fully observed and the estimation method in Section 2.2 is used, and (c) individuals with masked causes of deaths are removed from the data and the estimation method in Section 2.2 is used. In scenario (a) we assessed the performance of the proposed maximum likelihood estimation method through the EM algorithm in Section 2.3 in the presence of the masked causes of deaths. In scenario (b) we assessed the performance of the proposed maximum likelihood estimation method fitting Figure 2.1 in Section 2.2 in the absence of masked causes of deaths when the data was generated from Figure 2.2. In scenario (c) we showed the inaccuracy in parameter estimates when the data with masked causes of deaths are omitted in analysis, that is a frequently applied analysis approach by practitioners. Results of the three scenarios are shown in Table 2. It gives the empirical mean and standard deviation and mean standard error of maximum likelihood estimates over $1,000$ replications. The supplemented EM algorithm is used to obtain the standard error estimates in scenario (a).

The results in Table 2.2 show that the point estimates obtained by the proposed method taking the masked causes of deaths into account under scenario (a) are closer to the true values of the parameters except the parameter estimates for the transition $1 \to 3$. We obtain slightly biased estimates of the parameters and standard errors of estimators for the transition $1 \to 3$ since we generated the data from the multi-state model in Figure 2.2 but fitted the Figure 2.1 model. The amount of bias in parameter

estimates for transition $1 \rightarrow 3$ increased when a higher probability of masked causes of deaths was considered. The other parameter estimates are more accurate in scenario (a) than scenarios (b) and (c) with valid standard error estimates obtained by the supplemented EM algorithm.

In scenario (b) when no masked causes of deaths exist, because the Figure 2.2 model data was fitted to the Figure 2.1 model, the maximum likelihood estimates of the cure probability and the parameters in transition $1 \rightarrow 3$ are slightly biased. The estimation methods fitting the Figure 2.1 model in Sections 2.2 and 2.3 do not consider the transition from treatment to death due to other causes than cancer since our interest is only to model time-to-disease related events. For deaths due to other causes than cancer, the cure status is usually unknown. The estimation methods fitting the Figure 2.1 model in Sections 2.2 and 2.3 assume that the distribution of time-to-death due to other causes follows the distribution of censoring times, and time-to-deaths due to other causes are treated as censoring times. Although this assumption holds in the data generation algorithm in Section 2.4.1, it does not hold in the data generation algorithm described in Section 2.4.2. Therefore, we observe little bias in the estimates of cure probability and parameters in transition $1 \rightarrow 3$ in Table 2.2.

In scenario (c) when the individuals' data with masked causes of deaths are omitted in the analysis, we observe significant bias in the maximum likelihood estimates of the parameters in the distribution of time-to-first disease related event including the cure probability. The amount of bias increases when the percentage of the masked causes of deaths increases. This indicates that it is not a good practice to exclude data with masked causes of death.

	True	Scenario (a)			Scenario (b)			Scenario (c)		
	Value	$Mean(Est)$	$SD(Est)$	$\widehat{\overline{SE}}(Est)$	$Mean(Est)$	$SD(Est)$	$\widehat{\overline{SE}}_{boot}(Est)$	$Mean(Est)$	$SD(Est)$	$\widehat{\overline{SE}}_{boot}(Est)$
					Lower probabilities of masked causes of deaths					
α_{12}	2.000	1.934	0.062	0.062	1.922	0.060	0.061	1.902	0.059	0.060
β_{12}	4.000	4.142	0.286	0.278	4.192	0.283	0.281	4.238	0.288	0.286
α_{13}	3.500	3.236	0.307	0.367	3.321	0.306	0.291	3.599	0.393	0.377
β_{13}	3.000	3.353	0.438	0.421	3.153	0.418	0.390	3.119	0.441	0.430
α_{23}	2.500	2.510	0.155	0.161	2.514	0.155	0.160	2.520	0.156	0.160
β_{23}	1.500	1.525	0.108	0.106	1.533	0.108	0.107	1.539	0.108	0.107
ϕ	0.857	0.853	0.172	0.170	0.841	0.170	0.168	0.832	0.169	0.166
p	0.700	0.645	0.030	0.029	0.639	0.029	0.029	0.623	0.030	0.030
$P_{g_m\|g_{III}}$	0.200	0.227	0.077	0.072	-	-	-	-	-	-
$P_{g_m\|g_{IV}}$	0.100	0.084	0.044	0.046	-	-	-	-	-	-
					Higher probabilities of masked causes of deaths					
α_{12}	2.000	1.958	0.065	0.061	1.922	0.058	0.058	1.880	0.056	0.057
β_{12}	4.000	4.059	0.289	0.259	4.201	0.270	0.270	4.305	0.282	0.280
α_{13}	3.500	3.156	0.227	0.258	3.316	0.286	0.275	3.995	0.544	0.505
β_{13}	3.000	3.534	0.441	0.366	3.145	0.403	0.377	3.095	0.506	0.473
α_{23}	2.500	2.509	0.144	0.146	2.513	0.144	0.145	2.530	0.145	0.145
β_{23}	1.500	1.523	0.095	0.097	1.534	0.095	0.097	1.548	0.095	0.097
ϕ	0.857	0.862	0.165	0.164	0.841	0.161	0.160	0.821	0.157	0.157
p	0.700	0.655	0.029	0.028	0.639	0.028	0.028	0.606	0.030	0.029
$P_{g_m\|g_{III}}$	0.400	0.457	0.084	0.080	-	-	-	-	-	-
$P_{g_m\|g_{IV}}$	0.300	0.265	0.049	0.051	-	-	-	-	-	-

Table 2.2: Monte-Carlo simulation study results under the scenarios (a) masked causes are present, (b) causes of deaths are fully observed, (c) masked causes are discarded. Simulation study was conducted using $1,000$ replications with sample size $n = 400$. Censoring rates are approximately 59% for $1 \to 2$ transition, 89% for $1 \to 3$ transition and 66% for $2 \to 3$ transition. Est refers to estimate of the corresponding parameter, $Mean(Est)$ refers to the mean of the estimates, $SD(Est)$ refers to the standard deviation of the estimates, $\widehat{\overline{SE}}(Est)$ refers to the average standard error estimates over $1,000$ replications.

2.5 Application to Colon Cancer Data

We considered the data from a clinical trial on patients with colon cancer provided by Moertel et al. (1990). The data is available in the "survival" package in R. After surgical removal of their diseased tissue, early diagnosed patients were randomly assigned into a control placebo group or a drug therapy group, which can be Levamisole plus 5-FU group or Levamisole alone group. Earlier studies were interested in assessing effectiveness of therapies on time to cancer recurrence as well as on overall survival time (Moertel et al., 1990). Later, Lin et al. (1999) and Lawless and Yilmaz (2011) studied whether there is any effect of the treatments on survival time after cancer recurrence. They considered the sequential modeling of time from registration to cancer recurrence and time from recurrence to death. Multi-state modeling of this data with an all cause mortality state was considered in de Uña-Álvarez and Meira-Machado (2015) and Meira-Machado and Sestelo (2019). In our study, we fit the multi-state model in Figure 2.1 with the cancer death state and aim to assess the performance of the proposed estimation method when there are masked causes of deaths.

There are 619 patients in the combination of the control placebo and the Levamisole plus 5-FU treatment groups in the clinical trial. We excluded 31 patients who had stage 4 colon cancer at diagnosis since they already had cancer recurrence at diagnosis. For illustration purposes, we considered the 588 patients as a single group and ignored the treatment differences. Among 588 patients, 276 patients had cancer recurrence and 242 patients died after experiencing recurrence. By the end of the study, 25 patients died without experiencing cancer recurrence. Time to death for those 25 individuals varies from 23 days to 2789 days. Causes of their death were not recorded. These patients might be cured or not cured and their causes of death could be due to cancer or due to other reasons. There are also patients who neither had

recurrence nor died during the followup time. The maximum followup time is about 9 years.

For patients who had the corresponding event(s), time from registration to cancer recurrence and time from cancer recurrence to death or time from registration to death without recurrence were provided. For other patients, their followup times were provided. For patients who died without experiencing cancer recurrence, if there are more than 2 lymph nodes with detectable cancer, cause of death is assumed to be due to cancer ($\delta_{13} = 1$) and if the number of lymph nodes with detectable cancer is less than 3, then cause of death is assumed to be due to other reasons ($\delta_{14} = 1$). There are 10 patients who had more than 2 lymph nodes. Thus, deaths of 10 patients are assumed to be due to cancer ($\delta_{13} = 1$) and deaths of 15 patients are assumed to be due to other reasons ($\delta_{14} = 1$).

To assess the performance of our proposed method, we considered a scenario with masked causes of deaths. We randomly assigned the masked causes of deaths for each group g_{III} and g_{IV} . We obtained 2 masked causes from 10 patients who died due to cancer ($\delta_{13} = 1$) and 6 masked causes from 15 patients who died due to other causes ($\delta_{14} = 1$). Thus, the masked probabilities are $P_{g_m|g_{III}} = 0.2$ and $P_{g_m|g_{IV}} = 0.4$. We obtained the last observed cancer related event time as $\tau_{\max} = 2.695$, which allowed us to assign one of the masked causes as $\delta_{14} = 1$ and $\delta_{13} = 0$, since the individual died after $\tau_{\max}$.

There is empirical evidence that some subjects are cured and did not have a cancer related event in a long followup time. Lawless (2003) and Lawless and Yilmaz (2011) used a mixture-cure model with the log-logistic distribution to model time-to-recurrence for not cured. We also considered the mixture cure model in (2.1) with the log-logistic distribution to model the cause specific hazard function for cause $M = k$, $k = 2, 3$, in the form given in (2.8). The marginal distribution of T_2 for

subjects who experienced recurrence was considered as the log-logistic distribution with $F_2(t_2) = 1 - [1 + (t_2/\alpha_{23})\beta_{23}] - 1$ for $t_2 > 0$. We modeled the conditional joint distribution of T_1 and T_2 given that $M = 2$ with the Clayton copula in (2.23). For convenience, we scaled time-to-events t_1 and t_2 in days with $t_1 \times 10^{-3}$ and $t_2 \times 10^{-3}$.

First, in the absence of masked causes of deaths, we assessed the adequacy of the parametric model assumptions for the cumulative incidence functions of T_1 for subjects who had cancer recurrence ($F_{12}(t_1)$) and for subjects who died due to cancer without experiencing recurrence ($F_{13}(t_1)$) in Figure 2.3. We compared the parametric estimates of the cumulative incidence functions with their nonparametric estimates obtained by the method in Lin (1997). We observed that the parametric estimates of $F_{1k}(t_1)$ for $k = 2, 3$ are very close to their corresponding nonparametric estimate. Thus, we concluded that the log-logistic distribution provides an adequate fit for the cumulative incidence functions. In Figure 2.4, we compared the parametric estimates of $\Pr(T_2 > t_2 | T_1 \leq t_1, M = 2)$ with its nonparametric estimate using the method in Lin et al. (1999) to assess the parametric model assumption for $F_2(t_2)$ and the Clayton copula assumption in modeling the joint distribution of time-to-recurrence and time from recurrence to death. The parametric fits are close to the nonparametric fit. This provides support for the assumed distribution for $F_2(t_2)$ and the assumed copula function.

Table 2.3 shows the maximum likelihood and the two-stage pseudo-likelihood estimates of parameters and their standard error estimates (given in Section 2.2) in the absence of masked causes of deaths and the estimates obtained by the proposed EM algorithm (given in Section 2.3) in the presence of masked causes of deaths. Estimates of the standard errors for the two-stage pseudo-likelihood estimation were obtained by nonparametric bootstrap with $1,000$ samples. Estimates of standard errors under the masked causes scenario were obtained using the supplemented EM algorithm.

The maximum likelihood estimates of the uncured probability (p), parameters of the log-logistic cause specific hazard functions (α_{jk}, β_{jk} for $(j,k) = (1,2),(1,3),(2,3)$) and the copula parameter (ϕ) in Table 2.3 under the absence of masked causes of deaths indicate the following: About 47% of the considered sample is not susceptible for any cancer recurrence or cancer death. In the first 310 days after study registration there is a steep increase in the rate of cancer recurrence and later there is a rapid decrease. The rate of cancer death also initially increases but with a slower pace compared to the rate of cancer recurrence. The rate of death after experiencing cancer recurrence initially has a sudden increase in the first 210 days after recurrence and it later has a rapid decrease. Time to cancer recurrence and time from cancer recurrence to death have a mild level of positive but significant dependence with an estimated Kendall's $\tau = 0.2$. Having the Clayton copula parameter ϕ significantly different than 0 shows that the Markov assumption (Cook and Lawless, 2018, Chapter 2) between time to cancer recurrence and time from cancer recurrence to death does not hold and confirms using a semi-Markov model.

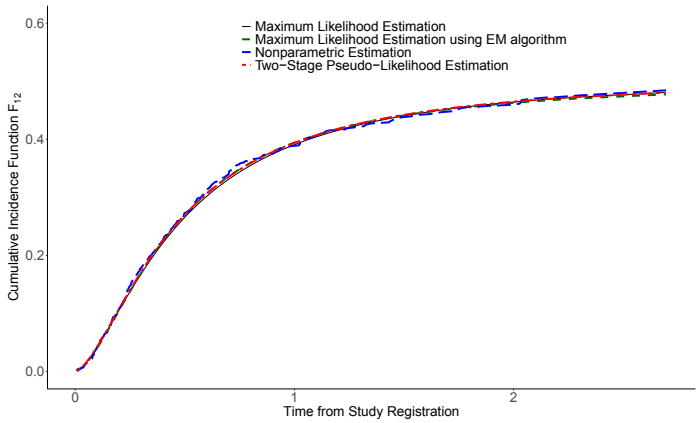
Results in Table 2.3 show that the proposed method in the presence of masked causes of deaths yielded estimates close to the maximum likelihood and two-stage pseudo-likelihood estimates in the absence of masked causes of deaths. These results are also supported by Figures 2.3 and 2.4 in which the maximum likelihood estimates of the cumulative incidence functions $F_{1k}(t_1)$ for $k = 2, 3$ and the conditional survival probability $\Pr(T_2 > t_2 | T_1 \leq t_1, M = 2)$ using the EM algorithm are given when masked causes of deaths exist and compared with their parametric and non-parametric estimates obtained under the absence of masked causes of deaths. The maximum likelihood methods under the presence and absence of masked causes of deaths provided almost the same standard error estimates except for one of the parameters in $1 \to 3$ transition. The two-stage pseudo-likelihood estimation method in

the absence of masked causes of deaths provided larger standard error estimates for the $1 \to 3$ transition than the maximum likelihood method. This may be due to the small number of events in transition $1 \to 3$.

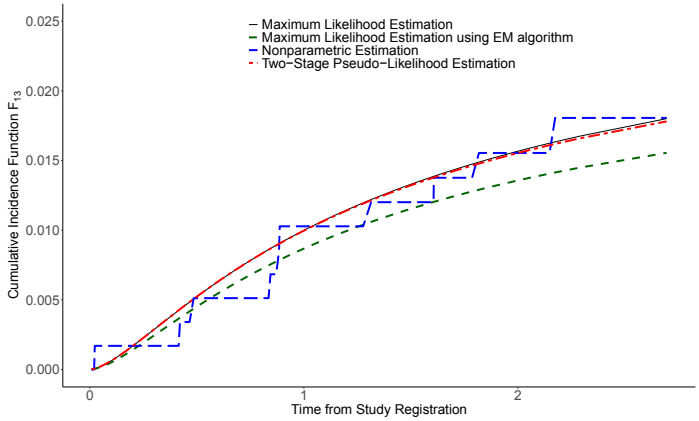
Although the number of masked causes of deaths is not large, when the individuals' data with masked causes of deaths were discarded and the estimation methods in Section 2.2 were applied to the remaining data, we obtained a biased estimate of the cumulative incidence function $F_{13}(t_1)$ in Figure 2.5. The maximum likelihood estimate of $F_{13}(t_1)$ in the lower panel of Figure 2.3 obtained through the EM algorithm which accounts for masked causes of deaths provided a better fit than the parametric estimates obtained when the data with masked causes of deaths are discarded in Figure 2.5.

	In the Absence of Masked Causes of Deaths				In the Presence of Masked Causes of Deaths	
	Maximum Likelihood		Two-Stage pseudo-likelihood		Using EM Algorithm	
Parameter	Est	$\widehat{SE}(Est)$	Est	$\widehat{SE}_{boot}(Est)$	Est	$\widehat{SE}_{SEM}(Est)$
α_{12}	0.494	0.039	0.483	0.039	0.471	0.037
β_{12}	1.533	0.098	1.548	0.098	1.599	0.105
α_{13}	8.021	3.979	7.748	5.357	7.769	4.099
β_{13}	1.526	0.357	1.538	0.618	1.582	1.244
α_{23}	0.414	0.031	0.411	0.033	0.400	0.030
β_{23}	1.393	0.077	1.392	0.091	1.443	0.083
ϕ	0.515	0.111	0.512	0.124	0.488	0.113
p	0.529	0.024	0.527	0.025	0.518	0.024
$P_{g_m\|g_{III}}$	.	.	.	.	0.203	0.122
$P_{g_m\|g_{IV}}$	.	.	.	.	0.398	0.109

Table 2.3: Maximum likelihood estimation and two-stage pseudo-likelihood estimation in the absence of masked causes of death and maximum likelihood estimation using EM algorithm in the presence of masked causes of death. $\widehat{SE}_{boot}$ are standard errors of the two-stage pseudo-likelihood estimator obtained by nonparametric bootstrap with 1,000 bootstrap samples.



12



13

Figure 2.3: Nonparametric, maximum likelihood and two-stage pseudo-likelihood estimates of $F_{12}(t_1)$ (upper panel) and $F_{13}(t_1)$ (lower panel) in the absence of masked causes of deaths and maximum likelihood estimates of $F_{12}(t_1)$ (upper panel) and $F_{13}(t_1)$ (lower panel) through EM algorithm in the presence of masked causes of deaths. t_1 is scaled to $t_1 \times 10^{-3}$.

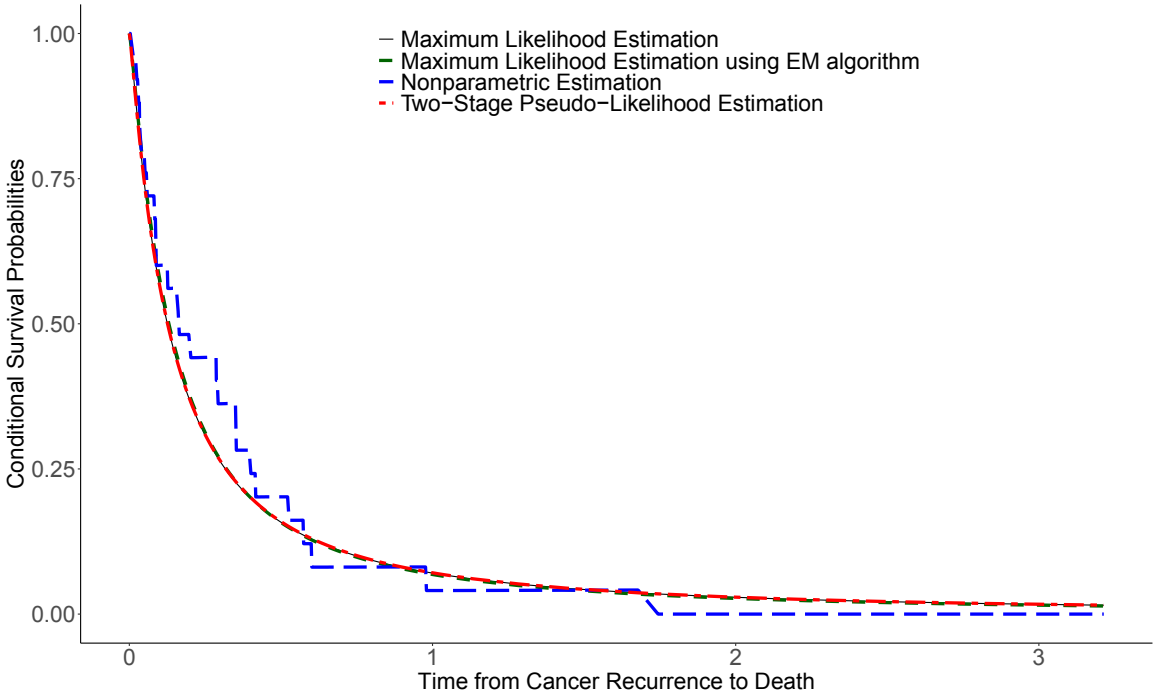


Figure 2.4: Nonparametric, maximum likelihood and two-stage pseudo-likelihood estimates of conditional probability $P(T_2 > t_2|T_1 \leq 0.1, M = 2)$ in the absence of masked causes of deaths and maximum likelihood estimates of $P(T_2 > t_2|T_1 \leq 0.1, M = 2)$ through EM algorithm in the presence of masked causes of deaths. t_1 and t_2 are scaled to $t_1 \times 10^{-3}$ and $t_2 \times 10^{-3}$.

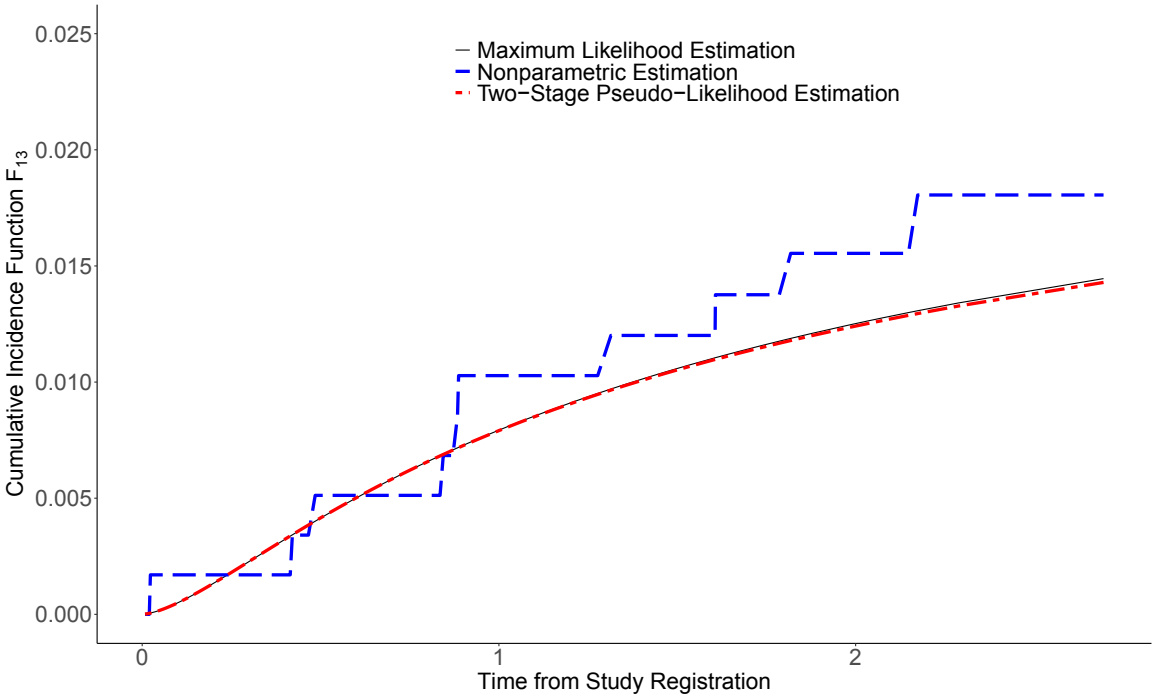


Figure 2.5: Maximum likelihood and two-stage pseudolikelihood estimates of $F_{13}(t_1)$ when the data with masked causes of deaths are discarded. Nonparametric estimate (Lin, 1997) of $F_{13}(t_1)$ is obtained in the absence of masked causes of deaths. t_1 is scaled to $t_1 \times 10^{-3}$.

Chapter 3

Introduction to Mediation Analysis Methods for Time-to-Event Outcomes

In many studies, establishing a causal relation is one of the crucial goals to achieve. Various areas including epidemiology, medicine and public heath give significant attention to understand causal relations and measure causal effects. For example, it is important to investigate how genetic markers affect primary phenotypes in genetic association studies. Measuring causal effects of environmental pollution on respiratory health outcomes is important in environmental studies. Studying the effect of a new drug treatment on a specific outcome of interest plays an important role in medicine.

To understand the true causal effect of an exposure on an outcome of interest, it is essential to consider other factors that may affect the outcome of interest. More specifically, it is important to consider indirect effects of the exposure on the outcome that may result from intermediate variables. Failure to account for indirect effects may give misleading results when measuring the direct effect of an exposure on the

outcome of interest. In order to measure causal effects, mediation analysis methods need to be applied.

In this chapter, we first review some regression models for time-to-event outcomes which play an important role in survival analysis. Then, we review some mediation analysis methods for time-to-event outcomes.

3.1 Regression Models for Time-to-Event Outcomes

We consider three modeling approaches to construct regression models for time-to-event outcomes: accelerated failure time models, proportional hazards models and additive hazards models. In this section, we give a review of the three regression models for time-to-event outcome.

3.1.1 Accelerated Failure Time Model

Accelerated failure time (AFT) model is one of the widely used regression models for time-to-event data. In AFT model, the time-to-event follows a distribution such as Weibull or log-normal and covariates' effects can be measured on the logarithm of time-to-event. It accelerates or decelerates the time-to-event.

We suppose that an individual has time-to-event T and a covariate vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^T$. We let $\boldsymbol{\beta}$ be a regression coefficient of $\boldsymbol{x}$ where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$. The logarithm of the time-to-event T is usually modeled in AFT model, since a real line on linear model can be used where $-\infty < Y = \log(T) < \infty$. Survival function of

$Y = \log(T)$ given $\boldsymbol{x}$ is of the form

$$S(Y|\boldsymbol{x}) = S_0\left(\frac{Y - u(\boldsymbol{x})}{\sigma}\right), \tag{3.1}$$

where $S_0(\epsilon)$ is independent of $\boldsymbol{x}$, σ is a scale parameter, and $u(\boldsymbol{x})$ is a function of $\boldsymbol{x}$. A linear specification $u(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x}$ is usually used. An AFT model can also be written in the form of

$$Y = \log(T) = \boldsymbol{\beta}^T\boldsymbol{x} + \sigma\epsilon, \tag{3.2}$$

where the error term ϵ is a random variable with the survival function $S_0(\epsilon)$ and σ is a scale parameter. One of the commonly used distributions for ϵ is the standard normal distribution which gives a log-normal regression model. Also, extreme value distribution and logistic distribution are other commonly used distributions of ϵ depending on the modeling purposes.

We let T_i be the time-to-event outcome which is subject to right-censoring and C_i be the censoring time for $i = 1, 2, \ldots, n$. We let $t_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$ for $i = 1, 2, \ldots, n$. We suppose $\boldsymbol{X}$ be the $n \times p$ matrix with entries x_{ij} and $u_i = u(\boldsymbol{x_i})$ where $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$. For example, $u(\boldsymbol{x_i})$ is $\boldsymbol{\beta}^T\boldsymbol{x}_i$ in (3.2). The likelihood function for the observed data $\{(t_i, \boldsymbol{x}_i, \delta_i), i = 1, 2, \ldots, n\}$ can be written as (Lawless, 2003)

$$L(\boldsymbol{\beta}, \sigma) = \prod_{i=1}^{n} \{\sigma^{-1} f_0(e_i)\}^{\delta_i} \{S_0(e_i)\}^{(1-\delta_i)}, \tag{3.3}$$

where $f_0(e) = -\frac{\partial S_0(e)}{\partial e}$ and $e_i = (\log(t_i) - u_i)/\sigma$.

To obtain the estimates of parameters in an AFT model, the maximum likelihood estimation method is used. The maximum likelihood estimate of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)^T$ is

obtained by solving $U(\boldsymbol{\theta}) = \mathbf{0}$ where $U(\boldsymbol{\theta})$ is the score function

$$U(\boldsymbol{\theta}) = \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{3.4}$$

The observed information matrix $\mathcal{I}(\hat{\boldsymbol{\theta}})$ can be obtained by

$$\mathcal{I}(\hat{\boldsymbol{\theta}}) = \left(\begin{array}{cc} \frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} & \frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \sigma} \\ \frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \sigma \partial \boldsymbol{\beta}^T} & \frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \sigma^2} \end{array} \right) \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}.} \tag{3.5}$$

Under the regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically $N_{(p+1)}(\mathbf{0}, \mathcal{J}_1{}^{-1}(\boldsymbol{\theta}))$ where $\mathcal{J}_1(\boldsymbol{\theta})$ is the Fisher information matrix and the observed information matrix $\mathcal{I}(\hat{\boldsymbol{\theta}})$ is a consistent estimator of $\mathcal{J}(\boldsymbol{\theta}) = n\mathcal{J}_1(\boldsymbol{\theta})$.

Although AFT model is flexible in modeling time-to-event distribution given covariates, it is a fully parametric model and results of the estimation can be sensitive to the distributional assumption on the time-to-event variable. If the distribution of time-to-event is not plausible, the results may be biased.

3.1.2 Proportional Hazards Model

Another important regression model for time-to-event data is the proportional hazards (PH) model. The PH model has been used in a large number of areas such as medicine, biochemistry and social sciences. The PH model assumes that a change in an explanatory variable has a multiplicative effect on the hazard rate by a constant.

The hazard function for time-to-event T given covariates $\boldsymbol{x}$ where $\boldsymbol{x}$ can be possibly time-dependent covariates is of the form

$$h(t|\boldsymbol{x}) = h_0(t)\gamma(\boldsymbol{x}), \tag{3.6}$$

where $h_0(t)$ is the baseline hazard function, $\gamma(\boldsymbol{x})$ is a positive valued function of $\boldsymbol{x}$. The baseline hazard function $h_0(t)$ is the hazard function when $\gamma(\boldsymbol{x}) = 1$. A common specification of $\gamma(\boldsymbol{x})$ is $\gamma(\boldsymbol{x}) = \exp(\boldsymbol{\beta}^T\boldsymbol{x})$. Then, $h_0(t)$ is the hazard function when $\boldsymbol{x} = \mathbf{0}$.

The survival function of T given $\boldsymbol{x}$ is of the form

$$S(t|\boldsymbol{x}) = S_0(t)^{\gamma(\boldsymbol{x})}, \tag{3.7}$$

where $S_0(t) = \exp(-\int_0^t h_0(u)du)$ is the baseline survival function. The survival function of T given $\boldsymbol{x}$ is the baseline survival function to a power. This means that the hazard ratio with different values of $\boldsymbol{x}$ is constant over time.

The PH model can be generalized using time-dependent covariates. Consider (3.6) with $\gamma(\boldsymbol{x}(t)) = \exp(\boldsymbol{\beta}^T\boldsymbol{x}(t))$ then the hazard function becomes

$$h(t|\boldsymbol{x}(t), \boldsymbol{\alpha}, \boldsymbol{\beta}) = h_0(t; \boldsymbol{\alpha}) \exp(\boldsymbol{\beta}^T\boldsymbol{x}(t)). \tag{3.8}$$

Then, the log-likelihood function based on a censored random sample $\{(t_i, \delta_i, \boldsymbol{x}_i(t_i), i = 1, 2, \ldots, n\}$ where $t_i = \min(T_i, C_i)$, C_i is right censoring time and $\delta_i = I(T_i \leq C_i)$ is given as

$$l(\alpha, \boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \{\log h_0(t_i; \boldsymbol{\alpha}) + \boldsymbol{\beta}^T\boldsymbol{x}(t)\} + \sum_{i=1}^{n} \{H_0(t_i; \boldsymbol{\alpha}) \exp(\boldsymbol{\beta}^T\boldsymbol{x}(t))\}, \tag{3.9}$$

where $H_0(t_i; \boldsymbol{\alpha})$ is a baseline cumulative hazard function. The maximum likelihood estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are obtained by maximizing (3.9). The asymptotic variance-covariance matrix of $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ can be estimated as $I(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})^{-1}$ which is obtained from the

information matrix $I(\boldsymbol{\alpha}, \boldsymbol{\beta})$ where

$$I(\boldsymbol{\alpha}, \boldsymbol{\beta}) = - \begin{pmatrix} \frac{\partial^2 l(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \boldsymbol{\alpha}^2} & \frac{\partial^2 l(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\beta}} \\ \frac{\partial^2 l(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}} & \frac{\partial^2 l(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \end{pmatrix}. \tag{3.10}$$

Under the regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, where $\theta = (\boldsymbol{\alpha}, \boldsymbol{\beta})^T$, is asymptotically $N_p(\mathbf{0}, {\mathcal{J}_1}^{-1}(\boldsymbol{\theta}))$ where $\mathcal{J}_1(\boldsymbol{\theta})$ is the Fisher information matrix and $\mathcal{J}(\boldsymbol{\theta}) = n\mathcal{J}_1(\boldsymbol{\theta})$ can be consistently estimated by $\mathcal{I}(\hat{\boldsymbol{\theta}}) = \mathcal{I}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$.

Cox PH Model and Semiparametric Estimation

A special case of PH model is the Cox PH model (Cox, 1972) where the baseline hazard $h_0(t)$ is left arbitrary. We define $Y(t)$ as "at risk" process $\{Y(t), 0 \leq t\}$ where

$$Y(t) = I(\text{process is observed at time } t). \tag{3.11}$$

Then, the hazard function of T given covariates $\boldsymbol{x}$ with $\gamma(\boldsymbol{x}) = \exp(\boldsymbol{\beta}^T \boldsymbol{x})$ can be written as

$$h(t|\boldsymbol{x}) = Y(t)\, h_0(t)\, \exp(\boldsymbol{\beta}^T \boldsymbol{x}), \tag{3.12}$$

where the baseline hazard $h_0(t)$ is unspecified.

The probability of having an event for ith individual at time t given the past and that the event is observed at that time is

$$\frac{Y_i(t)\, h_0(t)\, \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{\sum_{l=1}^n Y_l(t)\, h_0(t)\, \exp(\boldsymbol{\beta}^T \boldsymbol{x}_l)} = \frac{Y_i(t)\, \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{\sum_{l=1}^n Y_l(t)\, \exp(\boldsymbol{\beta}^T \boldsymbol{x}_l)}. \tag{3.13}$$

We let ordered observed times $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ assuming that there are no tied event times. We let $Y_i(t_{(i)})$ be 1 if individual i is at risk at time $t_{(i)}$ and

0 otherwise. Because the Cox PH model can easily be generalized to include time-varying covariates, we let $\boldsymbol{x}_i(t_{(i)})$ as the value of the covariates for the individual failing at $t_{(i)}$ (Fisher and Lin, 1999). Cox (1972) suggested the following partial likelihood function

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{Y_i(t_{(i)})\ \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i(t_{(i)}))}{\sum_{l=1}^{n} Y_l(t_{(i)})\ \exp(\boldsymbol{\beta}^T \boldsymbol{x}_l(t_{(i)}))}. \tag{3.14}$$

We let R_i as the risk set at $t_{(i)}$ which is the set of individuals who have not failed or censored until $t_{(i)}$. Then, (3.14) can be rewritten as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i(t_{(i)}))}{\sum_{l \in R_i} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_l(t_{(i)}))}. \tag{3.15}$$

Since $i \in R_l$ if and only if $Y_l(t_i) = 1$, the partial likelihood in (3.15) becomes

$$L(\boldsymbol{\beta}) = \left(\prod_{i=1}^{n} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i(t_i))}{\sum_{l=1}^{n} Y_l(t_i) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_l(t_i))} \right)^{\delta_i}. \tag{3.16}$$

Then, the logarithm of the partial likelihood function (3.16) can be written as

$$l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \{ \boldsymbol{\beta}^T \boldsymbol{x}_i(t_i) - \log[\sum_{l=1}^{n} Y_l(t_i)\ \exp(\boldsymbol{\beta}^T \boldsymbol{x}_l(t_i))] \}. \tag{3.17}$$

The maximum likelihood estimator of $\boldsymbol{\beta}$ can be obtained by maximizing (3.17). The score vector $\boldsymbol{S}(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$ is given by

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left(\boldsymbol{x}_i(t_i) - \frac{\sum_{l=1}^{n} Y_l(t_i)\ \exp(\boldsymbol{\beta}^T \boldsymbol{x}_l(t_i))\, \boldsymbol{x}_l(t_i)}{\sum_{l=1}^{n} Y_l(t_i)\ \exp(\boldsymbol{\beta}^T \boldsymbol{x}_l(t_i))} \right). \tag{3.18}$$

Under regularity conditions, $\hat{\boldsymbol{\beta}}$ has approximately $N(\boldsymbol{\beta}, E(I(\boldsymbol{\beta})^{-1}))$ distribution

where $I(\boldsymbol{\beta})$ is

$$\sum_{i=1}^{n}\delta_i\left(\frac{\sum_{l=1}^{n}Y_l(t_i)\,\exp(\boldsymbol{\beta}^T\boldsymbol{x}_l(t_i))\,\boldsymbol{x}_l(t_i)\,\boldsymbol{x}_l(t_i)^T}{\sum_{l=1}^{n}Y_l(t_i)\,\exp(\boldsymbol{\beta}^T\boldsymbol{x}_l(t_i))}\right.$$
$$\left.-\frac{\{\sum_{l=1}^{n}Y_l(t_i)\,\exp(\boldsymbol{\beta}^T\boldsymbol{x}_l(t_i))\,\boldsymbol{x}_l(t_i)\}\{\sum_{l=1}^{n}Y_l(t_i)\,\exp(\boldsymbol{\beta}^T\boldsymbol{x}_l(t_i))\,\boldsymbol{x}_l(t_i)\}^T}{\sum_{l=1}^{n}Y_l(t_i)\,\exp(\boldsymbol{\beta}^T\boldsymbol{x}_l(t_i))}\right). \tag{3.19}$$

$E(I(\boldsymbol{\beta}))$ can be consistently estimated by $I(\hat{\boldsymbol{\beta}})$.

While it does not require any distributional assumption for the baseline hazard function and thus flexible, it has some limitations. The PH model assumes the hazard ratio for any two sets of covariate values remain constant over time. If the proportional assumption is violated, the results may be misleading.

3.1.3 Additive Hazards Model

Alternative to AFT and PH models, additive hazards model is another choice in regression modeling of failure time in which the effect of covariates is measured directly on the hazard function. Additive hazards model was first introduced by Aalen (1980). In additive hazards model, it is assumed that the effects of covariates are additive on the hazard function. The hazard function for an individual with additive hazards model is defined by

$$h(t|\mathcal{H}(t)) = \alpha_0(t) + \alpha_1(t)x_1(t) + \alpha_2(t)x_2(t) + \cdots + \alpha_p(t)x_p(t), \tag{3.20}$$

where the history $\mathcal{H}(t)$ contains the covariate paths up to time t and $\alpha_j(t)$ are time-varying regression coefficients for $j = 0, 1, \ldots, p$.

We let $\mathbf{z}(t) = (1, x_1(t), x_2(t), \ldots, x_p(t))$ be a $(p+1)$-dimensional time-dependent

covariate vector and $\boldsymbol{\alpha}(t) = (\alpha_0(t), \alpha_1(t), \ldots, \alpha_p(t))^T$ be a $(p+1)$-dimensional time-varying regression coefficient vector. We let $Y(t)$ be a risk-indicator. The hazard function can be written as

$$h(t|\mathcal{H}(t)) = Y(t)\,\mathbf{z}(t)\,\boldsymbol{\alpha}(t). \tag{3.21}$$

We use the counting process notation since it provides a precise and concise way to formulate the additive hazards model and explain its properties. We explain the estimation of $\boldsymbol{\alpha}(t)$ in (3.21). We let $\boldsymbol{N}(t) = (N_1(t), N_2(t), \ldots, N_n(t))^T$ be counting processes for individuals $i = 1, 2, \ldots, n$. We let $\boldsymbol{X}(t)$ be $n \times (p+1)$ matrix as following

$$\boldsymbol{X}(t) = \begin{pmatrix} Y_1(t) & Y_1(t)\,x_{11}(t) & \ldots & Y_1(t)\,x_{1p}(t) \\ Y_2(t) & Y_2(t)\,x_{21}(t) & \ldots & Y_2(t)\,x_{2p}(t) \\ \vdots & \vdots & & \vdots \\ Y_n(t) & Y_n(t)\,x_{n1}(t) & \ldots & Y_n(t)\,x_{np}(t) \end{pmatrix}, \tag{3.22}$$

where $Y_i(t) = I[\text{individual } i \text{ is at risk at time } t]$. Then, we define an $n \times 1$ vector of martingales as

$$\boldsymbol{M}(t) = \boldsymbol{N}(t) - \int_0^t \boldsymbol{X}(u)\boldsymbol{\alpha}(u)du. \tag{3.23}$$

The equation in (3.23) implies that

$$d\boldsymbol{N}(t) = \boldsymbol{X}(t)d\mathcal{A}(t) + d\boldsymbol{M}(t), \tag{3.24}$$

where $d\mathcal{A}(t) = \boldsymbol{\alpha}(t)dt$ and $\mathcal{A}(t) = \int_0^t \boldsymbol{\alpha}(u)du$. Using the least squares estimation method, (3.24) has the solution for $d\mathcal{A}(t)$ as

$$d\hat{\mathcal{A}}(t) = (\boldsymbol{X}^T(t)\boldsymbol{X}(t))^{-1}\boldsymbol{X}^T(t)d\boldsymbol{N}(t). \tag{3.25}$$

Then, $\mathcal{A}(t)$ can be estimated by

$$\hat{\mathcal{A}}(t) = \int_0^t (\boldsymbol{X}^T(u)\boldsymbol{X}(u))^{-1}\boldsymbol{X}^T(u)d\boldsymbol{N}(u). \tag{3.26}$$

The variance-covariance matrix of $\mathcal{A}(t)$ is estimated by

$$\widehat{Var}(\hat{\mathcal{A}}(t)) = n\int_0^t \left((\boldsymbol{X}^T(u)\boldsymbol{X}(u))^{-1}\boldsymbol{X}^T(u)\right) diag(d\boldsymbol{N}(u)) \left((\boldsymbol{X}^T(u)\boldsymbol{X}(u))^{-1}\boldsymbol{X}^T(u)\right)^T. \tag{3.27}$$

Under some conditions (see, Martinussen and Scheike (2006), pp. 109-112, condition 5.1 and theorem 5.1.1), $\sqrt{n}(\hat{\mathcal{A}}(t) - \mathcal{A}(t))$ is asymptotically Normal with mean zero and the estimated asymptotic variance $\widehat{Var}(\hat{\mathcal{A}}(t))$ in (3.27).

Lin and Ying's Additive Hazards Model

Additional to Aalen's additive hazards model, Lin and Ying (1994) proposed an additive hazards regression model which was motivated by Cox semiparametric regression model. The hazard function in Lin and Ying's additive hazards model of ith independent individual is given by

$$h_i(t|\mathcal{H}(t)) = h_0(t) + \boldsymbol{\alpha}^T\boldsymbol{x}_i(t), \tag{3.28}$$

where $h_0(t)$ is an unknown and unspecified baseline hazard function, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T$ is a p-dimensional regression coefficient and $\boldsymbol{x}_i(t) = (x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t))^T$ is a time-dependent covariate vector for the ith individual. Lin and Ying's (1994) additive hazards model assumes the regression coefficients are constant over time.

We let $H_0(t) = \int_0^t h_0(u)du$ and $H(t; \boldsymbol{x}_i(t)) = \int_0^t h(u; \boldsymbol{x}_i(u))du$. The hazard function

can be written as

$$Y_i(t)\, dH(t; \boldsymbol{x}_i(t)) = Y_i(t) \left(dH_0(t) + \boldsymbol{\alpha}^T \boldsymbol{x}_i(t) \right). \tag{3.29}$$

Similar to (3.23), the counting process $N_i(t)$ can be decomposed into

$$N_i(t) = M_i(t) + \int_0^t Y_i(u)\, dH(u; \boldsymbol{x}_i). \tag{3.30}$$

Then, the estimating function to estimate $\boldsymbol{\alpha}$ is given by Lin and Ying (1994) as

$$S(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \int_0^{\infty} \boldsymbol{x}_i(t) \{ dN_i(t) - Y_i(t) d\hat{H}_0(\alpha, t) - Y_i(t) \boldsymbol{\alpha}^T \boldsymbol{x}_i(t) dt \}, \tag{3.31}$$

where

$$\hat{H}_0(\boldsymbol{\alpha}, t) = \int_0^t \frac{\sum_{j=1}^n dN_j(u) - Y_j(u) \boldsymbol{\alpha}^T \boldsymbol{x}_j(u) du}{\sum_{j=1}^n Y_j(u)}. \tag{3.32}$$

After some algebra, the estimating function (3.31) can be shown as

$$S(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \int_0^{\infty} \left(\boldsymbol{x}_i(t) - \bar{\boldsymbol{x}}(t) \right) \{ dN_i(t) - Y_i(t) \boldsymbol{\alpha}^T \boldsymbol{x}_i(t) dt \}, \tag{3.33}$$

where

$$\bar{\boldsymbol{x}}(t) = \frac{\sum_{j=1}^n Y_j(t) \boldsymbol{x}_j(t)}{\sum_{j=1}^n Y_j(t)}. \tag{3.34}$$

The estimator of $\boldsymbol{\alpha}$ can be obtained by solving $S(\hat{\boldsymbol{\alpha}}) = 0$. Then, the explicit solution is

$$\hat{\boldsymbol{\alpha}} = \left[\sum_{i=1}^{n} \int_0^{\infty} Y_i(t) \{ \boldsymbol{x}_i(t) - \bar{\boldsymbol{x}}(t) \} \{ \boldsymbol{x}_i(t) - \bar{\boldsymbol{x}}(t) \}^T dt \right]^{-1} \left[\sum_{i=1}^{n} \int_0^{\infty} \{ \boldsymbol{x}_i(t) - \bar{\boldsymbol{x}}(t) \} dN_i(t) \right]. \tag{3.35}$$

It can be easily seen that

$$S(\boldsymbol{\alpha}_0) = \sum_{i=1}^{n} \int_0^{\infty} (\boldsymbol{x}_i(t) - \bar{\boldsymbol{x}}(t))\, dM_i(t), \tag{3.36}$$

where $dM_i(t) = dN_i(t) - Y_i(t)\boldsymbol{\alpha}^T \boldsymbol{x}_i(t)\, dt$ is a martingale. Therefore, the random vector $n^{-1/2}\, S(\boldsymbol{\alpha}_0)$ converges to Normal distribution with mean zero and the variance covariance matrix that can be consistently estimated by (Andersen and Gill, 1982; Lin and Ying, 1994)

$$\boldsymbol{B} = n^{-1} \sum_{i=1}^{n} \int_0^{\infty} (\boldsymbol{x}_i(t) - \bar{\boldsymbol{x}}(t)) \left(\boldsymbol{x}_i(t) - \bar{\boldsymbol{x}}(t)\right)^T dN_i(t). \tag{3.37}$$

Then, $n(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)$ converges in Normal distribution with mean zero and the variance covariance matrix that can be consistently estimated by

$$\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}, \tag{3.38}$$

where

$$\boldsymbol{A} = n^{-1} \sum_{i=1}^{n} \int_0^{\infty} Y_i(t)\{\boldsymbol{x}_i(t) - \bar{\boldsymbol{x}}(t)\}\{\boldsymbol{x}_i(t) - \bar{\boldsymbol{x}}(t)\}^T dt. \tag{3.39}$$

3.2 Review of Mediation Analysis Methods for Time-to-Event Outcomes

Mediation analysis is considered to understand and measure a cause and effect relationship between observed data. It is one of the main goals in the fields of genetics, medicine and social sciences. There have been many methods developed for mediation analysis.

Structural equation model is widely used to infer the effect of an exposure in mediation analysis studies (Bollen, 1989; De Stavola et al., 2015). G-estimation is another widely considered method, alternative to structural equation model (Robins, 1986). Along with G-estimation method, marginal structural model with inverse probability weighting method (Robins et al., 2000; Madden et al., 2020) and structural nested model with inverse-weight estimator were developed (VanderWeele, 2009). Recently, VanderWeele and Tchetgen Tchetgen (2017) used G-estimation to infer the effects of an exposure on the time-to-event outcome in the presence of a time-varying mediator.

A Directed Acyclic Graph (DAG) is an effective tool for visualizing the causal relation (Greenland et al., 1999; Pearl, 1995). An illustrative example with the DAG in Figure 3.1 is given. There is a direct effect of an exposure X on the primary outcome of interest Y and indirect effect of X through the intermediate variable K on Y and there are unmeasured and measured factors U and L, respectively, confounding the effect of K on Y. The aim is to measure the direct effect of X on Y.

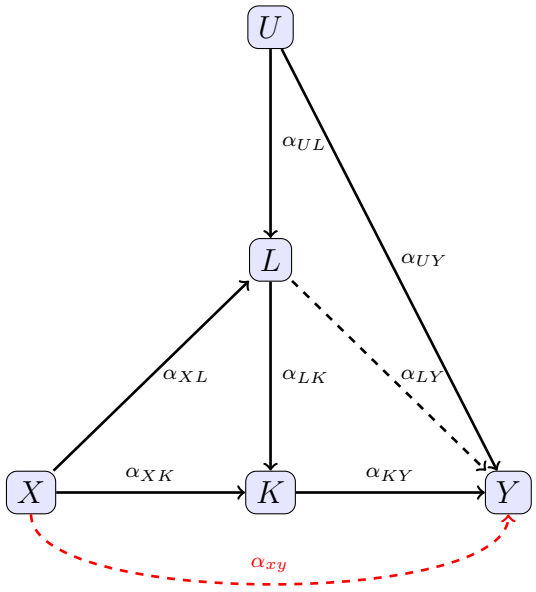


Figure 3.1: Directed acyclic graph with mediator K, measured confounder L and unmeasured confounder U. X is an exposure and Y is the outcome of interest. There is an indirect effect of X on Y through K. α_{XY} denotes the direct effect of X on Y.

Some methods are available to consider to measure the direct effect. As an extension of G-estimation method, sequential G-estimation method was applied to reveal the effects of genetic markers on continuous primary phenotypes (Vansteelandt et al., 2009; Vansteelandt, 2009). In sequential G-estimation method, two linear regression models are sequentially fitted to distinguish the direct genetic effect on primary phenotype (Vansteelandt, 2009). Along with sequential G-estimation method, Konigorski et al. (2018) introduced the method called causal inference based on estimating equations (CIEE) solving one set of estimating equations to remove indirect effects through intermediate phenotypes on a primary phenotype and to obtain the direct genetic effect on the primary phenotype.

In the following sections, we give a brief review of some available mediation analysis methods in which the DAG in Figure 3.1 can be modeled. The outcome of interest Y in Figure 3.1 is considered as time-to-event in Chapter 4.

3.2.1 Structural Equation Modeling

Structural equation modeling (SEM) is widely used to infer the effects of exposures in mediation analysis studies (Bollen, 1989; De Stavola et al., 2015). SEM was applied for time-to-event data. Recently, Vansteelandt et al. (2019) considered structural equation modeling to infer the effect of a treatment on time-to-event outcome under a setting with a repeatedly measured mediator.

SEM includes multiple regression models where a response variable in one regression model becomes an explanatory variable in another regression model. SEM may include unmeasured variables which are called latent variables. The details of SEM are given in Bollen (1989) and Pearl (1998).

Using SEM method, the DAG in Figure 3.1 can be for example modelled by

$$\begin{aligned} L =& \alpha_{XL}\, x + \epsilon_L, & \epsilon_L \sim N(0, \sigma_L^2) \\ K =& \alpha_{XK}\, x + \alpha_{LK}\, l + \epsilon_K, & \epsilon_K \sim N(0, \sigma_K^2) \\ Y =& \alpha_{XY}\, x + \alpha_{KY}\, k + \epsilon_Y, & \epsilon_Y \sim N(0, \sigma_Y^2) \end{aligned} \tag{3.40}$$

where ϵ_L, ϵ_K and ϵ_Y are error terms for L, K and Y, respectively. The model in (3.40) can be re-written in matrix form as

$$\tilde{\underset{\sim}{\boldsymbol{Y}}} = \boldsymbol{\alpha_Y}\tilde{\underset{\sim}{\boldsymbol{Y}}} + \underset{\approx}{\boldsymbol{\alpha_X}}\tilde{\boldsymbol{X}} + \underset{\approx}{\boldsymbol{\epsilon}}, \tag{3.41}$$

where $\tilde{\underset{\sim}{\boldsymbol{Y}}} = \begin{pmatrix} L \\ K \\ Y \end{pmatrix}$, $\tilde{\boldsymbol{X}} = x$, $\boldsymbol{\alpha_Y} = \begin{pmatrix} 0 & 0 & 0 \\ \alpha_{LK} & 0 & 0 \\ 0 & \alpha_{KY} & 0 \end{pmatrix}$, $\underset{\approx}{\boldsymbol{\alpha_X}} = \begin{pmatrix} \alpha_{XL} \\ \alpha_{XK} \\ \alpha_{XY} \end{pmatrix}$ and $\underset{\approx}{\boldsymbol{\epsilon}} = \begin{pmatrix} \epsilon_L \\ \epsilon_K \\ \epsilon_Y \end{pmatrix}$. The direct effect α_{XY} can be obtained using maximum likelihood estimation to minimize the function (Bollen, 1989, Chapter 4)

$$l_{ML} = \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + tr(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})) - \log|\boldsymbol{S}| - (p+q), \tag{3.42}$$

where $\boldsymbol{\theta}$ contains the parameters in $\boldsymbol{\alpha_Y}$ and $\underset{\approx}{\boldsymbol{\alpha_X}}$, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the variance-covariance matrix that is a function of $\boldsymbol{\theta}$, $\boldsymbol{S}$ is the sample variance-covariance matrix of $\tilde{\underset{\sim}{\boldsymbol{Y}}}$ and $\tilde{\boldsymbol{X}}$, and $p = 3$ and $q = 1$. The asymptotic variance-covariance matrix for the maximum likelihood estimator of $\boldsymbol{\theta}$ is (Bollen, 1989, Appendix 4B)

$$\left(\frac{2}{n-1}\right) E\left[\frac{\partial^2 l_{ML}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]^{-1}. \tag{3.43}$$

Unlike traditional regression models, SEM allows to model the direct and indirect relations among the measured and unmeasured (latent) variables. However, there are some limitations in SEM. If multivariate normal assumption of error terms is violated when maximum likelihood estimation is used under SEM, the estimates of the effects may not be accurate. Also, Konigorski et al. (2018) showed that when the DAG in Figure 3.1 was analyzed using SEM, inaccurate direct effect estimates can be obtained when there is significant amount of unmeasured confounding.

3.2.2 Sequential G-estimation Method

Vansteelandt (2009) proposed the sequential G-estimation method. Sequential G-estimation method fits two linear models sequentially. The direct effect of an exposure on the outcome of interest is obtained after its indirect effect is removed from the outcome.

In order to estimate the direct effect of an exposure X on the outcome Y in the DAG in Figure 3.1, sequential G-estimation method involves two stages as follows :

The first stage is to estimate the adjusted effect of K on Y by fitting the model

$$Y = \gamma_0 + \gamma_1 X + \gamma_2 K + \gamma_3 L + \epsilon_Y, \tag{3.44}$$

where ϵ_Y has mean 0 and constant variance. The estimators $\hat{\gamma_0}$, $\hat{\gamma_1}$, $\hat{\gamma_2}$ and $\hat{\gamma_3}$ are obtained by using ordinary least square estimation method.

In the second stage, the outcome is adjusted by removing the effect of the mediator K as

$$\tilde{Y} = Y - \hat{\gamma_2} K, \tag{3.45}$$

and the direct effect α_{XY} is estimated by fitting the following model with the adjusted

outcome $\tilde{Y}$ as

$$\tilde{Y} = \alpha_0 + \alpha_{XY}X + \epsilon_{\tilde{Y}}, \tag{3.46}$$

where $\epsilon_{\tilde{Y}}$ has mean 0 and constant variance. Least square estimation is used one more time to obtain the direct effect estimate $\hat{\alpha}_{XY}$. Due to the extra variability that occurs in two-stage estimation, the standard error of $\hat{\alpha}_{XY}$ can be estimated by a nonparametric bootstrap.

3.2.3 Sequential G-estimation Method using Aalen's Additive Hazards Model

Martinussen et al. (2011) proposed a sequential G-estimation method under Aalen's additive hazards model. It first postulates a model considering the DAG in Figure 3.1 for the conditional distribution of T given X, L and K by using Aalen's additive hazards model

$$h_{X,L,K}(t) = \psi_0(t) + \psi_X(t)X + \psi_L(t)L + \psi_K(t)K. \tag{3.47}$$

We let $R_i(t) = I(t \leq T_i)$ be the at-risk indicator for ith individual, $\boldsymbol{N}(t) = (N_1(t), N_2(t), \ldots, N_n(t))^T$ be the counting process, $\boldsymbol{X}(t) = (R_1(t)X_1, R_2(t)X_2, \ldots, R_n(t)X_n)^T$, $\boldsymbol{L}(t) = (R_1(t)L_1, R_2(t)L_2, \ldots, R_n(t)L_n)^T$ and $\boldsymbol{K}(t) = (R_1(t)K_1, R_2(t)K_2, \ldots, R_n(t)K_n)^T$.

Similar to the idea from the sequential G-estimation (Vansteelandt, 2009), the direct effect of X is obtained from a regression of the resulting adjusted outcome where the indirect effect of X through K is removed from the outcome. Adjusting the outcome involves adjusting the increment $dN(t)$ and the at-risk indicator $R(t)$. This can be done by substituting $dN(t)$ by $dN(t) - K(t)d\Psi_K(t)$ where $\Psi_K(t) = \int_0^t \psi_K(s)ds$

and $R(t)$ by $R(t)\exp(\Psi_K(t)K)$.

It involves two steps. The first step is to obtain the Aalen estimator of $\Psi_K(t)$ by using least square estimation

$$\hat{\boldsymbol{\Psi}}(t) = \int_0^t \{\boldsymbol{Y}^T(s)\boldsymbol{Y}(s)\}^{-1}\boldsymbol{Y}^T(s)d\boldsymbol{N}(s), \tag{3.48}$$

where $\boldsymbol{Y}(t)$ has ith row $R_i(t)(1, x_i, l_i, k_i)$. Then, the cumulative controlled direct effect $\boldsymbol{\Gamma}_{X,k}(t)$ is the second component of the following (see Martinussen et al., 2011, section 2.2)

$$\hat{\boldsymbol{\Gamma}}(t) = \int_0^t \boldsymbol{Z}^-_{\hat{H}}(s)\{d\boldsymbol{N}(s) - \boldsymbol{K}(s)d\hat{\boldsymbol{\Psi}}_K(s)\}, \tag{3.49}$$

where $\boldsymbol{Z}^-_{\hat{H}}(t) = (\boldsymbol{Z}^T(t)\boldsymbol{H}(t)\boldsymbol{Z}(t))^{-1}\boldsymbol{Z}^T(t)\boldsymbol{H}(t)$ is an estimate of a weighted generalized inverse of $\boldsymbol{Z}(t)$ with $\boldsymbol{H}(t)$ is replaced by $\hat{\boldsymbol{H}}(t)$ and $\boldsymbol{Z}(t)$ is the $n \times 2$ matrix with ith row $R_i(t)(1, X_i)$ and $\boldsymbol{H}(t)$ is the diagonal matrix with ith diagonal entry $R_i(t)\exp\{\Psi_K(t)K_i\}$.

Martinussen et al. (2011) showed that $\boldsymbol{W_n} = n^{-1/2}(\hat{\boldsymbol{\Gamma}}(t) - \boldsymbol{\Gamma}(t))$ is asymptotically normal with mean $\mathbf{0}$ and variance covariance matrix $\boldsymbol{\Sigma}(t)$. The variance covariance matrix $\boldsymbol{\Sigma}(t)$ can be consistently estimated. The result is shown in section A.3.2 of Martinussen et al. (2011).

3.2.4 Causal Inference Estimating Equation Method

Konigorski et al. (2018) introduced a causal inference method called causal inference based on estimating equations (CIEE). The general idea behind CIEE follows the two-stage sequential G-estimation method. However, their proposed method is a one-stage estimation. They solve estimating equations to obtain the indirect effect of exposure under an AFT model of outcome and to obtain direct effect of exposure on adjusted time-to-event. Since the estimating equation methodology was used, a closed-form

standard error estimator can be provided using CIEE.

First, suppose Y is a continuous random variable not subject to censoring. Then, in CIEE, the following model is considered under the DAG in Figure 3.1

$$Y_i = \alpha_0 + \alpha_k k_i + \alpha_x x_i + \alpha_l l_i + \epsilon_i, \tag{3.50}$$

where $\epsilon_i \sim N(0, \sigma_1^2)$ for $i = 1, 2, \ldots, n$. Here the outcome Y is considered as a completely observed normally distributed quantitative outcome.

The effect of the mediator K is removed from the outcome in order to block the indirect effect of X through K on the outcome

$$\tilde{Y}_i = Y_i - \overline{Y} - \alpha_k(k_i - \bar{k}), \tag{3.51}$$

where $\overline{Y} = \frac{1}{n}\sum_i^n Y_i$ and $\bar{k} = \frac{1}{n}\sum_i^n k_i$. Then, the direct effect of the exposure X on Y can be estimated using the model

$$\tilde{Y}_i = \alpha_0^{'} + \alpha_{XY} x_i + \epsilon_i^{'}, \tag{3.52}$$

where $\epsilon_i^{'} \sim N(0, \sigma_2^2)$ for $i = 1, 2, \ldots, n$.

We let $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \boldsymbol{\theta_2})^T$ where $\boldsymbol{\theta_1} = (\alpha_0, \alpha_k, \alpha_x, \alpha_l, \sigma_1)^T$ and $\boldsymbol{\theta_2} = (\alpha_0^{'}, \alpha_{XY}, \sigma_2)^T$. The following unbiased estimating equations are obtained to estimate $\boldsymbol{\theta}$

$$S(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \log L_1(\boldsymbol{\theta_1})}{\partial \boldsymbol{\theta_1}} \\ \frac{\partial \log L_2(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})}{\partial \boldsymbol{\theta_2}} \end{pmatrix} = \mathbf{0}, \tag{3.53}$$

where

$$L_1(\boldsymbol{\theta_1}) = \prod_{i=1}^{n} \left[\frac{1}{\sigma_1} \phi \left(\frac{y_i - \alpha_0 - \alpha_k k_i - \alpha_x x_i - \alpha_l l_i}{\sigma_1} \right) \right], \tag{3.54}$$

and

$$L_2(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \prod_{i=1}^{n} \left[\frac{1}{\sigma_2} \phi \left(\frac{Y_i - \overline{Y} - \alpha_k(k_i - \bar{k}) - \alpha_0^{'} - \alpha_{XY} x_i}{\sigma_2} \right) \right], \tag{3.55}$$

and $\Phi(\cdot)$, $\phi(\cdot)$ are the standard normal cumulative distribution function and standard normal probability density function, respectively.

Under mild regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normal with mean vector $\mathbf{0}$ and variance covariance matrix $C(\boldsymbol{\theta}) = A(\boldsymbol{\theta})^{-1} B(\boldsymbol{\theta}) [A(\boldsymbol{\theta})^{-1}]^T$ (White, 1982). The variance covariance matrix $C(\boldsymbol{\theta})$ can be consistently estimated by $C_n(\hat{\boldsymbol{\theta}})$ with $C_n(\boldsymbol{\theta}) = A_n(\boldsymbol{\theta})^{-1} B_n(\boldsymbol{\theta}) [A_n(\boldsymbol{\theta})^{-1}]^T$ where

$$A_n(\boldsymbol{\theta}) = -\frac{1}{n} \left(\frac{\partial S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right), \tag{3.56}$$

and

$$B_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} S_i(\boldsymbol{\theta}) S_i(\boldsymbol{\theta})^T. \tag{3.57}$$

Here, $S(\boldsymbol{\theta}) = \sum_{i=1}^{n} S_i(\boldsymbol{\theta})$. Therefore, a consistent estimator of $C(\boldsymbol{\theta})$ is given by

$$C_n(\hat{\boldsymbol{\theta}}) = A_n(\hat{\boldsymbol{\theta}})^{-1} B_n(\hat{\boldsymbol{\theta}}) [A_n(\hat{\boldsymbol{\theta}})^{-1}]^T. \tag{3.58}$$

Time-to-event outcome T was also considered. Here, (3.50) becomes an AFT model with Y be the logarithm of time-to-event outcome T. Under right-censored data setting, in order to remove the effect of the mediator K from the true underlying $Y^{'} = \log(T)$, the conditional expectation of Y given that Y is greater than the right-censoring time and given covariates is obtained as

$$Y_i^{'} = \delta_i \, y_i + (1 - \delta_i) \, E[Y_i | Y_i > y_i, k_i, x_i, l_i], \tag{3.59}$$

where $y_i = \log(t_i)$, $t_i = \min(T_i, C_i)$, C_i is the right censoring time and $\delta_i = I(T_i \leq C_i)$ for $i = 1, 2, \ldots, n$. Then, the adjusted outcome can be computed using

$$\tilde{Y}_i = Y_i^{'} - \overline{Y}_n^{'} - \alpha_k(k_i - \bar{k}), \tag{3.60}$$

where $\overline{Y}_n^{'} = \frac{1}{n}\sum_i^n Y_i^{'}$ and $\bar{k} = \frac{1}{n}\sum_i^n k_i$. Then, the direct effect of the exposure X on $Y^{'}$ can be estimated using the model

$$\tilde{Y}_i = \alpha_0^{'} + \alpha_{XY} x_i + \epsilon_i^{'}, \tag{3.61}$$

where $\epsilon_i^{'} \sim N(0, \sigma_2^2)$ for $i = 1, 2, \ldots, n$.

The estimating equations for estimating $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix}$ with $\boldsymbol{\theta_1} = (\alpha_0, \alpha_k, \alpha_x, \alpha_l, \sigma_1)^T$ and $\boldsymbol{\theta_2} = (\alpha_0^{'}, \alpha_{XY}, \sigma_2)^T$ are

$$S(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \log L_1(\boldsymbol{\theta_1})}{\partial \boldsymbol{\theta_1}} \\ \frac{\partial \log L_2(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})}{\partial \boldsymbol{\theta_2}} \end{pmatrix} = \mathbf{0}, \tag{3.62}$$

where

$$\begin{aligned} L_1(\boldsymbol{\theta_1}) = &\prod_{i=1}^{n} \left[\frac{1}{\sigma_1} \phi \left(\frac{y_i - \alpha_0 - \alpha_k k_i - \alpha_x x_i - \alpha_l l_i}{\sigma_1} \right) \right]^{\delta_i} \\ &\times \left[1 - \Phi \left(\frac{y_i - \alpha_0 - \alpha_k k_i - \alpha_x x_i - \alpha_l l_i}{\sigma_1} \right) \right]^{1-\delta_i}, \end{aligned} \tag{3.63}$$

and

$$L_2(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \prod_{i=1}^{n} \left[\frac{1}{\sigma_2} \phi \left(\frac{Y_i^{'} - \overline{Y^{'}} - \alpha_k(k_i - \bar{k}) - \alpha_0^{'} - \alpha_{XY} x_i}{\sigma_2} \right) \right], \tag{3.64}$$

and $\Phi(\cdot)$, $\phi(\cdot)$ are the standard normal cumulative distribution function and standard

normal probability density function, respectively. Under mild regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normal with mean vector $\mathbf{0}$ and variance covariance matrix $C(\boldsymbol{\theta})$ which can be estimated through (3.58).

3.3 Outline of Research

In this chapter, we reviewed some regression models for time-to-event data and some mediation analysis methods. In Chapter 4, we propose a mediation analysis method to make inference about direct exposure effects on time-to-event outcome under additive hazards model using estimating equations methodology. We examine properties of the proposed method and compare them with traditional survival analysis methods and the existing two-stage G-estimation method using additive hazards model by conducting simulation studies with various scenarios. A real-life application with the proposed method is provided.

Chapter 4

Estimation of Controlled Direct Exposure Effects on Time-to-Event Outcomes Using Additive Hazards Model

4.1 Introduction

In observational studies, it is well known that inference obtained about associations could be misleading due to possible confounding. Exposure effects could be mediated through other variables called mediators and the association between a mediator and the outcome can be confounded by measured and unmeasured factors. It is important to separate the direct exposure effects on outcome from indirect effects through mediators.

We consider the directed acyclic graph (DAG) in Figure 4.1 where there is a direct effect of an exposure X on time-to-event outcome T, and there is an indirect exposure

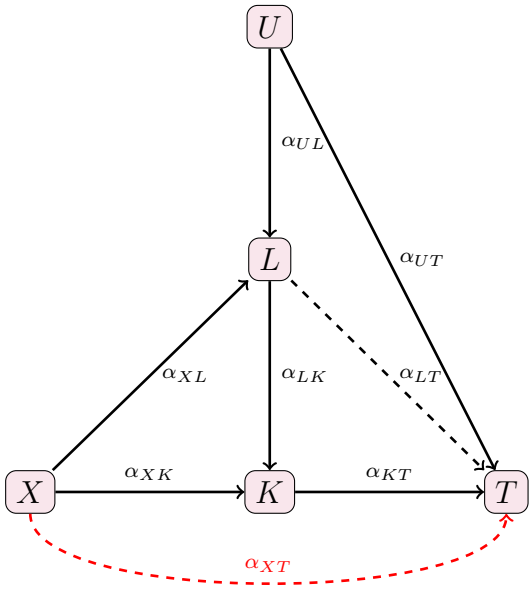


Figure 4.1: The overview of considered causal directed acyclic graph in this study. X is an exposure variable, T is the primary outcome of interest and K is a mediator. There is an indirect effect of X on T through K. α_{XT} denotes the direct effect of X on T. We assume that $\alpha_{LT} = 0$ so that L is a measured factor of K. U includes unmeasured factors and potential confounders that influence L and T.

effect of X on T through a mediator K. Measured factors L and unmeasured factors U are also included in the model which could confound the effect of K on time-to-event T. Our goal is to estimate and test the controlled direct exposure effect α_{XT} on T by removing the indirect effect of X on T through the mediator K.

Martinussen et al. (2011) inferred about α_{XT} using a two-stage G-estimation method under Aalen's additive regression method. We propose a one-stage method to estimate the controlled direct exposure effect on time-to-event outcome by using an estimating equation approach under Lin and Ying's additive hazards model (Lin and Ying, 1994). Lin and Ying's additive hazards model uses a semiparametric estimation procedure similar to the partial likelihood-based method for the proportional hazards (PH) regression model. There is no distributional assumption for the baseline hazard function under Lin and Ying's additive hazards model. The effect of K on T is

estimated using a set of unbiased estimating functions under Lin and Ying's additive hazards model. The controlled direct effect of X is estimated by another set of unbiased estimating functions using the adjusted T obtained by removing the effect of K from T. We consider a one-stage estimation by simultaneously solving the two sets of unbiased estimating equations. This allows us to use the Huber–White variance estimator to obtain the standard error of the controlled direct effect estimator. We use the well-developed asymptotic results for unbiased estimating equations to obtain the asymptotic results for the controlled direct effect estimator. To test the absence of the controlled direct effect, we use a Wald-type test statistic. We provide the asymptotic properties of the controlled direct effect estimator and the test statistic. A simulation study is carried out to assess the small sample properties of the method. We check the validity of the asymptotic results for finite samples. We apply the method to colon cancer data to estimate the controlled direct effect of having more than 4 positive lymph nodes on time from cancer recurrence to death.

The remainder of this chapter is organized as follows. Section 4.2 provides the notation and gives the novel method for estimation and testing of the controlled direct effect. Section 4.3 gives the simulation study results to assess the properties and the performance of the proposed method and to compare its performance with Lin and Ying's additive hazards model, Aalen's additive hazards model, Cox PH model and the sequential G-estimation method introduced by Martinussen et al. (2011). Section 4.4 illustrates how the proposed method is applied to a real life data. We make inference on the controlled direct effect of having more than 4 positive lymph nodes on time from cancer recurrence to death under a DAG model for colon cancer data where the intermediate variable is time from cancer diagnosis to cancer recurrence.

4.2 Notation and Method

We consider the directed acyclic graph in Figure 4.1. We let X_i, K_i and T_i be the exposure, mediator and time-to-event for individual i, $i = 1, 2, \ldots, n$, respectively. We let L_i be the collection of measured confounders for the ith individual. We denote $\boldsymbol{Z_i} = (X_i, L_i, K_i)^T$ as a covariate vector for ith individual. We assume that T_i is subject to right-censoring. The observed time-to-event and its event indicator becomes $t_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$ where C_i is the right censoring time for ith individual. We let $N_i(t)$ be the number of events over the interval $[0, t]$ for ith individual. This means in the current setting that $N_i(t) = I[T_i \leq t, \delta_i = 1]$. We let $dN_i(t)$ be the number of events in the small interval $[t, t + dt)$.

We assume that the hazard function at time t given $\boldsymbol{Z}$ is a linear combination of $\boldsymbol{Z}$ in the following (Lin and Ying, 1994)

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t) + \boldsymbol{\alpha}^T \boldsymbol{Z}, \tag{4.1}$$

where $\lambda_0(t)$ is the baseline hazard function and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$ is the regression coefficient vector associated with $\boldsymbol{Z}$.

Note that the estimation of (4.1) by Lin and Ying (1994) is based on an estimating equation using partial likelihood score function. The baseline hazard function $\lambda_0(t)$ is an unknown and unspecified function. The idea comes from semi-parametric proportional hazards regression model estimation (Cox, 1972), where the baseline hazard function is also unspecified.

Our estimation method follows the idea of the two-stage sequential G-estimation method (Vansteelandt et al., 2009) to measure the controlled direct effect of the exposure on the outcome. Unlike the two-stage sequential G-estimation method, our proposed method is a one-stage estimation method. Therefore, it is analytically

feasible to obtain the standard error estimator of the direct effect estimator. In our implementation, we construct two sets of unbiased estimating functions. The effect of the mediator K on time-to-event T is estimated by using the first set of estimating equations. The controlled direct effect of exposure X on T is estimated by using the second set of estimating equations which use adjusted counting process for the effect of the mediator K. We estimate the effect of the mediator and the controlled direct effect of exposure simultaneously by solving the two sets of unbiased estimating equations in one stage.

The first set of estimating functions to obtain the indirect effect of K on T are given as

$$S_1(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \int_0^{\infty} \boldsymbol{Z_i} \{dN_i(t) - Y_i(t) d\hat{\Lambda}_0(\boldsymbol{\alpha}, t) - Y_i(t) \boldsymbol{\alpha}^T \boldsymbol{Z_i} dt\}, \tag{4.2}$$

where $Y_i(t) = I[\text{individual } i \text{ is at risk at time } t]$ and $\Lambda_0(\boldsymbol{\alpha}, t)$ is estimated by

$$\hat{\Lambda}_0(\boldsymbol{\alpha}, t) = \int_0^t \frac{\sum_{j=1}^n \{dN_j(u) - Y_j(u) \boldsymbol{\alpha}^T \boldsymbol{Z_j} du\}}{\sum_{j=1}^n Y_j(u)}. \tag{4.3}$$

The estimating functions in (4.2) are equivalent to

$$S_1(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \int_0^{\infty} \left(\boldsymbol{Z_i} - \bar{\boldsymbol{Z}}(t)\right) \{dN_i(t) - Y_i(t) \boldsymbol{\alpha}^T \boldsymbol{Z_i} dt\}, \tag{4.4}$$

where

$$\bar{\boldsymbol{Z}}(t) = \frac{\sum_{j=1}^n Y_j(t) \boldsymbol{Z_j}}{\sum_{j=1}^n Y_j(t)}. \tag{4.5}$$

We estimate $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$ by solving $S_1(\boldsymbol{\alpha}) = \mathbf{0}$ for $\boldsymbol{\alpha}$ where $S_1(\boldsymbol{\alpha})$ is defined in equation (4.4). The adjustment for the effect of the intermediate variable K on $dN_i(t)$ is done by $dN_i(t) - \alpha_3 K_i dt$ by following the approach in Martinussen et al. (2011). The controlled direct effect of X on T, α_{XT}, can be estimated by solving the

following estimating equation,

$$S_2(\alpha_3, \alpha_{XT}) = \sum_{i=1}^{n} \int_0^{\infty} X_i\{dN_i(t) - \alpha_3 K_i\, dt - Y_i(t) d\hat{\Lambda'}_0(\alpha_3, \alpha_{XT}, t) - Y_i(t)\alpha_{XT} X_i dt\} = 0, \tag{4.6}$$

where

$$d\hat{\Lambda'}_0(\alpha_3, \alpha_{XT}, t) = \frac{\sum_{j=1}^{n} \left(dN_j(t) - \alpha_3 K_j\, dt - Y_j(t)\alpha_{XT} X_j\, dt\right)}{\sum_{j=1}^{n} Y_j(t)}. \tag{4.7}$$

Algebraically, the estimating function in (4.6) is equivalent to

$$S_2(\alpha_3, \alpha_{XT}) = \sum_{i=1}^{n} \int_0^{\infty} \left(X_i - \bar{X}(t)\right) \{dN_i(t) - Y_i(t)\alpha_3 K_i\, dt - Y_i(t)\alpha_{XT} X_i\, dt\}, \tag{4.8}$$

where

$$\bar{X}(t) = \frac{\sum_{j=1}^{n} Y_j(t) X_j}{\sum_{j=1}^{n} Y_j(t)}. \tag{4.9}$$

By combining the estimating functions in (4.4) and (4.8), we obtain the unbiased estimating equations $\tilde{S}(\boldsymbol{\theta}) = \begin{pmatrix} S_1(\boldsymbol{\alpha}) \\ S_2(\alpha_3, \alpha_{XT}) \end{pmatrix} = \mathbf{0}$ for a consistent estimation of the unknown parameter vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \alpha_{XT})^T$ with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$ with $S_1(\boldsymbol{\alpha})$ given in (4.4) and $S_2(\alpha_3, \alpha_{XT})$ given in (4.6).

The explicit solution for the estimator of $\boldsymbol{\alpha}$ is

$$\hat{\boldsymbol{\alpha}} = \left[\sum_{i=1}^{n} \int_0^{\infty} Y_i(t)(\boldsymbol{Z_i} - \bar{\boldsymbol{Z}}(t))(\boldsymbol{Z_i} - \bar{\boldsymbol{Z}}(t))^T dt\right]^{-1} \left[\sum_{i=1}^{n} \int_0^{\infty} (\boldsymbol{Z_i} - \bar{\boldsymbol{Z}}(t)) dN_i(t)\right], \tag{4.10}$$

and for the controlled direct exposure effect, α_{XT}, is

$$\begin{aligned} \hat{\alpha}_{XT} = &\left[\sum_{i=1}^{n} \int_0^{\infty} Y_i(t)(X_i - \bar{X}(t))^2\, dt\right]^{-1} \\ &\times \left[\sum_{i=1}^{n} \int_0^{\infty} (X_i - \bar{X}(t))(dN_i(t) - Y_i(t)\hat{\alpha}_3 K_i\, dt)\right]. \end{aligned} \tag{4.11}$$

Under regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ is asymptotically normal with mean vector $\mathbf{0}$ and variance-covariance matrix $C(\boldsymbol{\theta})$ which can be consistently estimated by $C_n(\hat{\boldsymbol{\theta}})$ (White, 1982) which is

$$C_n(\hat{\boldsymbol{\theta}}) = [A_n(\hat{\boldsymbol{\theta}})^{-1}]B_n(\hat{\boldsymbol{\theta}})[A_n(\hat{\boldsymbol{\theta}})^{-1}]^T, \tag{4.12}$$

where

$$A_n(\hat{\boldsymbol{\theta}}) = -\frac{1}{n}\left(\frac{\partial \tilde{S}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right)\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \qquad B_n(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}[\tilde{S}_{ji}(\hat{\boldsymbol{\theta}})\,\tilde{S}_{ki}(\hat{\boldsymbol{\theta}})^T]_{j,k=1,2,\ldots,p} \tag{4.13}$$

where $\tilde{S}_{ki}(\boldsymbol{\theta})$ is the kth element of $\tilde{S}_i(\boldsymbol{\theta})$ where $\tilde{S}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\tilde{S}_i(\boldsymbol{\theta})$ and $p = 4$.

For testing absence of the controlled direct effect, $H_0 : \alpha_{XT} = 0$, we use the Wald-type test statistic

$$Z = \frac{\hat{\alpha}_{XT}}{\widehat{SE}\{\hat{\alpha}_{XT}\}}, \tag{4.14}$$

where $\widehat{SE}\{\hat{\alpha}_{XT}\} = \frac{1}{\sqrt{n}}\sqrt{C_n(\hat{\boldsymbol{\theta}})_{4,4}}$. The asymptotic distribution of (4.14) is standard normal under the null hypothesis.

4.3 Simulation Study

We conducted a Monte Carlo simulation study to investigate properties of the proposed controlled direct effect estimator and to assess properties of the test statistic for testing the absence of the direct effect. Empirical type I error and empirical power of the Wald-type test statistic are examined for testing $H_0 : \alpha_{XT} = 0$ versus $H_a : \alpha_{XT} \neq 0$.

We generated the data $\{(t_i, \delta_i, l_i, k_i, x_i, u_i), i = 1, \ldots, n\}$ using the following setting

:

$$
\begin{aligned}
&U_i \sim N(\mu_U, \sigma_U^2), \\
&X_i \sim Ber(p), \\
&L_i = \alpha_{UL} u_i + \alpha_{XL} x_i + \epsilon_{L,i}, \qquad \text{where} \quad \epsilon_{L,i} \sim N(\mu_L, \sigma_L^2), \\
&K_i = \alpha_{XK} x_i + \alpha_{LK} l_i + \epsilon_{K,i}, \qquad \text{where} \quad \epsilon_{K,i} \sim N(\mu_K, \sigma_K^2),
\end{aligned}
\tag{4.15}
$$

where X is a binary exposure variable which takes the value 0 or 1, K is a mediator, L is an measured confounder, U is the unmeasured confounder under the DAG in Figure 4.1, and the time-to-event T_i was generated from

$$
\lambda(t_i|x_i, l_i, k_i, u_i) = \lambda_0 + \alpha_{XT}\, x_i + \alpha_{LT}\, l_i + \alpha_{KT}\, k_i + \alpha_{UT}\, u_i, \tag{4.16}
$$

assuming constant baseline hazard λ_0 for simplicity.

We considered 5 different scenarios. The overview of the scenarios is depicted with the DAGs in Figure 4.2. The models in Figure 4.2 are submodels of the DAG in Figure 4.1 where some of the effects are set to 0. Scenario 5 is to check robustness against model misspecification where there is a non-zero effect of L on T. To investigate the power of the test statistics, non-zero controlled direct effects of X on T are also considered for each scenario. Table 4.1 gives the effect sizes in (4.15) and (4.16) for scenarios in Figure 4.2.

We first considered the null hypothesis model that there is no direct effect of X on T ($\alpha_{XT} = 0$) for scenarios in Table 4.1. In each scenario, data was generated for $n = 1,000$ individuals with $m = 1,000$ replications. We generated right-censoring times C_i from the Uniform$(0.5, 15)$ distribution. The exposure X was generated from the Bernoulli distribution with $p = 0.25$ and $p = 0.50$. In each scenario, we considered $\lambda_0 = 0.1$, $\mu_U = 1, \sigma_U = 0.3, \mu_L = 0, \sigma_L = 0.3, \mu_K = 0$ and $\sigma_K = 0.3$. Under the

alternative hypothesis, we considered $\alpha_{XT} = 0.10$ and $\alpha_{XT} = 0.50$ for scenarios from 1 to 4 with the same values used for other parameters in the setting above.

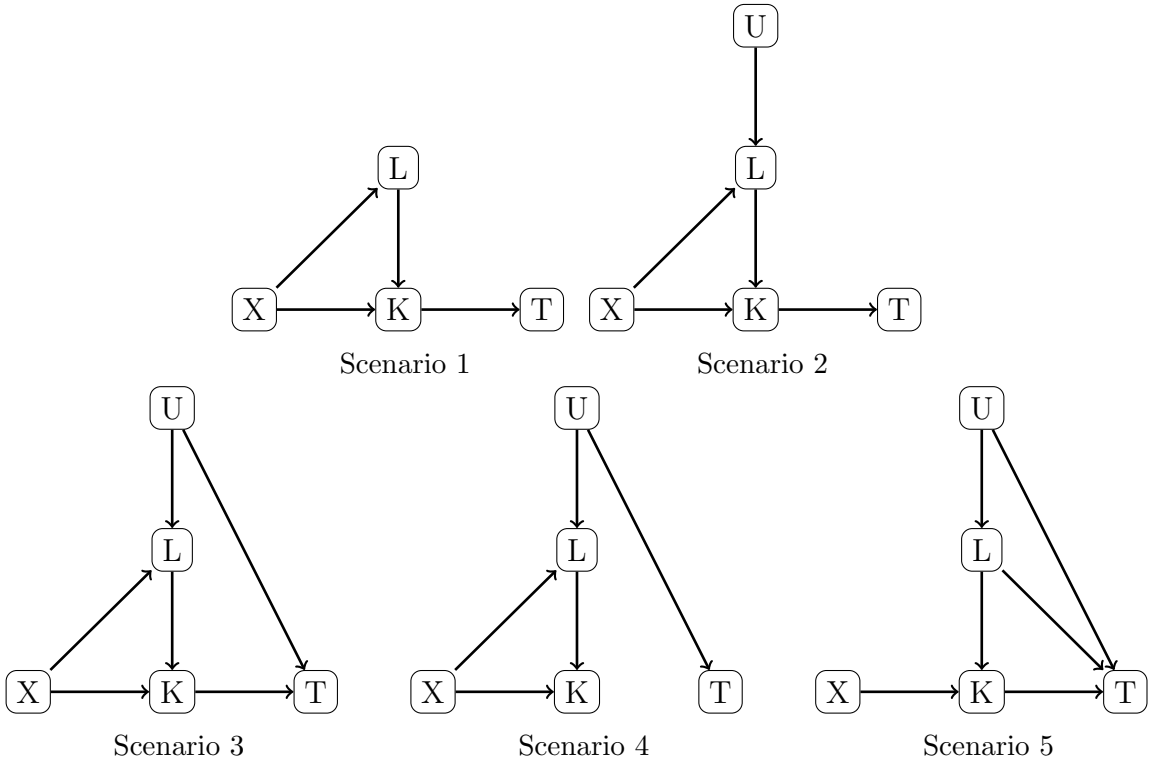


Figure 4.2: Overview of the scenarios

Scenario	α_{UL}	α_{XL}	α_{XK}	α_{LK}	α_{LT}	α_{UT}	α_{KT}
1	0	0.2	0.25	0.25	0	0	0.3
2	0.3	0.2	0.25	0.25	0	0	0.3
3	0.3	0.2	0.25	0.25	0	0.3	0.3
4	0.3	0.2	0.25	0.25	0	0.3	0
5	0.2	0	0.25	0.25	0.20	0.3	0.3

Table 4.1: The parameter values considered in each scenario

We compared the proposed method with the Aalen's additive hazards model, Lin and Ying's additive hazards model, Cox PH model and sequential G-estimation

method considered in Martinussen et al. (2011). Under Lin and Ying's additive hazards model, the hazard function considered is given as

$$\lambda(t_i|k_i, x_i, l_i) = \lambda_0(t_i) + \alpha_1 x_i + \alpha_2 l_i + \alpha_3 k_i. \tag{4.17}$$

The effect of X on T is estimated by solving the estimating equation

$$U(\boldsymbol{\alpha}) = \sum_{j=1}^{n} \int_0^{\infty} \left(\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t)\right) \{dN_i(t) - Y_i(t)\boldsymbol{\alpha}^T Z_i dt\} = \mathbf{0}, \tag{4.18}$$

where

$$\bar{Z}(t) = \frac{\sum_{j=1}^{n} Y_j(t)\boldsymbol{Z}_j}{\sum_{j=1}^{n} Y_j(t)}, \tag{4.19}$$

$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$ and $\boldsymbol{Z}_i = (X_i, L_i, K_i)^T$. Then, $H_0 : \alpha_1 = 0$ vs $H_A : \alpha_1 \neq 0$ was tested using the Wald-type test.

Under the Cox PH model, the effect of X on T was obtained from the hazard function

$$\lambda(t_i|k_i, x_i, l_i) = \lambda_0(t_i) \exp(\alpha_1 x_i + \alpha_2 l_i + \alpha_3 k_i), \tag{4.20}$$

and the Wald-type test was performed for testing the null hypothesis $H_0 : \alpha_1 = 0$.

Aalen's additive hazards model assumes that the hazard function as

$$\lambda(t_i|k_i, x_i, l_i) = \psi_0(t_i) + \psi_X(t_i)x_i + \psi_L(t_i)l_i + \psi_K(t_i)k_i, \tag{4.21}$$

where $\psi_0(t)$ is a baseline function and $\psi_j(t)$ is a regression coefficient that measures the time-varying association of factor j with T for $j = X, L, K$. The cumulative regression coefficient is

$$\Psi_j(t) = \int_0^t \psi_j(s)ds \qquad \text{for } j = X, L, K, \tag{4.22}$$

and $\boldsymbol{\psi}(t) = (\psi_0(t), \psi_X(t), \psi_L(t), \psi_K(t))^T$. The Aalen least squares estimator of $\boldsymbol{\Psi}(t) = \int_0^t \boldsymbol{\psi}(s)ds$ can be obtained from

$$\hat{\boldsymbol{\Psi}}(t) = \int_0^t \{\boldsymbol{J}^T(s)\boldsymbol{J}(s)\}^{-1}\boldsymbol{J}^T(s)d\boldsymbol{N}(s), \tag{4.23}$$

where $R_i(t) = I(t \leq T_i)$, $\boldsymbol{J}(t)$ has ith row $R_i(t)(1, x_i, l_i, k_i)$ and $\boldsymbol{N}(t) = (N_1(t), N_2(t), \ldots, N_n(t))^T$.

The sequential G-estimation method (Martinussen et al., 2011) assumes the additive hazards function in (4.21). Then the controlled direct effect is obtained from the adjusted counting process. We let $\boldsymbol{K}(t) = (R_1(t)K_1, R_2(t)K_2, \ldots, R_n(t)K_n)^T$, $\boldsymbol{G}(t)$ be the $n \times 2$ matrix with ith row $R_i(t)(1, X_i)$ and $\boldsymbol{H}(t)$ be the diagonal matrix with ith diagonal entry $R_i(t)\exp\{\Psi_K(t)K_i\}$. $\hat{\Psi}_K(t)$ is first obtained from (4.23). Then, the cumulative controlled direct effect $\Gamma_{X,k}(t)$ is the second component of the following (see Martinussen et al., 2011, section 2.2)

$$\hat{\boldsymbol{\Gamma}}(t) = \int_0^t \boldsymbol{G}^-_{\hat{H}}(s)\{d\boldsymbol{N}(s) - \boldsymbol{K}(s)d\hat{\Psi}_K(s)\}, \tag{4.24}$$

where

$$\boldsymbol{G}^-_{\hat{\boldsymbol{H}}}(t) = \{\boldsymbol{G}^T(t)\hat{\boldsymbol{H}}(t)\boldsymbol{G}(t)\}^{-1}\boldsymbol{G}^T(t)\hat{\boldsymbol{H}}(t).$$

Note that, in this simulation study, we obtained $\hat{\Gamma}_{X,k}(t)$ and $\hat{\Psi}_X(t)$ at $t = 1$, which provides $d\hat{\Gamma}_{X,k}(t)$ and $d\hat{\Psi}_X(t)$ for comparison. This was also considered in section 4.2 in Martinussen et al. (2011).

4.3.1 Simulation Results

The controlled direct effect (α_{XT}) of the exposure X on time-to-event outcome T and the standard error of the estimator of α_{XT} were estimated by the proposed method.

We also provided results for the other methods, Cox PH, Lin and Ying's additive hazards model, Aalen's additive hazards model and sequential G-estimation method considered in Martinussen et al. (2011). Under the null hypothesis $H_0 : \alpha_{XT} = 0$, Table 4.2 shows that our proposed method provides unbiased controlled direct effect estimates and valid standard error estimates of the controlled direct effect estimator.

The traditional survival analysis methods, additive hazard model by Lin and Ying (1994) and Aalen's additive hazards model do not provide valid inference when there is an unmeasured confounder affecting the time-to-event variable (see Table 4.2). Also, the estimates obtained from Cox PH model in Table 4.2 are biased throughout scenarios from 1 to 4 compared to other considered models. This suggests that Cox PH model does not provide valid inference when time-to-event comes from additive hazards model.

The proposed method gave unbiased results even under Scenario 5 where the model is misspecified. This suggests that the proposed method is robust when L affects T.

The sequential G-estimation method provided unbiased controlled direct effect estimates but gave larger standard deviations of controlled direct effect estimates than the mean standard error estimates obtained from the proposed method (Table 4.2). This suggests that the proposed method provides a more efficient controlled direct effect estimator than the sequential G-estimation method proposed by Martinussen et al. (2011) (Table 4.2).

We further considered the scenarios in Table 4.1 under the alternative hypotheses with $\alpha_{XT} = 0.10$ and $\alpha_{XT} = 0.50$ each with $p = 0.25$ and $p = 0.50$. Tables 4.3 and 4.4 display the results with $\alpha_{XT} = 0.10$ and $\alpha_{XT} = 0.50$ when $p = 0.25$, respectively. Tables 4.5 and 4.6 show the results with $\alpha_{XT} = 0.10$ and $\alpha_{XT} = 0.50$ when $p = 0.50$, respectively. The proposed method gave unbiased estimates of controlled direct effect and valid standard estimates when $\alpha_{XT} = 0.10$ and $\alpha_{XT} = 0.50$. Overall, larger

			Standard multiple regression method			Causal inference method	
Scenario	Censoring Rates %	p	Aalen's Least Square $\widehat{Est}$ $(\widehat{\mathrm{SE}}(\widehat{Est}))$ $[Sd(\widehat{Est})]$	Lin and Ying's Semiparametric $\widehat{Est}$ $(\widehat{\mathrm{SE}}(\widehat{Est}))$ $[Sd(\widehat{Est})]$	Cox PH $\widehat{Est}$ $(\widehat{\mathrm{SE}}(\widehat{Est}))$ $[Sd(\widehat{Est})]$	Proposed Method $\widehat{\alpha_{XT}}$ $(\widehat{\mathrm{SE}}(\hat{\alpha}_{XT}))$ $[Sd(\hat{\alpha}_{XT})]$	Sequential G-estimation $\widehat{Est}$ $[Sd(\widehat{Est})]$
1	14.20 %	0.25	7×10^{-5} (0.0615) [0.0596]	0.0016 (0.0386) [0.0367]	−0.0116 (0.0866) [0.0825]	0.0011 (0.0381) [0.0365]	−0.0006 [0.0601]
	13.28 %	0.50	2×10^{-5} (0.0537) [0.0521]	−0.0003 (0.0333) [0.0326]	−0.0024 (0.0776) [0.0754]	−0.0004 (0.0328) [0.0315]	−0.0007 [0.0524]
2	13.27 %	0.25	−0.0041 (0.0631) [0.0647]	0.0004 (0.0403) [0.0406]	−0.0126 (0.0861) [0.0868]	0.0008 (0.0399) [0.0399]	−0.0031 [0.0641]
	13.27 %	0.50	−0.0041 (0.0631) [0.0647]	0.0004 (0.0403) [0.0406]	−0.0126 (0.0861) [0.0868]	0.0008 (0.0399) [0.0399]	−0.0031 [0.0641]
3	6.75 %	0.25	−0.0182 (0.0876) [0.0871]	−0.0161 (0.0637) [0.0638]	−0.0283 (0.0834) [0.0832]	0.0001 (0.0631) [0.0633]	−0.0016 [0.0869]
	6.43 %	0.5	−0.0146 (0.0768) [0.0763]	−0.0167 (0.0556) [0.0546]	−0.0227 (0.0742) [0.0727]	0.0001 (0.0548) [0.0539]	0.0029 [0.0746]
4	14.93 %	0.25	−0.0192 (0.0562) [0.0553]	−0.0158 (0.0336) [0.0338]	−0.0430 (0.0873) [0.0877]	0.0008 (0.0333) [0.0339]	−0.0024 [0.0548]
	14.95 %	0.50	−0.0175 (0.0501) [0.0492]	−0.0172 (0.0298) [0.0290]	−0.0450 (0.0778) [0.0757]	−0.0006 (0.0295) [0.0289]	−0.0014 [0.0495]
5	6.50 %	0.25	0.0054 (0.0864) [0.0870]	0.0046 (0.0629) [0.0647]	0.0013 (0.0805) [0.0825]	0.0046 (0.0627) [0.0648]	0.0056 [0.0876]
	6.29 %	0.50	2×10^{-5} (0.0750) [0.0756]	0.0014 (0.0544) [0.0570]	0.0014 (0.0712) [0.0746]	0.0011 (0.0542) [0.0570]	−0.0003 [0.0762]

Table 4.2: Empirical mean and standard deviation of estimates of coefficients of α_{XT} and their mean standard error estimates when $\alpha_{XT} = 0$. Est stands for the estimate of the coefficient of X. Data was generated for $n = 1,000$ individuals over the $m = 1000$ replicates.

standard error estimates were obtained when $\alpha_{XT} = 0.50$ (Table 4.3 versus Table 4.4 and Table 4.5 versus Table 4.6) and smaller standard error estimates were obtained when $p = 0.50$ (Table 4.3 versus Table 4.5 and Table 4.4 versus Table 4.6). The standard error estimates from the proposed method are smaller than the standard deviations of the estimates obtained using the sequential G-estimation method proposed by Martinussen et al. (2011). The results show that Aalen's least square estimation method and Lin and Ying's semiparametric estimation method are not valid methods to estimate the controlled direct effect when an unmeasured confounder exists.

Also, the estimates obtained from Cox PH model in Tables 4.3, 4.4, 4.5 and 4.6 gave biased results throughout scenarios from 1 to 4. This indicates that Cox PH model does not provide valid inference when time-to-event comes from additive hazards model.

The proposed method gave empirical type I error close to 5% throughout all scenarios considered, while Lin and Ying's additive hazards model and Cox PH model yielded inflated type I error when an unmeasured confounder exists (Table 4.7).

Empirical power estimates are presented in Table 4.8. We considered the scenarios in Table 4.1 under the alternative hypotheses with $\alpha_{XT} = 0.10$ and $\alpha_{XT} = 0.20$ each with $p = 0.25$ and $p = 0.50$. Power gets higher as p gets higher. Overall, the proposed method gave higher power than all traditional methods across all considered scenarios even in the scenarios where two traditional methods had inflated type I error estimates. This suggests that our proposed method is the only currently available powerful test for the controlled direct effect.

			Standard multiple regression method			Causal inference method	
Scenario	Censoring Rates %	p	Aalen's Least Square $\widehat{Est}$ $(\widehat{\text{SE}}(\widehat{Est}))$ $[Sd(\widehat{Est})]$	Lin and Ying's Semiparametric $\widehat{Est}$ $(\widehat{\text{SE}}(\widehat{Est}))$ $[Sd(\widehat{Est})]$	Cox PH $\widehat{Est}$ $(\widehat{\text{SE}}(\widehat{Est}))$ $[Sd(\widehat{Est})]$	Proposed Method $\widehat{\alpha_{XT}}$ $(\widehat{\text{SE}}(\hat{\alpha}_{XT}))$ $[Sd(\hat{\alpha}_{XT})]$	Sequential G-estimation $\widehat{Est}$ $[Sd(\widehat{Est})]$
1	13.55 %	0.25	0.0974 (0.0673) [0.0682]	0.1009 (0.0444) [0.0454]	0.1852 (0.0864) [0.0878]	0.1010 (0.0439) [0.0447]	0.0985 [0.0674]
2	12.65 %	0.25	0.0996 (0.0689) [0.0675]	0.1019 (0.0461) [0.0464]	0.1795 (0.0857) [0.0854]	0.1020 (0.0458) [0.0457]	0.0997 [0.0676]
3	6.53 %	0.25	0.0807 (0.0931) [0.0925]	0.0859 (0.0695) [0.0694]	0.0964 (0.0834) [0.0826]	0.1021 (0.0689) [0.0691]	0.0977 [0.0922]
4	14.04 %	0.25	0.0792 (0.0620) [0.0618]	0.0826 (0.0392) [0.0396]	0.1857 (0.0865) [0.0866]	0.0990 (0.0389) [0.0399]	0.0961 [0.0619]
5	6.3 %	0.25	0.0962 (0.0919) [0.0940]	0.1024 (0.0687) [0.0703]	0.1181 (0.0804) [0.0822]	0.1022 (0.0685) [0.0697]	0.0962 [0.0946]

Table 4.3: Empirical mean and standard deviation of estimates of coefficients of α_{XT} and their mean standard error estimates when $\alpha_{XT} = 0.10$. Est stands for the estimate of the coefficient of X. Data was generated for $n = 1,000$ individuals over the $m = 1000$ replicates with $p = 0.25$.

			Standard multiple regression method			Causal inference method	
Scenario	Censoring Rates %	p	Aalen's Least Square $\widehat{Est}$ $(\widehat{\text{SE}}(\widehat{Est}))$ $[Sd(\widehat{Est})]$	Lin and Ying's Semiparametric $\widehat{Est}$ $(\widehat{\text{SE}}(\widehat{Est}))$ $[Sd(\widehat{Est})]$	Cox PH $\widehat{Est}$ $(\widehat{\text{SE}}(\widehat{Est}))$ $[Sd(\widehat{Est})]$	Proposed Method $\widehat{\alpha_{XT}}$ $(\widehat{\text{SE}}(\hat{\alpha}_{XT}))$ $[Sd(\hat{\alpha}_{XT})]$	Sequential G-estimation $\widehat{Est}$ $[Sd(\widehat{Est})]$
1	12.3 %	0.25	0.5096 (0.0912) [0.0917]	0.5053 (0.0684) [0.0677]	0.7303 (0.0874) [0.0864]	0.5057 (0.0681) [0.0667]	0.5101 [0.0912]
2	11.6 %	0.25	0.5003 (0.0926) [0.0934]	0.5033 (0.0702) [0.0717]	0.7045 (0.0870) [0.0884]	0.5026 (0.0699) [0.0710]	0.5009 [0.0927]
3	5.95 %	0.25	0.4912 (0.1177) [0.1234]	0.4918 (0.0931) [0.0945]	0.4777 (0.0839) [0.0839]	0.5082 (0.0926) [0.0949]	0.5084 [0.1233]
4	12.49 %	0.25	0.4816 (0.0852) [0.0849]	0.4849 (0.0629) [0.0635]	0.7871 (0.0868) [0.0881]	0.5016 (0.0626) [0.0633]	0.4995 [0.0846]
5	5.71 %	0.25	0.4992 (0.1167) [0.1153]	0.5052 (0.0928) [0.0921]	0.4886 (0.0811) [0.0792]	0.5051 (0.0926) [0.0921]	0.5004 [0.1157]

Table 4.4: Empirical mean and standard deviation of estimates of coefficients of α_{XT} and their mean standard error estimates when $\alpha_{XT} = 0.50$. Est stands for the estimate of the coefficient of X. Data was generated for $n = 1,000$ individuals over the $m = 1000$ replicates with $p = 0.25$.

			Standard multiple regression method			Causal inference method	
Scenario	Censoring Rates %	p	Aalen's Least Square $\widehat{Est}$ $(\widehat{\text{SE}}(\widehat{Est}))$ $[Sd(\widehat{Est})]$	Lin and Ying's Semiparametric $\widehat{Est}$ $(\widehat{\text{SE}}(\widehat{Est}))$ $[Sd(\widehat{Est})]$	Cox PH $\widehat{Est}$ $(\widehat{\text{SE}}(\widehat{Est}))$ $[Sd(\widehat{Est})]$	Proposed Method $\overline{\hat{\alpha}_{XT}}$ $(\widehat{\text{SE}}(\hat{\alpha}_{XT}))$ $[Sd(\hat{\alpha}_{XT})]$	Sequential G-estimation $\widehat{Est}$ $[Sd(\widehat{Est})]$
1	12.04 %	0.50	0.1008 (0.0577) [0.0600]	0.1013 (0.0374) [0.0379]	0.2088 (0.0776) [0.0778]	0.1010 (0.0368) [0.0371]	0.1009 [0.0597]
2	12.65 %	0.50	0.0996 (0.0689) [0.0675]	0.1019 (0.0461) [0.0464]	0.1795 (0.0857) [0.0854]	0.1020 (0.0458) [0.0457]	0.0997 [0.0676]
3	5.99 %	0.50	0.0837 (0.0805) [0.0822]	0.0855 (0.0595) [0.0599]	0.1063 (0.0743) [0.0746]	0.1023 (0.0587) [0.0590]	0.1020 [0.0810]
4	13.06 %	0.50	0.0839 (0.0541) [0.0528]	0.0845 (0.0337) [0.0342]	0.1942 (0.0772) [0.0783]	0.1008 (0.0334) [0.0332]	0.1003 [0.0514]
5	5.8 %	0.50	0.0990 (0.0789) [0.0813]	0.1002 (0.0583) [0.0596]	0.1224 (0.0712) [0.0727]	0.1001 (0.0581) [0.0590]	0.0990 [0.0812]

Table 4.5: Empirical mean and standard deviation of estimates of coefficients of α_{XT} and their mean standard error estimates when $\alpha_{XT} = 0.10$. Est stands for the estimate of the coefficient of X. Data was generated for $n = 1,000$ individuals over the $m = 1000$ replicates with $p = 0.50$.

			Standard Multiple Regression Method			Causal Inference Method	
Scenario	Censoring Rates %	p	Aalen's Least Square $\widehat{Est}$ $(\widehat{\text{SE}}(\widehat{Est}))$ $[Sd(\widehat{Est})]$	Lin and Ying's Semiparametric $\widehat{Est}$ $(\widehat{\text{SE}}(\widehat{Est}))$ $[Sd(\widehat{Est})]$	Cox PH $\widehat{Est}$ $(\widehat{\text{SE}}(\widehat{Est}))$ $[Sd(\widehat{Est})]$	Proposed Method $\overline{\hat{\alpha}_{XT}}$ $(\widehat{\text{SE}}(\hat{\alpha}_{XT}))$ $[Sd(\hat{\alpha}_{XT})]$	Sequential G-estimation $\widehat{Est}$ $[Sd(\widehat{Est})]$
1	12.41 %	0.50	0.5034 (0.0910) [0.0926]	0.5052 (0.0686) [0.0679]	0.7288 (0.0875) [0.0860]	0.5047 (0.0682) [0.0675]	0.5040 [0.0919]
2	11.6 %	0.50	0.5003 (0.0926) [0.0934]	0.5033 (0.0702) [0.0717]	0.7045 (0.0870) [0.0884]	0.5026 (0.0699) [0.0710]	0.5009 [0.0927]
3	4.9 %	0.50	0.4807 (0.0969) [0.0923]	0.4833 (0.0762) [0.0747]	0.4897 (0.0757) [0.0734]	0.4992 (0.0754) [0.0739]	0.4978 [0.0911]
4	9.97 %	0.50	0.4852 (0.0699) [0.0728]	0.4836 (0.0504) [0.0508]	0.7947 (0.0790) [0.0795]	0.4999 (0.0500) [0.0509]	0.5023 [0.0721]
5	4.77 %	0.50	0.5006 (0.0949) [0.0990]	0.5025 (0.0747) [0.0762]	0.5011 (0.0724) [0.0730]	0.5020 (0.0746) [0.0761]	0.5019 [0.0994]

Table 4.6: Empirical mean and standard deviation of estimates of coefficients of α_{XT} and their mean standard error estimates when $\alpha_{XT} = 0.50$. Est stands for the estimate of the coefficient of X. Data was generated for $n = 1,000$ individuals over the $m = 1000$ replicates with $p = 0.50$.

		Standard multiple regression method			Causal inference method
Scenario	p	Aalen's Least Square	Lin and Ying's Semiparametric	Cox PH	Proposed Method
1	0.25	4.30 %	3.40 %	3.50 %	4.10 %
	0.50	4.10 %	5.00 %	4.80 %	4.10 %
2	0.25	5.79 %	5.39 %	5.89 %	5.19 %
	0.50	5.79 %	5.39 %	5.89 %	5.19 %
3	0.25	5.20 %	6.30 %	6.70 %	5.10 %
	0.50	4.70 %	5.21 %	5.41 %	4.80 %
4	0.25	5.91 %	7.01 %	6.81 %	5.01 %
	0.50	5.81 %	8.31 %	8.11 %	5.41 %
5	0.25	4.60 %	5.30 %	5.20 %	6.30 %
	0.50	5.30 %	6.00 %	6.10 %	6.10 %

Table 4.7: Type I error estimates under the null hypothesis $H_0 : \alpha_{XT} = 0$. Data was generated for $n = 1,000$ individuals over the $m = 1000$ replicates.

			Standard multiple regression method		Causal inference method	
Scenario	α_{XT}	p	Aalen's Least Square	Lin and Ying's Semiparametric	Cox PH	Proposed Method
1	0.10	0.25	28.50 %	61.70 %	56.60 %	62.90 %
	0.20		80.10 %	98.50 %	98.10 %	99.20 %
	0.10	0.50	42.60 %	78.80 %	78.00 %	79.70 %
	0.20		89.90 %	99.99 %	99.99 %	99.99 %
2	0.10	0.25	28.20 %	60.40 %	55.90 %	60.10 %
	0.20		77.90 %	99.00 %	98.50 %	98.70 %
	0.10	0.50	39.10 %	72.80 %	72.00 %	75.50 %
	0.20		88.20 %	99.80 %	99.80 %	99.80 %
3	0.10	0.25	13.00 %	23.50 %	22.50 %	30.90 %
	0.20		47.00 %	71.60 %	70.00 %	80.90 %
	0.10	0.50	19.00 %	29.50 %	29.10 %	41.00 %
	0.20		57.40 %	83.00 %	83.10 %	87.60 %
4	0.10	0.25	23.90 %	54.90 %	57.20 %	72.00 %
	0.20		80.70 %	99.40 %	99.60 %	99.90 %
	0.10	0.50	33.30 %	71.50 %	72.20 %	86.20 %
	0.20		89.90 %	99.70 %	99.70 %	99.90 %
5	0.10	0.25	18.00 %	30.10 %	31.00 %	30.90 %
	0.20		51.10 %	80.20 %	80.40 %	80.00 %
	0.10	0.50	25.30 %	41.90 %	41.80 %	40.90 %
	0.20		68.50 %	89.40 %	89.20 %	89.20 %

Table 4.8: Power estimates under the alternative model when $\alpha_{XT} = 0.10$ and $\alpha_{XT} = 0.20$. Data was generated for $n = 1,000$ individuals over the $m = 1000$ replicates.

4.4 Application to Colon Cancer Data

We considered a clinical trial data on Duke's Stage C colon cancer patients treated with therapy with levamisole plus fluorouracil relative to a placebo which was also considered by Moertel et al. (1990), Lin et al. (1999), He and Lawless (2003) and Lawless and Yilmaz (2011). The data set is obtained from "colon" in R survival library. Patients were assigned to the treatment (levamisole only or levamisole plus fluorouracil) group or to the control (placebo) group. The data includes the patients' information on whether having more than four positive nodes or not, type of treatment received, time from registration to cancer recurrence and time from recurrence to death. It is of interest to infer the direct effect of having more than 4 positive lymph nodes on time to death after cancer recurrence which is not mediated through time to recurrence from the registration of the study. Therefore, we only considered the patients who experienced cancer recurrence in the data analysis.

There are 468 patients who experienced cancer recurrence. Among those patients, 414 of the patients died after cancer recurrence and 54 patients were censored after the recurrence. This gives the censoring rate of 11.53 % for observing death among those who experienced cancer recurrence. Whether or not having more than four positive lymph nodes (Node4) is a binary variable. There are 180 patients with more than four positive lymph nodes and 288 patients with less than or equal to four positive lymph nodes. For patients with cancer recurrence observed, time to cancer recurrence varies from 0.02 to 7.38 years and time to death from cancer recurrence varies from 0 to 7.47 years.

We considered standard regression methods which are Cox PH regression model, Lin and Ying's additive hazards model and Aalen's additive hazards model. Also, the sequential G-estimation method by Martinussen et al. (2011) and the proposed

method were considered for mediation analysis.

Patients were randomly assigned to the control group, treatment with levamisole alone group and treatment with levamisole plus fluorouracil group after stratification according to several factors including the primary lesion and the interval since surgery and whether or not having more than four positive lymph nodes (Node4). Therefore, the treatment could be affected by whether the number of lymph nodes is greater than 4 or not (Node4) because of the stratification. Also, in the treatment group, treatments were either deferred or discontinued if there were side effects (Moertel et al., 1990). This suggests that there could be unmeasured factors affecting the treatment.

Figure 4.3 describes the assumed DAG in which time from cancer recurrence to death T is the primary time-to-event outcome, having more than four positive nodes X may affect T and time from registration to cancer recurrence K could mediate the exposure effect X on time from recurrence to death T. The primary goal was to measure the controlled direct effect of having more than four positive nodes X on time from recurrence to death T that is not mediated through time from registration to cancer recurrence K. Based on preliminary analysis with Lin and Ying's additive hazards model to check whether the treatment L has an effect on time from reccurence to death T, we found that the treatment L does not have an effect on T.

We applied the sequential G-estimation method under the DAG in Figure 4.3 to make a comparison with our proposed method. The hazard function for the Aalen additive model is in (4.21) was considered. Note that the controlled direct effect $(\Gamma_{X,k}(t))$ using the sequential G-estimation method by Martinussen et al. (2011) is obtained by using the equation (4.24). We obtained $\hat{\Gamma}_{X,k}(t)$ at $t = 1$ which gives $d\hat{\Gamma}_{X,k}(t)$. With the proposed method, the hazard function in (4.17) is considered and the controlled direct effect α_{XT} is obtained from (4.11).

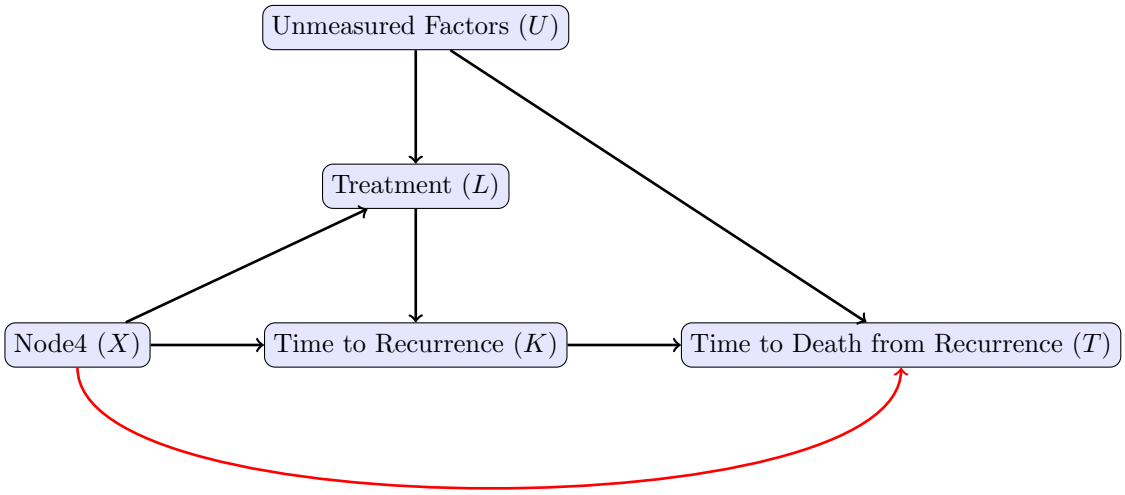


Figure 4.3: Overview of the assumed DAG for colon cancer data. X is whether or not having more than four positive nodes (Node4), K is time to recurrence, L is treatment received (levamisole plus fluorouracil and levamisole only) and T is time to death from recurrence.

Table 4.9 shows the results of the data analysis using Cox PH model, Lin and Ying's additive hazards model, Aalen's additive hazards model, sequential two-stage G-estimation method and the proposed method. To test the absence of the effect of the factors on time from recurrence to death T including the controlled direct effect of X on T, Wald-type test statistics were used in each model. The estimates of the effect of having more than four positive nodes X on time from cancer recurrence to death T under Aalen's additive hazards model and sequential G-estimation method were obtained at $t = 1$ year in order to compare with other models. The effect of having more than 4 positive lymph nodes X on time from cancer recurrence to death T is statistically significant for all the methods considered. Positive estimates of the direct effect of having more than 4 positive lymph nodes on time to death from recurrence suggest that having more than 4 positive lymph nodes increases the probability of death after the recurrence. The estimates of the association between having more than 4 positive lymph nodes and time from cancer recurrence to death using standard

regression methods are slightly less than the estimates of the controlled direct effect of X on T using the sequential G-estimation method and the proposed method. This suggests that the estimation of the effect of having more than 4 positive lymph nodes on time from recurrence to death could be misled using standard regression methods assuming DAG in Figure 4.3 is correct.

Note that we conducted a proportionality test where a formal score test for time-dependent coefficient is used from "cox.zph()" function R survival library. The proportionality test shows global p-value being 0.0065 which indicates the violation of the proportional assumption.

Methods	Model	Estimate	$\widehat{SE}$	Z	p-value
Standard Regression Methods	Cox PH	0.3857	0.1030	3.745	0.0001
	Lin and Ying's Additive Hazards	0.2700	0.0745	3.410	0.0006
	Aalen's Additive Hazards	0.2485	0.1178	2.109	0.0349
Causal Inference Methods	Proposed Method	0.3402	0.0733	4.636	3.5×10^{-6}
	Sequential G-estimation	0.3221	0.1064	3.027	0.0024

Table 4.9: Estimates of association between X and T using Cox PH regression model, Lin and Ying's additive hazards model and Aalen's additive hazards model and estimates of controlled direct effect of X on T using the sequential two-stage G-estimation method and the proposed method. The standard error estimates of sequential G-estimation and Aalen's least square estimation were obtained by a nonparametric bootstrap based on $B = 500$ resamples.

Chapter 5

Summary and Conclusions

5.1 Statistical Inference in Multi-State Semi-Markov Models with a Cured Fraction

In certain cancer types patients who were treated for their primary cancer might be cured or might be susceptible to experience cancer related events. Cured individuals do not experience any cancer related event, and eventually die due to other causes. Individuals who are not cured may die after experiencing cancer recurrence or without experiencing any recurrence. Cure status is a partially latent variable and is only known if a disease related event, cancer recurrence or cancer death, is observed. In this study, we considered the multi-state model in Figure 2.1 with the initial "Treatment" state which represents patients who have been treated for their primary cancer. It includes both "cured" and "not cured" states. Not cured patients may have a transition from the "Treatment" state to the "Cancer Recurrence" state or to the "Cancer Death" state. Both cured and not cured individuals may die due to other causes. Although the data generation process includes the transition from the "Treatment" state to the "Death Due to Other Causes" state for both cured and

not cured individuals (Figure 2.2), our interest only lies in modeling the transitions to the disease related events in Figure 2.1.

To model disease progression events, we considered a semi-Markov multi-state model including partially latent cured and not cured states. We used mixture cure model to model the distribution of time to first cancer related event. We took into account the possible dependence between successive event times which are time to cancer recurrence and time from cancer recurrence to death. We utilized the marginal modeling approach using a copula function to model the joint distribution of time to cancer recurrence and time from cancer recurrence to cancer death for not cured patients. Copula modeling allows to model marginal distributions of time-to-events separately from the dependence structure (Nelsen, 2006; Joe, 2014). Thus, the marginal distributions can be selected based on modeling needs and can be combined using a copula function to obtain the joint distribution of the sequential gap times. We applied the maximum likelihood estimation and the two-stage pseudo-likelihood estimation to obtain fits of each time-to-cancer related event distribution.

Previous studies, Conlon et al. (2014) and Beesley and Taylor (2018), considered multi-state cure modeling with an all cause mortality state. Conlon et al. (2014) considered fully parametric models for cure probability and transition intensities and applied Bayesian estimation for their semi-Markov multi-state cure model. They used conditional modeling approach to incorporate the effect of time to recurrence on time from recurrence to death in their semi-Markov modeling. Their Bayesian method requires assumptions on prior distributions for parameters in each of the regression model for transition intensities and probability of being cured. Under the conditional modeling approach, the obtained marginal distribution for time from recurrence to death may not be in a simple form and the effect of covariates may not be easily interpreted (Cook and Lawless, 2007, Chapter 4). Under a similar multi-state model

to Conlon et al. (2014) with an all-cause mortality state, Beesley and Taylor (2018) conducted maximum likelihood estimation using a Monte Carlo EM algorithm. They assumed the successive time-to-events, time to recurrence and time from recurrence to death are not associated. Markov assumption may lead to biased parameter estimators when there is dependence between sequential time-to-events (Yilmaz et al., 2017).

In our multi-state model in Figure 2.1, there is a cause specific "Cancer Death" state instead of an all cause death. It is important to consider meaningful endpoints like "Cancer Death" which provide time to disease related events having homogeneous definition and would lead to more meaningful and efficient estimators for measures of associations between prognostic factors and time-to-event traits. Our multi-state model requires information on if the cause of death for observed deaths is due to cancer or not. It is common to have masked causes for some observed deaths in cancer patient cohort data. In this case, in addition to partially latent cure status, there are also masked causes of deaths. For individuals with masked causes of death, their cure status are unknown too. We assume masked causes of deaths are missing at random. We use empirical evidence to determine true causes of certain masked causes of deaths. If a death occurs after the last observed cancer related event time, we assume the individual is cured and death is due to other causes than cancer. To estimate models, we use an EM algorithm assuming no cancer death after the last observed cancer related event time.

The multi-state semi-Markov modeling with latent cured or not cured states provides an informative approach to model cancer progression events. This would allow us to identify prognostic factors associated with being susceptible to a cancer event, timing of recurrence, time from recurrence to death, and time to cancer death without experiencing recurrence. The latter might allow us to detect the risk factors associated with serious adverse reactions of treatments used leading to death. Since the

marginal modeling approach was considered for time-to-event distributions, covariate information could be included using regression modeling for the cure probability and each time-to-event distribution separately. Parametric estimation methods in Sections 2.2 and 2.3 can then be applied. Semi-parametric regression models would be more robust to distribution misspecifications. Thus, the proposed estimation method accounting for masked causes of deaths will be extended to handle covariates through the adoption of Cox models for time-to-events.

5.2 Estimation of Controlled Direct Exposure Effects on Time-to-Event Outcomes Using Additive Hazards Model

We proposed a new method to estimate the controlled direct exposure effect on time-to-event outcomes using estimating equation approach under Lin and Ying's additive hazards model (Lin and Ying, 1994).

Using Lin and Ying's additive hazards model gives an advantage that distributional assumption on the baseline hazard function can be relaxed. The proposed method does not require the censoring to be adjusted, as it naturally adjusts for censoring (Martinussen et al., 2011). Multiple influencing factors can be included in the model which makes it flexible when modeling a variety of circumstances. The proposed method gives a consistent estimator of the controlled direct effect and its standard error in a variety of scenarios in the presence of confounding of indirect effects due to measured and unmeasured factors. Closed form of standard error estimates is obtained through Huber-White standard error estimation which is computationally much less intensive than using nonparametric bootstrap standard error estimation.

In addition, the proposed method yields a simple test statistic for testing absence of the direct effect.

Simulation studies justified that the proposed method removes the indirect exposure effect due to the mediator and provided unbiased direct exposure effects. It is shown that traditional multiple regression methods including Lin and Ying's additive hazards model, Aalen's additive hazards model and Cox PH model are not valid methods to estimate the controlled direct effect when there is an unmeasured confounding of indirect effects. In addition, the proposed method gave smaller standard error estimates of the controlled direct effect estimator than the standard deviation of the estimator obtained using the sequential G-estimation method proposed by Martinussen et al. (2011). Empirical type I error under the proposed test statistic is close to 5% throughout all scenarios considered in the simulation study while Lin and Ying's additive hazards model and Aalen's additive hazards model yielded inflated type I error rates when there is an unmeasured confounding of indirect effects. It is also shown that our proposed method is more powerful than the sequential G-estimation method to test the absence of the controlled direct effect.

We conducted an in-depth analysis of clinical trial data on Duke's Stage C colon cancer patients. We inferred the controlled direct effect of having more than 4 positive lymph nodes on time to death after cancer recurrence. Utilizing various statistical models including the Cox PH model, Lin and Ying's additive hazards model, Aalen's least square estimation method, the sequential G-estimation method by Martinussen et al. (2011) and our proposed approach, we found that the controlled direct effect estimates varied between standard regression methods and causal inference methods while Cox PH model is not valid for the data. This suggests that inferring the controlled direct effect requires valid causal inference methods such as our proposed

method or the sequential G-estimation method especially when there is an unmeasured confounding of indirect effects.

In this thesis, we considered the Lin and Ying's additive hazards model with constant coefficient parameters for simplicity. It is also important to explore controlled direct effect estimation with time-varying effects of factors on hazard function. Therefore, the additive hazards model with time-varying coefficients will be considered in a future work.

Bibliography

Aalen, O. (1978). Nonparametric estimation of partial transition probabilities in multiple decrement models. *The Annals of Statistics*, 6(3):534–545.

Aalen, O. (1980). A model for nonparametric regression analysis of counting processes. *In Mathematical Statistics and Probability Theory: Proceedings, Sixth International Conference, Wisła (Poland), 1978*, pages 1–25.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4):1100–1120.

Basu, S. and Tiwari, R. C. (2010). Breast cancer survival, competing risks and mixture cure model: a bayesian analysis. *Journal of the Royal Statistical Society: Series A*, 173(2):307–329.

Beesley, L. J. and Taylor, J. M. G. (2018). EM algorithms for fitting multistate cure models. *Biostatistics*, 20(3):416–432.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons, Incorporated.

Bryant, J. and Dignam, J. J. (2004). Semiparametric models for cumulative incidence functions. *Biometrics*, 60(1):182–190.

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.

Conlon, A., Taylor, J. M., and Sargent, D. J. (2014). Multi-state models for colon cancer recurrence and death with a cured fraction. *Statistics in Medicine*, 33(10):1750–1766.

Cook, R. J. and Lawless, J. F. (2007). *The Statistical Analysis of Recurrent Events*. Springer Science & Business Media.

Cook, R. J. and Lawless, J. F. (2018). *Multistate Models for the Analysis of Life History Data*. CRC Press.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–202.

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

Craiu, R. V. and Duchesne, T. (2004). Inference based on the EM algorithm for the competing risks model with masked causes of failure. *Biometrika*, 91(3):543–558.

De Stavola, B. L., Daniel, R. M., Ploubidis, G. B., and Micali, N. (2015). Mediation analysis with intermediate confounding: structural equation modeling viewed through the causal inference lens. *American Journal of Epidemiology*, 181(1):64–80.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–22.

Dillekås, H., Rogers, M. S., and Straume, O. (2019). Are 90% of deaths from cancer caused by metastases? *Cancer Medicine*, 8(12):5574–5576.

Dinse, G. E. (1986). Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause-of-death data. *Journal of the American Statistical Association*, 81(394):328–336.

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

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

Fisher, L. D. and Lin, D. Y. (1999). Time-dependent covariates in the Cox proportional-hazards regression model. *Annual Review of Public Health*, 20(1):145–157.

Fix, E. and Neyman, J. (1951). A simple stochastic model of recovery, relapse, death and loss of patients. *Human Biology*, 23(3):205–241.

Flehinger, B. J., Reiser, B., and Yashchin, E. (1998). Survival with competing risks and masked causes of failures. *Biometrika*, 85(1):151–164.

Flehinger, B. J., Reiser, B., and Yashchin, E. (2002). Parametric modeling for survival with competing risks and masked failure causes. *Lifetime Data Analysis*, 8(2):177–203.

Genest, C. and Rivest, L.-P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association*, 88(423):1034–1043.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48.

Gumbel, E. (1960). Distributions des valeurs extrêmes en plusieurs dimensions. *Publications de l'Institut de Statistique de l'Université de Paris*, 9:171–173.

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

Jeong, J.-H. and Fine, J. (2006). Direct parametric inference for the cumulative incidence function. *Journal of the Royal Statistical Society: Series C*, 55(2):187–200.

Joe, H. (2014). *Dependence Modeling with Copulas*. Chapman and Hall/CRC, Boca Raton, FL.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Konigorski, S., Wang, Y., Cigsar, C., and Yilmaz, Y. E. (2018). Estimating and testing direct genetic effects in directed acyclic graphs using estimating equations. *Genetic Epidemiology*, 42(2):174–186.

Lambert, P. C., Dickman, P. W., Weston, C. L., and Thompson, J. R. (2010). Estimating the cure fraction in population-based cancer studies by using finite mixture models. *Journal of the Royal Statistical Society: Series C*, 59(1):35–55.

Larson, M. G. and Dinse, G. E. (1985). A mixture model for the regression analysis of competing risks data. *Journal of the Royal Statistical Society: Series C*, 34(3):201–211.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons.

Lawless, J. F. and Yilmaz, Y. E. (2011). Semiparametric estimation in copula models for bivariate sequential survival times. *Biometrical Journal*, 53(5):779–796.

Lin, D. (1997). Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in Medicine*, 16(8):901–910.

Lin, D., Sun, W., and Ying, Z. (1999). Nonparametric estimation of the gap time distribution for serial events with censored data. *Biometrika*, 86(1):59–70.

Lin, D. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, 81(1):61–71.

Madden, J. M., Leacy, F. P., Zgaga, L., and Bennett, K. (2020). Fitting marginal structural and G-estimation models under complex treatment patterns: Investigating the association between de novo vitamin D supplement use after breast cancer diagnosis and all-cause mortality using linked pharmacy claim and registry data. *American Journal of Epidemiology*, 189(3):224–234.

Martinussen, T. and Scheike, T. H. (2006). *Dynamic Regression Models for Survival Data*. Springer.

Martinussen, T., Vansteelandt, S., Gerster, M., and Hjelmborg, J. v. B. (2011). Estimation of direct effects for survival data by using the Aalen additive hazards model. *Journal of the Royal Statistical Society: Series B*, 73(5):773–788.

McLachlan, G. J. and Krishnan, T. (2007). *The EM Algorithm and Extensions*, volume 382. John Wiley & Sons.

Meira-Machado, L. and Sestelo, M. (2019). Estimation in the progressive illness-death model: A nonexhaustive review. *Biometrical Journal*, 61(2):245–263.

Meng, X.-L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, 86(416):899–909.

Moertel, C. G., Fleming, T. R., Macdonald, J. S., Haller, D. G., Laurie, J. A., Goodman, P. J., Ungerleider, J. S., Emerson, W. A., Tormey, D. C., Glick, J. H., et al. (1990). Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *New England Journal of Medicine*, 322(6):352–358.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer, New York, 2nd edition.

Nicolaie, M. A., van Houwelingen, H. C., and Putter, H. (2010). Vertical modeling: A pattern mixture approach for competing risks modeling. *Statistics in Medicine*, 29(11):1190–1205.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27(2):226–284.

Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):541–554.

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430.

Putter, H., van der Hage, J., de Bock, G. H., Elgalta, R., and van de Velde, C. J. (2006). Estimation and prediction in a multi-state model for breast cancer. *Biometrical Journal*, 48(3):366–380.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512.

Robins, J. M., Hernan, and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Royston, P. and Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15):2175–2197.

VanderWeele, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20(1):18–26.

VanderWeele, T. J. and Tchetgen Tchetgen, E. J. (2017). Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society: Series B*, 79(3):917–938.

Vansteelandt, S. (2009). Estimating direct effects in cohort and case–control studies. *Epidemiology*, 20(6):851–860.

Vansteelandt, S., Goetgeluk, S., Lutz, S., Waldman, I., Lyon, H., Schadt, E. E., Weiss, S. T., and Lange, C. (2009). On the adjustment for covariates in genetic association analysis: a novel, simple principle to infer direct causal effects. *Genetic Epidemiology*, 33(5):394–405.

Vansteelandt, S., Linder, M., Vandenberghe, S., Steen, J., and Madsen, J. (2019). Mediation analysis of time-to-event endpoints accounting for repeatedly measured mediators subject to time-varying confounding. *Statistics in Medicine*, 38(24):4828–4840.

Voelkel, J. G. and Crowley, J. (1984). Nonparametric inference for a class of semi-Markov processes with censored observations. *The Annals of Statistics*, 12(1):142–160.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

Xu, C., Baines, P. D., and Wang, J.-L. (2014). Standard error estimation using the EM algorithm for the joint modeling of survival and longitudinal data. *Biostatistics*, 15(4):731–744.

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

Yilmaz, Y. E., Cigsar, C., and Mariathas, H. (2017). The use of copulas for modelling dependent gap times of recurrent event data. *Invited Paper in Proceedings of the 10th International Conference on Mathematical Methods in Reliability, Grenoble, France*.

Yilmaz, Y. E., Lawless, J. F., Andrulis, I. L., and Bull, S. B. (2013). Insights from mixture cure modeling of molecular markers for prognosis in breast cancer. *Journal of Clinical Oncology*, 31(16):2047–2054.