Semiparametric Regression Models for Disease Natural History and Multiple Events in Cancer Research

by
Chen Hu

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Biostatistics) in The University of Michigan 2012

Doctoral Committee:
Professor Alexander Tsodikov, Chair
Professor Jeremy M.G. Taylor
Associate Professor David Mendez
Associate Professor Douglas E. Schaubel
To my wife, daughter and parents
ACKNOWLEDGEMENTS

This dissertation could have never been completed without the help of many individuals. I want express my deepest gratitude to my advisor, Alex Tsodikov, for his continuous support, encouragement and patience throughout my Ph.D. study. His integrity, scholarship and personality have profound influences on my research and career, and his inspiring guidance has been crucial for the dissertation’s completion. Moreover, I received research assistantship through National Cancer Institute CIS-NET grant (CA157224) under his direction, where I gained substantial knowledge and experiences on statistical modeling in cancer studies. My current and former dissertation committee members, Jack Kalbfleisch, Douglas Schaubel, Jeremy Taylor, David Mendez, have contributed many insightful and constructive suggestions toward the completion of the dissertation, which greatly improve the quality of dissertation. I also feel fortunate enough to learn from them through their lectures and many other occasions. Last but not least, I would like to thank my wife, parents and in-laws for their constant support and love. Without them, I would have not been able to go through the entire journey.
# TABLE OF CONTENTS

DEDICATION ................................................. ii  
ACKNOWLEDGEMENTS ....................................... iii  
LIST OF FIGURES ......................................... vi  
LIST OF TABLES ........................................... vii  
LIST OF APPENDICES ..................................... viii  

CHAPTER

I. Introduction ........................................... 1  

II. Semiparametric Regression Analysis for Time-to-Event Marked Endpoints in Cancer Studies ........................................... 6  
   2.1 Introduction ........................................ 6  
   2.2 Notations, Model and Likelihood ................. 9  
      2.2.1 Notation and Model ......................... 9  
      2.2.2 Likelihood ................................ 10  
   2.3 Semiparametric Regression Model ................. 12  
      2.3.1 Counting Process Notation and Martingale Theory .......... 13  
      2.3.2 Estimating Equation Approach ............... 14  
      2.3.3 Nonparametric Maximum Likelihood Estimation ....... 16  
   2.4 Numerical Examples ................................ 20  
      2.4.1 Simulation Studies ....................... 20  
      2.4.2 Breast Cancer Adjuvant Therapy .......... 21  
   2.5 Discussion ....................................... 27  

III. Joint Modeling Approach for Semicompeting Risks Data with Missing Nonterminal Event Status ............... 29  
   3.1 Introduction ..................................... 29  
   3.2 Model and Data Structure ....................... 33  
      3.2.1 Notation and Model ..................... 33  
      3.2.2 Observed Data Structure and Counting Process Notation .... 37  
   3.3 Likelihood with Missing Nonterminal Event Status ...... 38  
   3.4 Nonparametric Maximum Likelihood Estimation ........ 41  
      3.4.1 Martingale Theory ...................... 41  
      3.4.2 Score Function and Estimating Equation ......... 42  
   3.5 Asymptotic Properties ......................... 44  
   3.6 Numerical Examples ............................ 46
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>An illness-death model with nonterminal event always missing for time-to-event marked endpoints</td>
</tr>
<tr>
<td>2.2</td>
<td>Kaplan-Meier plot for recurrence-free survival by treatment group</td>
</tr>
<tr>
<td>2.3</td>
<td>Cumulative incidence functions for local and distant recurrence by treatment group</td>
</tr>
<tr>
<td>3.1</td>
<td>A three-state illness-death model for semicompeting risks with partially missing nonterminal event</td>
</tr>
<tr>
<td>3.2</td>
<td>Copula plot when $\mu = 2, 1, 0.5$, representing a positive, neutral, and negative association between nonterminal and terminal events</td>
</tr>
<tr>
<td>3.3</td>
<td>Kaplan-Meier plot for time to relapse by treatment group</td>
</tr>
<tr>
<td>3.4</td>
<td>Kaplan-Meier plot for time to death by treatment group</td>
</tr>
<tr>
<td>3.5</td>
<td>Estimated Kendall’s $\tau$. Model I: accounting for possibly missing relapse; Model II: assuming no missing relapse except due to censoring</td>
</tr>
<tr>
<td>4.1</td>
<td>A progressive multistate model for progression and death with partially missing progression</td>
</tr>
<tr>
<td>4.2</td>
<td>Kaplan-Meier plot for recurrence-free survival by treatment group</td>
</tr>
<tr>
<td>4.3</td>
<td>Kaplan-Meier plot for time to death by treatment group</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Simulation results: increasing baseline hazard, 24% local recurrence, 41% distant recurrence</td>
<td>21</td>
</tr>
<tr>
<td>2.2</td>
<td>Simulation results: decreasing baseline hazard, 26% local recurrence, 38% distant recurrence</td>
<td>22</td>
</tr>
<tr>
<td>2.3</td>
<td>Characteristics of breast cancer study subjects</td>
<td>23</td>
</tr>
<tr>
<td>2.4</td>
<td>Breast cancer study analysis results: conventional and proposed approaches</td>
<td>26</td>
</tr>
<tr>
<td>3.1</td>
<td>Simulation results for a model with two covariates; no missing nonterminal events except by right censoring</td>
<td>48</td>
</tr>
<tr>
<td>3.2</td>
<td>Simulation results for a model with two covariates, missing nonterminal events present</td>
<td>49</td>
</tr>
<tr>
<td>3.3</td>
<td>Results of separate Cox PH analyses, and the proposed model accounting for possibly missing relapse</td>
<td>54</td>
</tr>
<tr>
<td>4.1</td>
<td>Event proportions in different simulation scenarios</td>
<td>76</td>
</tr>
<tr>
<td>4.2</td>
<td>Simulation results for gamma frailty variance $\theta = 0.5$</td>
<td>77</td>
</tr>
<tr>
<td>4.3</td>
<td>Simulation results for gamma frailty variance $\theta = 1$</td>
<td>78</td>
</tr>
<tr>
<td>4.4</td>
<td>Simulation results for gamma frailty variance $\theta = 2$</td>
<td>78</td>
</tr>
<tr>
<td>4.5</td>
<td>Summary of stage C colon cancer adjuvant therapy study</td>
<td>80</td>
</tr>
<tr>
<td>4.6</td>
<td>Analysis results of separate Cox marginal models and Cox frailty model</td>
<td>83</td>
</tr>
<tr>
<td>4.7</td>
<td>Analysis results of proposed model with unobserved recurrence</td>
<td>83</td>
</tr>
</tbody>
</table>
LIST OF APPENDICES

Appendix

A. Important Results for Asymptotic Properties ........................................ 91
   A.1 Property of Transform $\int_{0}^{t} \epsilon(u,t;\Omega)dM(u)$ ....................... 91

B. Asymptotic Results for Chapter II ...................................................... 92
   B.1 Asymptotic Results for Estimating Equation Based Estimator ............... 92
      B.1.1 Regularity Conditions ........................................................... 92
      B.1.2 Proof of Theorem II.1 ........................................................... 93
      B.1.3 Additional Notation .............................................................. 94
   B.2 Asymptotic Results for Nonparametric Maximum Likelihood Estimator .... 95
      B.2.1 Regularity Conditions ........................................................... 95
      B.2.2 Proof of Theorem II.2 ........................................................... 95
      B.2.3 Proof of Theorem II.3 ........................................................... 98
      B.2.4 Observed Information Matrix .................................................. 99

C. Asymptotic Results for Chapter III ..................................................... 101
   C.1 Proof of Theorem III.1 ............................................................... 101
   C.2 Proof of Theorem III.2 ............................................................... 104
   C.3 Information Matrix ................................................................. 105

D. Asymptotic Results for Chapter IV ...................................................... 107
   D.1 Proof of Theorem IV.1 ............................................................... 107
   D.2 Proof of Theorem IV.2 ............................................................... 110
   D.3 Information Matrix ................................................................. 111
CHAPTER I

Introduction

This dissertation is devoted to semiparametric joint models of disease natural history and its relationship with observed multiple events. These research work are primarily motivated by some interesting problems emerging from cancer studies, in which a better understanding of disease natural history is needed to better understand clinical outcomes and disease prognosis and design optimal treatments. However, this has never been an easy task, as the disease natural history is never directly observed due to its latency. The corresponding observation schemes thus play an important role in answering the questions of interest. Specifically, development of flexible statistical models for disease natural history jointly with observed information becomes a pivotal issue throughout this dissertation.

Event history analysis and multistate models ([3, 6]) deal with multiple events collected over time. Existing research in this area primarily focuses on the relationship between observed multiple events. In this dissertation, we extend this framework and invoke a mechanistic if simplified model of the latent disease natural history explaining the observed correlated phenotype. As we shall see in this dissertation, one way to design such models is to consider the observed events as part of the same disease progression process leading to natural assumptions on the association between the
observed endpoints. Such models have the potential to better utilize the observed information and achieve greater efficiency compared with traditional methods that typically treat part of the observed phenotype as nuisance. A partially observed latent disease progression process means that valid statistical inference requires appropriate consideration of missing data resulting from the discrepancy between the complete and observed data.

One immediate advantage of such modeling strategy is the formulation of the dependence between the observed surrogate of the disease natural history and the observed survival events. Technically, when modeling the disease process jointly with observed survival we use the concept of modulation that described how occurrence of a latent event affects the risk of occurrence of future events in a time-dependent fashion. This approach has been used in recurrent data analysis ([17] Section 5.3). The resultant dependence measures have clinically meaningful interpretation and accommodates both positive and negative associations. In this aspect conditional specification based on modulated point processes is more general than frailty ([31]) or copula ([49]) models that typically only handle positive (frailty) or weak negative (copula models) correlations.

Furthermore, the proposed mechanistic formulation leads to information sharing between the observed and the latent part of the disease process, and provides a way to avoid the multiple nonidentifiability issues encountered in previous work that treats one or the other aspect of the joint picture as disconnected nuisance, such as [37, 65, 26] in current status data ([73]) setting and [52] in semicompeting risks data setting. The joint modeling approaches also provides a predictive capability for all possible outcomes from the underlying disease progression process.

The issue of the latent disease process governing the observed responses is a chal-
lenging territory in semiparametric regression analysis where the cumulative baseline hazard function is fully unspecified and is affected by a latent stochastic process. We show that it is possible to adopt estimation procedures used in transformation models, e.g., [12, 61, 71, 13], and establish asymptotic properties for the corresponding estimators based on martingale ([3]) and empirical process ([62, 40]) machinery.

The rest of dissertation is organized as follows:

• In Chapter 2, I consider the situation when the disease natural history process is only observed at a fixed, random point in time (survival time), leading to a current status data ([73]) surrogate (a mark) ([35]). The observed survival time plays a role of informative censoring for the latent process. The observed marked survival data can also be viewed within a dependent competing risks framework ([36]). Examples of data types handled by this approach include time to recurrence and recurrence site, or time to cancer diagnosis and stage at diagnosis. Little has been done to provide statistical inference on the relationship between the observed marked endpoint and the latent disease natural history leading to it. I present a semiparametric regression model to assess the covariate effects on the observed marked endpoint explained by a latent disease natural history. Constructed through a nested series of Cox relative risk models with time-dependent covariates, the proposed semiparametric regression model can be represented as a transformation model in terms of outcome-specific hazards, induced by a complex non-proportional process-based frailty. Two estimation procedures for such transformation models are proposed: 1) an estimating equation based approach; 2) a nonparametric maximum likelihood estimation (NPMLE) approach. Large-sample and finite-sample properties of the proposed estimators are established. The methodology is illustrated by Monte Carlo sim-
ulation studies, and an application to a randomized clinical trial of adjuvant therapy for breast cancer.

- Chapter 3 deals with the case when the disease natural history of interest is observable in principle. However, the status of the dependent censoring by the terminal event is missing in a fraction of patients. The disease natural history represents an event of disease progression, a nonterminal event. The nonterminal event and terminal events together represent the so-called semicompeting risks data ([21]), where a subject may experience sequential nonterminal and terminal events, and the terminal event may censor the nonterminal event but not vice versa. When a proportion of subjects’ nonterminal events is missing, the disease natural history may be right- or left-censored by the terminal event in an informative way. Which of the two options manifests itself may be unknown if there is no indication whether the non-terminal event is in fact missing. Examples include cancer local recurrence (non-terminal) and death (terminal). While death may be due to metastases rather than local recurrence, when death is observed it may be unknown whether recurrence has not happened yet, or did happen but was missed in the data registration process. The observed data thus become a mixture of true semicompeting risks data and partially observed terminal event only data. An illness-death multistate model with proportional hazards assumptions is proposed to study the relationship between nonterminal and terminal events, and provide covariate-specific global and local association measures. Maximum likelihood estimation based on semiparametric regression analysis is used for statistical inference, and asymptotic properties of proposed estimators are studied using empirical process and martingale arguments. I illustrate the proposed method with simulation studies and data analysis of a
follicular cell lymphoma study.

- In Chapter 4, I investigate the scenario when disease progression always precedes death (the terminal event) in disease settings like advanced or adjuvant cancer studies, such that the progression-related events (e.g., progression-free or recurrence-free survival) and cancer death are sequentially observed representing recurrent event on the complete data level. The relationship between covariate (e.g., therapeutic intervention), progression, and death is often of interest, as it may provide a key to optimal treatment decisions. The evaluation of this relationship is often complicated by the latency of disease progression leading to undetected or missing progression-related events. I consider a progressive multistate model with a frailty and a modulation modeling the association between progression and death, and propose a semiparametric regression model for the joint distribution. An Expectation-Maximization (EM) approach is used to derive the maximum likelihood estimators of covariate effects on both endpoints, the probability of missing progression event, as well as the parameters involved in the association. The asymptotic properties of the estimators are studied. We illustrate the proposed method with Monte Carlo simulation and data analysis of a clinical trial of colorectal cancer adjuvant therapy.

- In Chapter 5, I present a brief summary of the strengths and limitations of the proposed semiparametric regression models, and lay out some future research areas that can be further explored based on the current framework for disease natural history and event history analysis.
CHAPTER II

Semiparametric Regression Analysis for Time-to-Event Marked Endpoints in Cancer Studies

2.1 Introduction

In cancer studies, understanding the disease natural history and its manifestation through the observed survival phenotype is essential for the design of optimal treatment strategies. Cancer endpoints are typically measured at an observed event and represent a time to event and a cross-sectional surrogate of the disease natural history measured at the event. It is crucial to elucidate the relationship between the marked survival phenotype and the latent disease natural history process and understand what can be learned from the observed data. Event history analysis and multistate models ([6]) that deal with events collected over time, received considerable attention in the past several decades. One common type of observed endpoint in the event history analysis is a time-to-event jointly observed with another outcome variables such as cancer stage. Examples include: 1) time to cancer recurrence and recurrence site (local or distant); 2) time to cancer diagnosis (incidence) and cancer stage and/or grade at diagnosis.

A number of generic methods have been proposed for marked survival endpoints. Continuous marks were considered in ([34]), and categorical marks were approached within the dependent competing risks framework ([36] chapter 8 and references
therein). Due to the non-identifiability issue of competing risks data ([60]), models based on observable quantities such as “crude” outcome-specific hazards ([36], chapter 8) or models based on latent failure times ([20, 14]) do not provide mechanistic information on the dependence between outcome variables. In this paper we recognize that a disease history mark measured at the survival event is often a categorical variable representing a snapshot of a compartmental disease progression process. The model for the outcome variables is then mechanistic in nature and depends on the latent association between the disease progression process and the survival time affected by it. Therapeutic interventions and prognostic factors have complex causality affecting the outcomes directly and through the latent disease process. Understanding this relationship may help design optimal therapy regimen and correctly assess and explain the diversity of observed treatment effects.

In past several decades, statistical mechanistic models have been used to extract information on the latent disease natural history from observed data. Much work has been done through progressive multistate models to determine the effects of screening interventions (e.g., [68, 2, 56]). Convolutions involved in the partially unobserved cancer progression through stages viewed as a renewal process present technical difficulties and require de-convolution and smoothing or a parametric analysis. Targeting a flexible semiparametric modeling approach, the relationship between disease progression (stage) and the survival time was treated within the current-status data framework [37, 65, 66, 26, 27]. This approach views the time to event as an informative censoring for the time-to-metastasis endpoint, the latter representing current status data on whether the metastatic transition has happened or not at the point of evaluation (the survival event). For reasons of nonidentifiability ([73]) two extreme scenarios were studied: 1) tumors are detected immediately when metastases occur;
2) tumor detection is not affected by the presence of metastases (noninformative censoring).

Identifiability issues can be resolved by making additional assumptions. In fact, both the time to diagnosis (a survival time playing the role of censoring event for the time to metastases) and the time to metastases are part of the same disease progression process, and it would be natural to build a joint model for it avoiding a consideration of censoring as a nuisance. This strategy adopted in this paper resolves the non-identifiability in a natural way and converts a current status formulation to a right-censored joint transformation model with dynamic frailty. Besides, this approach is generalizable to non-binary marks.

Two estimation procedures for the transformation model are pursued: 1) an estimating equation approach for nonparametric component based on martingale properties. The resulting estimator is shown to be a Breslow-type estimator ([10]). This approach is similar to the method for linear transformation models proposed by [12]. While being not fully efficient, the method has some computational advantage; 2) a fully efficient NPMLE approach that is a little more complex computationally. The resulting estimator is shown to follow a weighted Breslow-type approach [13], where weights depend on martingale residuals. Asymptotic properties are established using marked point processes ([35]), empirical processes and martingale theory ([3]). We assess the finite-sample properties of proposed estimators through extensive Monte Carlo simulation studies, and apply the proposed methods to a randomized clinical trial of adjuvant therapy for breast cancer.
2.2 Notations, Model and Likelihood

2.2.1 Notation and Model

Consider a time-to-event $T^*$ and a mark $S^*$ representing a state of the latent process measured at $T^*$. In this paper we assume the mark to be binary, and we model the joint distribution of random variables $(T^*, S^*)$ through latent variable $U$ that represents a surrogate of disease natural history. Such modeling strategy provides statistical inference on both disease natural history and dependence measure between outcome variables. For example, in the setting of cancer treatment studies where times to local or distant recurrence are of interest, $U$ may be viewed as time to detectable metastasis. We will identify $U$ with a time to a disease progression event in the rest of the paper.

Formally, let $T^*$ and $U$ be the underlying survival time and time to a progression event, respectively, and let $S^*(t) = I(U < t)$ be the underlying counting process for the occurrence of disease progression transition. Note that the binary mark observed at the failure event is simply the value of $S^*(T^*)$, hence the abuse of notation when both the mark and the latent process are denoted by the same $S^*$. We further denote by $Z(t)$ the covariates of interest, a vector of dimensions $p$. We specify the joint distribution of $(T^*, S^*)$ based on the following conditional and marginal hazard functions:

\begin{align}
\lambda_U(u|Z(u)) &= \lim_{\Delta \to 0} \frac{P(U \in [u, u + \Delta]|U \geq u, T^* \geq u, Z(u))/\Delta = h_0(u)\eta(Z(u)),}

\lambda_T(t|Z(t), S^*(t)) &= \lim_{\Delta \to 0} \frac{P(T^* \in [t, t + \Delta]|T^* \geq t, Z(t), S^*(t))/\Delta = h_0^*(t)\theta(Z(t))\mu^{S^*(t)},}
\end{align}

where $\eta(Z(t)) = e^{\beta_\eta Z(t)}, \theta(Z) = e^{\beta_\theta Z(t)}, \mu = e^{\beta_\mu}$. Denote by $\beta = (\beta_\eta, \beta_\theta, \beta_\mu)$ the
combined vector of regression coefficients. Covariates $Z$ will be time-independent in the rest of the paper for brevity. An extension to external time-dependent setting ([36] section 6.4) is straightforward.

The model presented above is based on a nested hierarchical Cox proportional hazard (PH) model, one for $U$ and one for $T^*$ with $U$ being a time-dependent covariate ([36] section 6.4). Alternatively, it can be viewed as an illness-death model as described in Figure 2.1 ([36] section 8.3, [6]), where $\lambda_U$ corresponds to the state transition hazard between state 0 and 1 ($\lambda_{01}$), and $\lambda_T$ corresponds to the state transition hazard between state 0 and 2 or 3, that may depend on whether state 1 occurs ($\lambda_{13}$) or not ($\lambda_{02}$).

It is noted that $\mu$ may be interpreted as a dependence measure between the latent disease natural history $U$ and the observed time-to-event process $T^*$. $\mu = 1$ represents no association between $U$ and $T^*$, while $\mu > 1$ or $\mu < 1$ represents a positive or negative association, such that the occurrence of $U$ accelerates or decelerates the occurrence of $T^*$.

2.2.2 Likelihood

Since both $U$ and $T^*$ are determined by the same disease progression process, it is natural to assume a common baseline hazard $h(\cdot) \equiv h_0^*(\cdot) = h_0(\cdot)$ summarizing the dynamic pattern of the disease. With this PH assumption, we derive the following probabilistic quantities in a closed form:

$$G(t|Z) = \Pr(T^* \geq t|Z) = \frac{\theta \bar{\mu}}{\theta \bar{\mu} + \eta} e^{-(\theta + \eta)H(t)} + \frac{\eta}{\theta \bar{\mu} + \eta} e^{-\theta \mu H(t)},$$

$$f(t, S^* = 0|Z) = h(t) \theta e^{-(\theta + \eta)H(t)},$$

$$f(t, S^* = 1|Z) = h(t) \frac{\theta \eta \mu}{\theta \bar{\mu} + \eta} (e^{-\theta \mu H(t)} - e^{-(\theta + \eta)H(t)}),$$
Figure 2.1: An illness-death model with nonterminal event always missing for time-to-event marked endpoints
and the corresponding outcome-specific hazards as

$$\lambda_s(t|Z) = \frac{f(t, S^* = s)}{G(t)} = h(t)\Theta_s(t; \beta, H|Z) = h(t)\Theta_s(t; \beta, H),$$

where

$$\Theta_0(t; \beta, H) = \frac{\theta e^{-(\theta + \eta)H(t)}}{\bar{\theta} \mu e^{-(\theta + \eta)H(t)} + \frac{\eta}{\bar{\theta} + \eta} e^{-\theta \mu H(t)}}, \quad \Theta_1(t; \beta, H) = \frac{\theta \mu e^{-(\theta + \eta)H(t)} - e^{-(\theta + \eta)H(t)}}{\bar{\theta} \mu e^{-(\theta + \eta)H(t)} + \frac{\eta}{\bar{\theta} + \eta} e^{-\theta \mu H(t)}},$$

with $H(t) = \int_0^t h(x)dx$ and $\bar{\mu} = 1 - \mu$.

Now we introduce the observed data structure. Without loss of generality, let $V^*$ be the censoring time which is independent of $(T^*, S^*)$ given $Z$; $\tau$ be the maximum follow-up time in the study such that $\tau = \inf\{t : \text{pr}(T^* > t) = 0\}$. Suppose $(T_i^*, S_i^*, V_i^*, Z_i), i = 1, \cdots, n$ are $n$ independent and identically distributed replicates of $(T^*, S^*, V^*, Z)$, such that the observed time-to-event marked data for subject $i, i = 1, \cdots, n$ consist of $\{T_i, \delta_i, S_i, Z_i\}$, where $T_i = T_i^* \wedge V_i$ with $V_i = V_i^* \wedge \tau$, $\delta_i = I(T_i = T_i^*)$, and $I(\cdot)$ is the indicator function, $S_i = \delta_i S_i^*$, and $S_i$ is unobserved and undefined if $\delta_i = 0$.

Therefore, the full likelihood of observed time-to-event marked data becomes

$$L = \prod_{i=1}^n \lambda_s(t_i|Z_i)^{\delta_i}G(t_i|Z_i)$$

### 2.3 Semiparametric Regression Model

Our model is semiparametric in the sense that the baseline cumulative hazard $H(\cdot)$ is unspecified and is treated as a nondecreasing step function with jumps at the times where events are observed. The full parameter set is $\Omega = (\beta, H(\cdot))$, where $\beta = (\beta_\eta, \beta_\theta, \beta_\mu)$ is finite-dimensional and $H(\cdot)$ is infinite-dimensional. This nonparametric maximum likelihood estimation (NPMLE) technique has been adopted in semiparametric transformation models [12, 61, 71, 13], to which our model bears
strong resemblance. We will understand the function $H$ as a set of jumps $\{dH\}$ of $H$ at the observed event times.

2.3.1 Counting Process Notation and Martingale Theory

In counting process notation, for subject $i$, let $Y_i(t) = I(T_i \geq t)$ be the at-risk process, $N_{si}(t) = \delta_i I(S_i = s)I(T_i \leq t)$ be the outcome-specific counting process with $N_i(t) = \sum_{s=0}^{1} N_{si}(t)$, such that $N_{si}(\cdot)$ records the time and outcome of subject $i$’s event, and $N_i(t)$ records the time of subject $i$’s event regardless of the outcome. It is noted that $S$ and $N_{si}(\cdot), s = 0, 1$ may be considered in the context of a marked point process ([35]).

We define the filtration as $\mathcal{F}_{t-} = \sigma\{N_{0i}(x), N_{1i}(x), Y_i(x), Z_i : x \in [0, t), i = 1, \cdots, n\}$, and consider the continuous (orthogonal) case where no two counting processes can jump simultaneously. Then, under independent censoring, we have

$$
P[dN_{si}(t) = 1|\mathcal{F}_{t-}] = Y_i(t)\Theta_{si}(t; \beta, H)dH_t,
$$

$$
G(t|Z_i) = \exp \left\{ - \int_0^t \sum_{s=0}^{1} Y_i(x)\Theta_{si}(x; \beta, H)dH_x \right\},
$$

where we assume $\Theta_s(t; \beta, H)$ is predictable. Construct the following orthogonal martingales based on the observed marked counting processes with respect to $\mathcal{F}_{t-}$,

$$(2.4) \quad dM_{si}(t) = dN_{si}(t) - Y_i(t)\Theta_{si}(t; \beta, H)dH_t, \quad s = 0, 1 \quad i = 1, \cdots, n$$

The log-likelihood takes the form

$$(2.5) \quad \ell = \sum_{i=1}^{n} \sum_{s=0}^{1} \int_0^x \{\log dH_x + \log \Theta_{si}(x; \beta, H)\} dN_{si}(x) - Y_i(x)\Theta_{si}(x; \beta, H)dH_x \}.$$
2.3.2 Estimating Equation Approach

Estimating Equations and Breslow-Type Estimator

By aggregating the marked martingales (2.4) and marked counting processes, i.e.,
\[ M(t) = \sum_{i=1}^{n} \sum_{s=0}^{1} M_{si}(t), \quad N(t) = \sum_{i=1}^{n} \sum_{s=0}^{1} N_{si}(t), \]
we have
\[ dM(t) = dN(t) - nS_0(t; \beta, H) dH_t, \tag{2.6} \]

where \( S_0(t; \beta, H) = n^{-1} \sum_{i=1}^{n} \sum_{s=0}^{1} Y_i(t) \Theta_{si}(t; \beta, H). \)

Based on martingale properties, it is natural to specify the following estimating equations
\[ dN(t) - nS_0(t; \beta, H) dH_t = 0, \]
or equivalently, a Breslow-type estimator ([10, 41]) for \( H_t \) as
\[ \hat{H}_t(\beta) = \int_0^t \frac{n^{-1} dN(x)}{S_0(x; \beta, \hat{H})}, \tag{2.7} \]
for fixed \( \beta \). This construction is similar to that used by [12] for linear transformation models. For any fixed \( \beta \), the Breslow-type estimator \( \hat{H}_t(\beta) \) can be computed recursively from (2.7), a computational advantage of the method.

By replacing \( dH_t \) and \( H_t \) with \( d\hat{H}_t(\beta) \) and \( \hat{H}_t(\beta) \), the solution of (2.6), in the log-likelihood (2.5), and dropping terms not depending on \( \beta \), we can obtain the quasi-profile likelihood
\[ \ell_{pr}(\beta) = \ell(\beta, \{d\hat{H}(\beta)\}) = \sum_{i=1}^{n} \sum_{s=0}^{1} \int_0^t \left[ \log \Theta_{si}(x; \beta, \hat{H}) - \log S_0(x; \beta, \hat{H}) \right] dN_{si}(x), \tag{2.8} \]

Recall that \( \Theta_s(t; \beta, \hat{H}) \) is a function of \( \beta \) as \( \hat{H} \) is also a function of \( \beta \), and denote its corresponding derivative with respect to \( \beta \) as \( \dot{\Theta}_s(t; \beta, \hat{H}) = \frac{d}{d\beta} \Theta_s(t; \beta, \hat{H}). \) Then,
the quasi-profile score function \( U_{pr}(\beta) \), the derivative of the quasi-profile likelihood \( \ell_{pr} \) with respect to \( \beta \), is

\[
U_{pr}(\beta) = \frac{d}{d\beta} \ell(\beta, \{d\hat{H}(\beta)\}) = \sum_{i=1}^{n} \sum_{s=0}^{1} \int_{0}^{\tau} \left[ \frac{\dot{\Theta}_{si}(x; \beta, \hat{H})}{\Theta_{si}(x; \beta, \hat{H})} - \frac{\dot{S}_{1}(x; \beta, \hat{H})}{S_{0}(x; \beta, \hat{H})} \right] dN_{si}(x),
\]

where \( \dot{S}_{1}(t; \beta, \hat{H}) = n^{-1} \sum_{i=1}^{n} \sum_{s=0}^{1} Y_{i}(t) \dot{\Theta}_{si}(t; \beta, \hat{H}). \)

The estimators \( \hat{\beta} \) may be obtained by maximizing of the quasi-profile likelihood (2.8) or solving the quasi-profile score function (2.9).

**Asymptotic Properties**

Let \( \beta^{0} \) and \( H^{0}_{t} \) be the true value of \( \beta \) and \( H_{t} \), and assume \( \Theta_{s}(\cdot), s = 0, 1 \) is positive; \( \dot{\Theta}_{s}(\cdot) / \Theta(\cdot) \) is continuous. For any \( t \in (0, \tau] \), let \( b \odot b' = bb' \) for any vector \( b \).

The complete regularity conditions are listed in Appendix, in which we essentially follow [22] (p289-p290) to ensure martingale central limit theorem. It is not difficult to verify the conditions of Lemma 8.3.1 in [22] as the profile likelihood 2.8 behaves like a regular parametric likelihood in survival analysis, and \( \hat{\beta} \) is the unique maximizer of \( \ell_{pr}(\beta) \), such that the consistency of \( \hat{\beta} \) can be established.

**Theorem II.1.** Under suitable regularity conditions, we have

\[
n^{1/2}(\hat{\beta} - \beta^{0}) \xrightarrow{D} N\left(0, \Sigma^{-1}_{\beta}(\Sigma^{-1}_{\beta})'\right)
\]

as \( n \to \infty \). Moreover, \( \Sigma^{*} \) and \( \Sigma_{*} \) can be consistently estimated by

\[
\hat{\Sigma}^{*} = n^{-1} \sum_{i=1}^{n} \sum_{s=0}^{1} \int_{0}^{\tau} \left[ \frac{\dot{\Theta}_{si}(x; \hat{\beta}, \hat{H})}{\Theta_{si}(x; \hat{\beta}, \hat{H})} - Q(x; \hat{\beta}, \hat{H}) \right] \otimes \Omega_{si}(x; \hat{\beta}, \hat{H}) d\hat{H}(x),
\]

\[
\hat{\Sigma}_{*} = n^{-1} \int_{0}^{\tau} \left[ \frac{\dot{S}_{1}(x; \hat{\beta}, \hat{H}) S_{0}(x; \hat{\beta}, \hat{H}) - \dot{S}_{1}^{\otimes 2}(x; \hat{\beta}, \hat{H})}{S^{2}(x; \hat{\beta}, \hat{H})} \right] d\hat{H}(x),
\]

where \( Q(\cdot), \tilde{S}_{2}(\cdot), \) and \( \dot{S}_{1}(\cdot) \) are defined in Appendix.
The proof is outlined in Appendix, where we follow the line of [8] for linear transformation models. The proof involves the following steps: (i) the uniform consistency of $\hat{H}_t(\beta_0)$; (ii) the weak convergence of $\hat{H}_t(\beta_0)$; (iii) the limiting value of Jacobian operator $\dot{\hat{H}}_t(\beta) = \frac{d}{d\beta} \hat{H}_t(\beta)$; (iv) the weak convergence of $n^{-1/2}U_{pr}(\beta^0)$, i.e., $n^{-1/2}U_{pr}(\beta^0) \xrightarrow{D} N(0, \Sigma^*)$; (v) $n^{-1}I_{pr}(\beta)\big|_{\beta=\beta_0} \xrightarrow{P} \Sigma_*$, where $I_{pr}(\beta) = I_{pr}(\tau, \beta) = -\frac{d}{d\beta} U_{pr}(\beta)$. The primary tools include Martingale Central Limit Theorem (MCLT), Weak Law of Large Number (WLLN), Lenglart’s inequality ([3]), and Slutsky’s theorem.

2.3.3 Nonparametric Maximum Likelihood Estimation

Score Functions and Weighted Breslow-Type Estimator

The above estimating equation approach based on martingale properties provides estimators that are consistent, computationally fast, yet not fully efficient.

The score $U_{H_t}$ represents a local directional functional derivative with respect to $H$ defined as follows. For any functional $f$

$$\frac{\partial f(H(\cdot))}{\partial dH(t)} = \left. \frac{df(H(\cdot) + a \times 1(\cdot - t))}{da} \right|_{a=0},$$

where $1(x) = 1$ when $x \geq 0$ and 0 otherwise (an impulse function). Note that this definition corresponds to taking a derivative with respect to a jump of $H$ at time $t$ if $H$ is a step-function.

Denote the derivatives of $\Theta_{si}(t; \beta, H)$ with respect to $\{dH\}$ and $\beta$ as: $\Theta_{si,H}(t; \beta, H) = \partial \Theta_{si}(t; \beta, H)/\partial dH$, $\Theta_{si,\beta}(t; \beta, H) = \partial \Theta_{si}(t; \beta, H)/\partial \beta$.

Based on likelihood (2.5), we can obtain score function for $dH(t_*) = dH_*$ as

$$U_{H_t} = \sum_{i=1}^{n} \sum_{s=0}^{\tau} \int_0^{\tau} \left\{ dN_{si}(x) - Y_i(x)\omega_{si}(x; \beta, H)\Theta_{si}(x_-; \beta, H)dH_x \right\},$$

(2.11)
where

\[
\omega_{si}(t^*_i; \beta, H) = 1 - \frac{\int_{t^*_i}^{\tau} \psi_{si}(x; \beta, H) dM_{si}(x)}{\Theta_{si}(t^*_i; \beta, H)}; \quad \psi_{si}(t; \beta, H) = \frac{\Theta_{si, H}(t; \beta, H)}{\Theta_{si}(t; \beta, H)}.
\]

(2.12)

The score function for \( \beta \) is

\[
U_{\beta} = \sum_{i=1}^{n} \sum_{s=0}^{1} \int_{0}^{\tau} \left\{ \frac{\Theta_{si, \beta}(x; \beta, H)}{\Theta_{si}(x; \beta, H)} dN_{si}(x) - Y_i(t) \Theta_{si, \beta}(t; \beta, H) dH_x \right\}
\]

(2.13)

Solutions to the above estimating equations (2.11) and (2.13), when set to zero, give nonparametric maximum likelihood estimators (NPMLE) \( \{d\tilde{H}\} \) and \( \tilde{\beta} \).

Note that the proposed method is semiparametric and thus the dimension of the parameter \( \Omega \) is of the same order as the sample size. Using common scientific computing software, it is now feasible to solve the large scale optimization problem directly. However a structured solution is more stable and computationally attractive. An EM algorithm or an iterative reweighting algorithm of [13] can be used. Based on \( U_{H} \), we can write the NPMLE of \( H \) as

\[
\tilde{H}(t; \beta) = \int_{0}^{t} \frac{n^{-1} dN_{i}(t)}{S_{0}^{w}(x; \beta, H)},
\]

(2.14)

where \( S_{0}^{w}(t; \beta, H) = n^{-1} \sum_{i=1}^{n} \sum_{s=0}^{1} Y_i(t) \omega_{si}(t; \beta, H) \Theta_{si}(t; \beta, H) \). By replacing \( dH(t) \) and \( H(t) \) by \( d\tilde{H}(t; \beta) \) and \( \tilde{H}(t; \beta) \) in \( U_{\beta} \) (2.13), we have the profile score function as

\[
\tilde{U}_{pr}(\beta) = \sum_{i=1}^{n} \sum_{s=0}^{1} \int_{0}^{\tau} \left\{ \frac{\Theta_{si, \beta}(x; \beta, H)}{\Theta_{si}(x; \beta, H)} - \frac{S_{1}^{w}(x; \beta, H)}{S_{0}^{w}(x; \beta, H)} \right\} dN_{si}(x)
\]

(2.15)

where \( S_{1}^{w}(t; \beta, H) = n^{-1} \sum_{i=1}^{n} \sum_{s=0}^{1} Y_i(t) \Theta_{si, \beta}(t; \beta, H) \).

The NPMLE thus can be obtained through solving the profile score equation \( (2.15)=0 \) based on the following iterative reweighting algorithm. Starting with initial weights \( \omega^{(0)} = 1 \) and initial values \( \beta^{(0)} = 0, dH^{(0)}(t) = N_{i}(t)/Y_{i}(t) \), the Nelson-Aalen estimator, for \( k = 0, 1, \cdots \),

1. fix weights \( \omega^{(k)} \) and solve the estimating equation \( (2.15)=0 \) for \( \beta \). Obtain the solution \( \beta^{(k+1)} \) and \( dH^{(k+1)}(t) \) from (2.14);
2. update weights $\omega^{(k+1)}$ by (2.12) with $\beta^{(k+1)}$ and $H^{(k+1)}$. Return to the previous step and continue until convergence.

Remark 1. It is intriguing to note the relationship between inefficient $\hat{H}$ and fully efficient $\tilde{H}$. $\tilde{H}$ can viewed as a weighted Breslow-type estimator $\hat{H}$, with weights $\omega_s(\cdot)$. Based on martingale properties, weights $\omega_s(\cdot)$ have unit expectations. This fact explains why both the Breslow estimators $\hat{H}(\cdot)$ and $\tilde{H}(\cdot)$ are consistent. The weights $\omega_s(t; \beta, H)$ depend on martingale residuals $\int_{t+} \psi_s(x; \beta, H) dM_s(x)$ evaluated over the future of $t$. Utilization of only the past information makes the estimating equations based $\hat{H}$ inefficient. The weighted Breslow estimator, on the contrary, utilizes the full information on the subject, yet is more complex computationally as values of $H$ in the future of $t$ distort the recursive structure of the score equation.

**Asymptotic Properties**

Based on the score functions (2.11) and (2.13), we can write the score functions for $\Omega = (\beta, \{dH\})$ as martingales (at the true $\Omega^*$)

\begin{equation}
U_\beta = \sum_{i=1}^{n} \sum_{s=0}^{1} \int_{0}^{T} \frac{\Theta_{si,\beta}(x; \beta, H)}{\Theta_{si}(x; \beta, H)} dM_{si}(x),
\end{equation}

and

\begin{equation}
U_{H_t} = \sum_{i=1}^{n} \sum_{s=0}^{1} \int_{0}^{T} \left[ dM_{si}(x) + \int_{x+}^{T} \psi_{si}(u; \beta, H) dM_{si}(u) dH_x \right].
\end{equation}

Exchanging the integrals, we have:

\begin{equation}
U_{H_t} = \sum_{i=1}^{n} \sum_{s=0}^{1} \left\{ \int_{0}^{T} \left[ 1 + \int_{0}^{u} \psi_{si}(u; \beta, H) dH_x dM_{si}(u) \right] + \int_{t+}^{T} \int_{0}^{T} \psi_{si}(u; \beta, H) dH_x dM_{si}(u) \right\}
\end{equation}

\begin{equation}
= \sum_{i=1}^{n} \sum_{s=0}^{1} \int_{0}^{T} \varepsilon_{si}(u, t; \beta, H) dM_{si}(u),
\end{equation}
where \( \varepsilon_{si}(u, t; \beta, H) = I(u \leq t) + \int_0^{u\wedge t} \psi_{si}(u; \beta, H) dH_x \). As we show in Appendix, martingale transform \( \int_0^\tau \varepsilon_{si}(u, t; \beta, H) dM_{si}(u) \) is a martingale as \( \varepsilon_{si}(u, t; \beta, H) \) does not depend on \( t \) for \( u \leq t \). Therefore, the score function \( U_\beta \) and \( U_{H_t} \) are both martingales at the true model.

In the following we present the consistency and weak convergence results for the proposed NPMLE \( \tilde{\Omega} = (\tilde{\beta}, \tilde{H}_t) \). The detailed proofs may be found in Appendix. An empirical process argument is used for consistency and a martingale argument is used for the weak convergence.

**Theorem II.2.** Assuming regularity conditions hold, with probability one, \( \tilde{\beta} \) converges to \( \beta^0 \), \( \tilde{H}_t \) converges to \( H^0_t \) uniformly in the interval \([0, \tau]\).

Consider a linear functional

\[
(2.18) \quad n^{1/2} \left\{ \int_0^\tau a^T (\tilde{\beta} - \beta^0) + b(t)^T (\tilde{H}_t - H^0_t) \right\},
\]

where \( a \) is real vector, \( b(t) \) is a function with bounded total variation in \([0, \tau]\), and let \( B \) be the vector consisting of the values of \( b(t) \) evaluated at the observed failure times corresponding to the set \( \{dH\} \), and \( \mathcal{E}^T = (a^T, B^T) \).

**Theorem II.3.** Assuming regularity conditions hold, \( n^{1/2} \{\tilde{\beta} - \beta^0, \tilde{H}_t - H^0_t\} \) converges weakly to a zero-mean Gaussian process. In addition, \( n\mathcal{E}^T (\tilde{\mathcal{I}}_n)^{-1} \mathcal{E} \) converges in probability to the asymptotic variance-covariance function of the linear functional (2.18), where \( \tilde{\mathcal{I}}_n \) is the negative Hessian matrix of the observed log-likelihood function with respect to \( \tilde{\Omega} \).

For a differentiable functional \( F(\Omega) \) of \( \Omega \), based on the functional delta method ([3] Section II.8), \( n^{1/2} \{F(\tilde{\Omega}) - F(\Omega)\} \) converges weakly to a zero-mean Gaussian process with variance-covariance function \( \hat{F}(\tilde{\Omega})^T (\tilde{\mathcal{I}}_0)^{-1} \hat{F}(\tilde{\Omega}) \), where \( \hat{F}(\Omega) \) is the gradient of
$F(\Omega)$ with respect to $\Omega$, $\tilde{I}^0$ can be consistently estimated by $n^{-1}\tilde{Z}_n$, with the explicit expression of $\tilde{Z}_n$ is derived in Appendix.

2.4 Numerical Examples

2.4.1 Simulation Studies

To assess the finite-sample properties of the parameter estimates obtained by the proposed methodology, we perform a series of Monte Carlo simulation studies under different scenarios. We consider two covariates $Z_1$ and $Z_2$, where $Z_1 \sim \text{Unif}(0, 1)$, and $Z_2 \sim \text{Binom}(0.5)$. To illustrate the flexibility of semiparametric regression, we assume the common baseline hazard function following Weibull distribution, $h(t) = 0.02t$ (increasing), and $h(t) = 0.05(t/10)^{-1/2}$ (decreasing). Using uniform random variables $U(0, 20)$ and $U(0, 10)$ for independent censoring yield approximately 35% censoring proportions for both scenarios. For each simulation scenario, 1,000 data sets are generated. Sample sizes $n = 250$ and 500 are considered. The results are summarized in Table 2.1 and 2.2. Based on the simulation results, both proposed estimation and inference procedures based on unweighted and weighted Breslow-type estimators appear to have satisfactory performance for the sample sizes considered, with diminishing bias for larger sample sizes and good coverage probability at 95% nominal level. When sample size is small, the weighted Breslow-type estimator based method has slightly better performance than the inefficient estimation approach, with slightly better coverage probability and agreement between empirical standard errors and asymptotic standard deviations. As sample size increases, we find both methods to be quite comparable, and the efficiency loss from inefficient estimators is almost negligible. In summary, based on our simulation studies, we find the estimating equation based approach (M1) attractive as its computational efficiency is achieved at only a small estimation efficiency penalty compared to the (NPMLE) estimator
Table 2.1: Simulation results: increasing baseline hazard, 24% local recurrence, 41% distant recurrence

<table>
<thead>
<tr>
<th>N</th>
<th>Statistics</th>
<th>Model</th>
<th>Detectable Mets</th>
<th>Recurrence</th>
<th>Interrelation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Z₁</td>
<td>Z₂</td>
<td></td>
</tr>
<tr>
<td>True</td>
<td></td>
<td></td>
<td>0.5</td>
<td>1</td>
<td>-0.5</td>
</tr>
<tr>
<td>Bias</td>
<td>M₁</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>M₂</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>ESE</td>
<td>M₁</td>
<td>0.32</td>
<td>0.22</td>
<td>0.26</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>M₂</td>
<td>0.30</td>
<td>0.21</td>
<td>0.25</td>
<td>0.17</td>
</tr>
<tr>
<td>ASE</td>
<td>M₁</td>
<td>0.32</td>
<td>0.23</td>
<td>0.27</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>M₂</td>
<td>0.31</td>
<td>0.22</td>
<td>0.25</td>
<td>0.17</td>
</tr>
<tr>
<td>CP(%)</td>
<td>M₁</td>
<td>94.0</td>
<td>93.9</td>
<td>95.5</td>
<td>94.6</td>
</tr>
<tr>
<td></td>
<td>M₂</td>
<td>93.8</td>
<td>94.0</td>
<td>95.1</td>
<td>94.4</td>
</tr>
<tr>
<td>Bias</td>
<td>M₁</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>M₂</td>
<td>0.01</td>
<td>0.00</td>
<td>-0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>ESE</td>
<td>M₁</td>
<td>0.23</td>
<td>0.16</td>
<td>0.19</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>M₂</td>
<td>0.23</td>
<td>0.16</td>
<td>0.19</td>
<td>0.12</td>
</tr>
<tr>
<td>ASE</td>
<td>M₁</td>
<td>0.24</td>
<td>0.16</td>
<td>0.20</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>M₂</td>
<td>0.23</td>
<td>0.15</td>
<td>0.19</td>
<td>0.12</td>
</tr>
<tr>
<td>CP(%)</td>
<td>M₁</td>
<td>95.0</td>
<td>95.4</td>
<td>95.1</td>
<td>94.4</td>
</tr>
<tr>
<td></td>
<td>M₂</td>
<td>94.8</td>
<td>96.2</td>
<td>94.4</td>
<td>94.5</td>
</tr>
</tbody>
</table>

(M2).

2.4.2 Breast Cancer Adjuvant Therapy

We apply the proposed method to a multi-center randomized clinical trial of early breast cancers in Canada diagnosed from 1992 to 2000 ([25, 53]). This study is designed to compare the effect of tamoxifen alone with radiotherapy plus tamoxifen as adjuvant therapy for breast-conserving surgery on disease-free survival and local relapse among women 50 years of age or older who had T1 or T2 node-negative breast cancers. Since [53] only provided a subset of the original study, our analysis results are thus not intended for efficacy interpretation, and should be considered for illustration purposes only. The events recorded in the dataset were local recurrence, axillary recurrence, distant recurrence, second malignancy of any type, and death. The time of the first occurrence of each type event was documented. We combine
Table 2.2: Simulation results: decreasing baseline hazard, 26% local recurrence, 38% distant recurrence

<table>
<thead>
<tr>
<th>N</th>
<th>Statistics</th>
<th>Model</th>
<th>Detectable Mets</th>
<th>Recurrence</th>
<th>Interrelation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True</td>
<td></td>
<td>Z₁</td>
<td>Z₂</td>
<td>Z₁</td>
</tr>
<tr>
<td>250</td>
<td>Bias</td>
<td>M1</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M2</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>ESE</td>
<td>M1</td>
<td>0.38</td>
<td>0.24</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M2</td>
<td>0.38</td>
<td>0.24</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>ASE</td>
<td>M1</td>
<td>0.36</td>
<td>0.24</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M2</td>
<td>0.36</td>
<td>0.23</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>M1</td>
<td>94.6</td>
<td>94.9</td>
<td>95.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M2</td>
<td>94.8</td>
<td>94.8</td>
<td>94.9</td>
</tr>
<tr>
<td>500</td>
<td>Bias</td>
<td>M1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M2</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>ESE</td>
<td>M1</td>
<td>0.26</td>
<td>0.16</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M2</td>
<td>0.26</td>
<td>0.16</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>ASE</td>
<td>M1</td>
<td>0.26</td>
<td>0.16</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M2</td>
<td>0.25</td>
<td>0.16</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>M1</td>
<td>95.0</td>
<td>94.8</td>
<td>95.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M2</td>
<td>95.2</td>
<td>94.6</td>
<td>95.0</td>
</tr>
</tbody>
</table>

M1: inefficient estimator; M2: weighted (efficient) estimator
ESE: empirical standard errors based on 1,000 estimates
ASE: average of estimated standard errors
CP: 95% Coverage Probability

local recurrence and axillary recurrence as local recurrence, distant recurrence and second malignancy as distant recurrence. In addition, the study population, women 50 years of age or older with early breast cancers (tumor size 5cm or less in diameter, pathological stage T1 or T2), is a relatively healthy patient group who are basically absent of immediate cancer mortality risks and have good survival. Therefore, we treat other-cause death as independent censoring for time to recurrence, the endpoint of interest.

The dataset consists of 641 women, including 321 who received tamoxifen alone (Tam) and 320 who received radiotherapy plus tamoxifen (RT+Tam). Patients’ baseline characteristics include age at diagnosis, tumor size, and hormone receptor status, which are summarized in Table 2.3. Chi-square tests show the baseline characteristics of the two groups to be similar ($p > 0.05$). The median follow-up
Table 2.3: Characteristics of breast cancer study subjects

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Tamoxifen (N=321)</th>
<th>Radiotherapy + Tamoxifen (N=320)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at diagnosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50-59 yr</td>
<td>79 (24.6)</td>
<td>84 (26.3)</td>
</tr>
<tr>
<td>60-69 yr</td>
<td>117 (36.4)</td>
<td>96 (30.0)</td>
</tr>
<tr>
<td>≥70 yr</td>
<td>125 (38.9)</td>
<td>140 (43.8)</td>
</tr>
<tr>
<td>Tumor size</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤1 cm</td>
<td>112 (34.9)</td>
<td>101 (31.6)</td>
</tr>
<tr>
<td>&gt;1-2 cm</td>
<td>106 (33.0)</td>
<td>124 (38.8)</td>
</tr>
<tr>
<td>&gt;2-5 cm</td>
<td>103 (32.1)</td>
<td>95 (29.7)</td>
</tr>
<tr>
<td>Hormone receptor status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>53 (16.5)</td>
<td>53 (16.6)</td>
</tr>
<tr>
<td>Positive</td>
<td>268 (83.5)</td>
<td>267 (83.4)</td>
</tr>
<tr>
<td>Local recurrence</td>
<td>41 (12.8)</td>
<td>5 (1.6)</td>
</tr>
<tr>
<td>Distant recurrence</td>
<td>35 (10.9)</td>
<td>45 (14.1)</td>
</tr>
<tr>
<td>Censored*</td>
<td>245 (76.3)</td>
<td>270 (84.4)</td>
</tr>
<tr>
<td>Median follow-up time (year)</td>
<td>5.1</td>
<td>5.3</td>
</tr>
</tbody>
</table>

*: including 12 non-cancer deaths (5 in Tam and 7 in RT+Tam)

time, numbers and proportions of local and distant recurrences between treatment arms are also summarized in Table 2.3. Within the study duration, 41(12.8%) vs. 5(1.6%) women in Tam and RT+Tam developed local recurrence, while 35(10.9%) and 45(14.1%) women had distant recurrence. Figure 2.2 presents Kaplan-Meier estimates of disease-free survival by treatment groups, and suggests that patients in RT+Tam group were significantly less likely to have cancer recurrences than Tam group 3 years after treatment (log-rank test \( p = 0.007 \)). Cumulative incidence functions of local recurrence by treatment groups are presented in Figure 2.3, in which RT+Tam group has lower cumulative incidence than Tam group throughout the study duration, and the difference is significant \( p < 0.001 \) based on Gray’s \( k \)-sample test for competing risks ([29]). However, more distant recurrence occurred in RT+Tam group than Tam group within the first 7 years, and the cumulative incidences of the two groups became similar afterwards. Gary’s test shows no difference between two groups in terms of cumulative incidences \( p = 0.24 \).
Figures 2.2 and 2.3 together reveal interesting patterns of disease natural history under different treatment options and prognostic factors. While RT+Tam signifi-
cantly reduced the incidence of local recurrence throughout the study duration, the overall cancer recurrence is similar between two groups for the first 3 years after treatment, and sharply diverges after 7 years. These results are accompanied by higher incidence of distant recurrence in RT+Tam group than in Tam group for the first 7 years after treatment. The role of RT+Tam in breast cancer local and overall recurrence control is of great interest. Does RT+Tam contribute to an overall reduction of local recurrence, or it masks some local recurrences that become metastatic as they surface later? More importantly, what in general is the mechanistic effect of treatment and prognostic factors here? To answer these questions, we analyzed this dataset with the proposed mechanistic model to investigate the relationship between recurrence, disease natural history and treatment/prognostic factors. For comparison purposes, we also analyze the crude local and distant recurrence hazards ([36], section 8.2) using the conventional Cox proportional hazard (PH) model and proportional odds (PO) model ([9]). The PO model is often considered as an alternative to the PH model, particularly when hazards show convergence over time, and its estimation is based on the self-consistent maximum likelihood estimator ([61]). The covariates included in above regression analysis are: treatment arm (RT+Tam vs. Tam, Z$_1$), hormone-receptor status (positive vs. negative, Z$_2$), tumor size ($\geq 2$cm vs. $< 2$cm, Z$_3$), and age at diagnosis ($50 - 59$ vs. $\geq 60$, Z$_4$). The results are summarized in Table 2.4.

The PH and PO crude hazard models both suggest that patients who received RT+Tam, with negative hormone receptor status, smaller tumor size and older age at diagnosis were less likely to develop local recurrence than their counterparts. Our proposed approach agrees with these findings, with a slightly attenuated treatment effect when comparing with crude hazard models. For distant recurrence, PH and
Table 2.4: Breast cancer study analysis results: conventional and proposed approaches

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Crude Hazard PH</th>
<th>Crude Hazard PO</th>
<th>Joint Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est.</td>
<td>SE</td>
<td>p-value</td>
</tr>
<tr>
<td>Local Recurrence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z_1</td>
<td>-2.22</td>
<td>0.48</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Z_2</td>
<td>-1.21</td>
<td>0.39</td>
<td>0.002</td>
</tr>
<tr>
<td>Z_3</td>
<td>0.78</td>
<td>0.30</td>
<td>0.010</td>
</tr>
<tr>
<td>Z_4</td>
<td>1.05</td>
<td>0.31</td>
<td>0.001</td>
</tr>
<tr>
<td>Distant Recurrence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z_1</td>
<td>0.19</td>
<td>0.23</td>
<td>0.418</td>
</tr>
<tr>
<td>Z_2</td>
<td>-0.44</td>
<td>0.38</td>
<td>0.251</td>
</tr>
<tr>
<td>Z_3</td>
<td>-0.02</td>
<td>0.26</td>
<td>0.938</td>
</tr>
<tr>
<td>Z_4</td>
<td>0.52</td>
<td>0.23</td>
<td>0.025</td>
</tr>
<tr>
<td>Detectable Metastasis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z_1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z_2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z_3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z_4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interrelation</td>
<td>2.07</td>
<td>0.48</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Z_1: randomization arm: RT+Tam vs. Tam  
Z_2: hormone-receptor status: positive vs. negative  
Z_3: tumor size ≥2cm vs. <2cm  
Z_4: age at diagnosis 50-59 vs. ≥60

PO crude hazard models also provide direct covariate effect estimates, and find older age at diagnosis associated with decreased incidence. In contrast, the proposed joint model provides statistical inference about disease natural history, e.g., time to detectable metastasis, and a measure of the association between metastasis and recurrence. We find that patients who received RT+Tam ($p < 0.001$) and were diagnosed at a younger age ($p = 0.022$) were more likely to develop detectable metastases, and the occurrence of metastasis would significantly accelerate the occurrence of (distant) recurrence ($p < 0.001$). We also conclude that age at diagnosis is negatively associated with distant recurrence ($-0.87 + 2.07 > 0$). In fact, our findings are more in line with observations in Figure 2.2 and 2.3, and explain the discrepancy in distant recurrence between treatment groups in the first 7 years. We may speculate that RT may only destroy more fragile less aggressive tumor cells leaving the most aggressive (perhaps dormant) subtype intact thus masking the aggressive malignant fraction that may metastasize sooner in the absence of competition from the destroyed cells. Our estimated association measure indicates that the occurrence of metastases is
positively related with recurrence.

In summary, our proposed model provides additional insights about the disease natural history and the treatment effect that may help understand the role of treatments and prognostic factors.

2.5 Discussion

In this paper, we have presented a general framework to analyze time-to-event marked endpoints in cancer studies using a semiparametric mechanistic regression model. By viewing the observed marked endpoints as cross-sectional observations of latent disease natural history, we aim to understand the disease progression and its relationship with observed data and covariates (e.g., treatment interventions and prognostic factors). Comparing with existing methods that analyze observed data, e.g., methods for univariate survival data or competing risks, our model provides clinically meaningful dependence measures for the outcome variables, and a predictive utility for all possible outcomes through a joint model. Unlike some previous mechanistic approaches, our method is not restricted to the tumor size-metastasis data only treating the point of diagnosis as censoring (nuisance). Instead we seek to explore the power of a joint model for both the event time as the mark viewed as a manifestation of the same disease. This approach, in particular, provides natural assumptions that make the model identifiable.

We develop two estimation procedures based on ideas previously proposed for a different class of models. One is based on estimating equations (EE) motivated by martingale property, and the other one is a maximum likelihood estimator (MLE). Asymptotic properties of both estimators are established and finite sample properties are studied by simulation. While the EE-based estimator is not fully efficient, we did
not observe much efficiency loss compared to MLE, and this made the computational speed and convenience of the inefficient estimator a good deal in this specific analysis.

Our framework offers a range of possible extensions for models used in event history analysis. Depending on the disease natural history of interest and the observed data, different multistate latent structures can be used to explain a more complex mark. For example, progression in stage and grade of cancer could be incorporated. The survival component could be extended to multivariate data, for example jointly modeling recurrence and cancer death. The proportionality assumption between latent progression and observed survival can be relaxed without modifying the general line of the paper by assuming that the two hazards are linked by a transformation model.
CHAPTER III

Joint Modeling Approach for Semicompeting Risks Data
with Missing Nonterminal Event Status

3.1 Introduction

In many biomedical studies, patients may experience multiple distinct causes of failures. With competing risks data, the observed failure time and cause may be considered as the minimum of a set of potential failure times corresponding to different causes, and the first occurrence of any cause censors the occurrences of other causes. When only one of the two failure causes (terminal) can censor the other but not vice versa, e.g., a nonterminal event is observed first and then a terminal event is observed subsequently, the risks are called semicompeting ([21]). It is noted that semicompeting risks data is more informative than competing risks data. An example of semicompeting risks data involves time to morbidity and time to death (mortality), where morbidity (nonterminal event) usually represents the observed outcome of disease progression, and mortality (terminal event) represents the endpoint of the disease natural history.

While semicompeting risks data often arise in biomedical context, different underlying models may lead to seemingly identical observations. The observation of terminal event such as death is usually well defined and objective. However, absence of nonterminal event may be interpreted as right censoring when the event has not
occurred yet, or left censoring when the non-terminal event occurred but was not registered. Further, upon absence of the non-terminal event it may be unknown which of the two possibilities takes place. In cancer, recurrence or progression and death often represent semicompeting risks data. Here disease progression may be established through a series of clinical and/or radiologic measurements. Observation of such nonterminal events may be missing (unobserved) dependent on what procedures are performed and how they are registered in the database. As a result, an unknown proportion of subjects may have their nonterminal events occurred but unobserved, such that the observed semicompeting risks data become a mixture of “true” semicompeting risks data and terminal event only data with unknown nonterminal event status. This situation may occur by design when a mortality database that does not have recurrence information is combined with a clinical dataset where such information is partially or fully available.

Because times to the nonterminal and terminal events of an individual are usually correlated, the censoring of the nonterminal event by the terminal event is informative. As a result, the conventional analysis based on the independent censoring assumption may be biased ([74, 33]), and efficiency loss may arise if the association between the events is ignored.

Understanding the association between the events is interesting in itself as an important component of the disease natural history useful in prognostic and therapeutic studies. Learning about the disease natural history through an observation of a non-terminal event and adapting treatment to this knowledge is an interesting problem.

As pointed out by [52], it is not possible to analyze semicompeting risks data in a full nonparametric model, as it inherits the non-identifiability issue of competing
Therefore semicompeting risks have been addressed in a semiparametric framework. A copula model for the two event times has been considered under a restriction on the upper wedge $T_1^* \leq T_2^*$ of the positive quadrant, where $T_1^*, T_2^*$ are times to nonterminal and terminal events, respectively. [21] proposed a semiparametric estimation approach based on Clayton copula ([16]), and [63] extended this framework to more general copula models. [52], [32] and more recently, [15] suggested using semiparametric transformation models for the marginal regressions, and a copula model for the joint distribution. All of these models proceed from latent failure times and are similar in spirit to competing risks models, e.g., [20]. Therefore, the interpretation of the marginal distribution of the nonterminal event in this framework is hypothetical.

Alternatively, a multistate illness-death models ([36, 31]), Figure 3.1, can be adopted. Associations can be incorporated through effects on transition intensity functions. Let $\lambda_{ij}(t|t_i)$ be the intensity of transition from state $i$ to state $j$, where $t_i$ is the time spent in state $i$. Markov or semi-Markov illness-death models represent specific assumption on the form of $\lambda_{ij}$. A variety of proportional hazards (PH) models have been considered by [5] and [36](p.270) (Markov models); [4] and [57] (semi-Markov models).

[19] proposed a parametric model based on the joint distribution of terminal and non-terminal events, treating unobserved non-terminal events as dependently right censored.

In this paper we extend the PH framework to incorporate a possibility that non-terminal event is unobserved either because it is censored or because the observation is missing, and it is generally unknown which of the two possibilities takes place.

A shared-frailty effect ([50, 28]) has been a popular tool to induce positive de-
Figure 3.1: A three-state illness-death model for semicompeting risks with partially missing non-terminal event

For semi-competing risks data, [64] incorporates gamma frailty into the illness-death model. However, conditional specification of the model where non-terminal event affects the risk for the terminal event may be more attractive when the dependence cannot be assumed positive. Specifically we use an idea similar to so-called Markov modulated models based on intensity processes (e.g., [36] section 9.2, [17] section 5.3) used with recurrent data analysis. This leads us to an approach similar to transformation models in survival analysis ([12, 61, 13]).

Partial-likelihood-based approaches conditioning on the past history become inappropriate when the past history is not fully observable ([6]). Therefore our paper is based on a full likelihood approach.
The paper is organized as follows. In section 3.2, we describe the illness-death model and its properties in terms of bivariate survival distributions, as well as the data structure to be used throughout the paper. In section 3.3 we introduce the likelihood construction when missing nonterminal event status is possible. The maximum likelihood estimation procedure and asymptotic properties of the estimators for regression coefficients are discussed in section 3.4. We present Monte Carlo simulation studies and a data analysis of follicular cell lymphoma patients in section 3.6, and conclude the article with a discussion in section 3.7.

3.2 Model and Data Structure

3.2.1 Notation and Model

Let $T^*_1$ and $T^*_2$ be the potential times to nonterminal and terminal events, respectively; $Z(t)$ be the covariate vector of dimension $p$; $N^*_1(t) = I(T^*_1 < t)$ be the counting processes for nonterminal event. We define the joint distribution of $T^*_1$ and $T^*_2$ through the marginal and conditional hazard functions, respectively:

\begin{align}
\lambda_1(t|Z(t)) &= \lim_{\Delta \to 0} P(T^*_1 \in [t, t + \Delta]|T^*_1 \geq t, T^*_2 \geq t, Z(t))/\Delta = h_0(t)\eta(Z(t)), \\
\lambda_2(t|Z(t), N^*_1(t)) &= \lim_{\Delta \to 0} P(T^*_2 \in [t, t + \Delta]|T^*_2 \geq t, Z(t), N^*_1(t))/\Delta = h_0^*(t)\theta(Z(t))\mu(Z(t))^{N^*_1(t)},
\end{align}

where $\eta(Z(t)) = e^{\beta_0 Z(t)}$, $\theta(Z(t)) = e^{\beta_0 Z(t)}$, $\mu(Z(t)) = e^{\beta_0 + \beta_0 Z(t)}$, and $h_0, h_0^*$ are the baseline hazard functions. For convenience, we consider time-independent covariate $Z$ from now on, and keep in mind that the proposed model may accommodate external time-dependent covariate $Z(t)$ ([36] section 6.4) as well.

The proposed specification can be viewed as a combination of two commonly used models: 1) a Cox proportional hazard (PH) model for $T^*_1$; and 2) a Cox PH model
for $T_2^*$ with the counting process for $T_1^*$ as a time-dependent covariate. Alternatively, one may view the proposed model in terms of an illness-death process (Figure 3.1), where $\lambda_1$ is the transition hazard between state 0 and 1 ($\lambda_{01}$), $\lambda_2$ is the path transition hazard between state 0 and 2 which may depend on whether state 1 occurs ($\lambda_{12}$) or not ($\lambda_{02}$). Furthermore, $\mu(Z)$ may be interpreted as the hazard ratio between $\lambda_2$ and $\lambda_{12}$, which defines the dependence between nonterminal and terminal events, as discussed in section 3.1. $\mu(Z) = 1$ corresponds to no correlation between $T_1^*$ and $T_2^*$; $\mu(Z) > 1$ or $\mu(Z) < 1$ represent positive or negative association between $T_1^*$ and $T_2^*$ as the occurrence of $T_1^*$ accelerates or decelerates the occurrence of $T_2^*$.

The model (3.1) and (3.2) defines a bivariate distribution for $(T_1^*, T_2^*)$ on the positive quadrant of $(t_1 \geq 0, t_2 \geq 0)$. For semicompeting risks (observed) data, observations are restricted to the upper wedge $0 \leq t_1 \leq t_2$ and are positively correlated even if the potential times are not. Following [64], define $T_1^* = \infty$ if terminal event occurs before nonterminal event, such that there is no probability content in the lower wedge $t_2 < t_1 < \infty$. The probability model for $(T_1^*, T_2^*)$ is taken to be absolutely continuous with joint density $f(t_1, t_2)$ defined in the upper wedge $0 < t_1 \leq t_2$, and continuous along the line at $t_1 = \infty$ with density $f_\infty(t_2), t_2 > 0$, such that $\int_0^\infty \int_{t_1}^\infty f(t_1, t_2) dt_2 dt_1 + \int_0^\infty f_\infty(t_2) dt_2 = 1$. This definition provides validity for the marginal distribution of $T_1^*$, which is not well defined in semicompeting risks data.

Assume within the PH framework that $h_0 = h_0^* = h$, i.e., proportionality relationship holds between $T_1^*$ and $T_2^*$. The proportionality assumption yields close-form expressions, in section 3.4. We will discuss approaches to relax this assumption in section 3.7. Introduce the following notation $\eta = \eta(Z), \theta = \theta(Z), \mu = \mu(Z)$, $\bar{\mu} = 1 - \mu$, $H(t) = \int_0^t h(x) dx$, and $D_k = -\frac{\partial}{\partial \nu_k}$ for $k = 1, 2$. We can subsequently derive the joint probability density functions ($f$) and survival functions ($S$) of $T_1^*$.
and $T^*_2$ on the upper wedge $0 \leq t_1 \leq t_2$ as

$$S_1(t_1, t_2) = S(t_1, t_2, t_1 \leq t_2) = \frac{\eta \theta \mu}{\eta + \theta \mu} \left\{ \frac{e^{-H(t_1)(\eta + \theta \mu) - H(t_2)\theta \mu}}{\theta \mu} - \frac{e^{-H(t_2)(\eta + \theta)}}{\eta + \theta} \right\},$$

$$f_1(t_1, t_2) = D_1D_2S_1(t_1, t_2) = h(t_1)h(t_2)\eta \theta \mu e^{-H(t_1)(\eta + \theta \mu) - H(t_2)\theta \mu},$$

$$D_1S_1(t_1, t_2) = h(t_1)\eta e^{-H(t_1)(\eta + \theta \mu) - H(t_2)\theta \mu},$$

$$D_2S_1(t_1, t_2) = \frac{h(t_2)\eta \theta \mu}{\eta + \theta \mu} \left\{ e^{-H(t_1)(\eta + \theta \mu) - H(t_2)\theta \mu} - e^{-H(t_2)(\theta + \eta)} \right\},$$

When the subject fails before the nonterminal event occurs, i.e., $t_2 < t_1 = \infty$ the probability density and survival functions are

$$f_\infty(t_2) = h(t_2)\theta e^{-H(t_2)(\theta + \eta)},$$

$$S_\infty(t_2) = \frac{\theta}{\theta + \eta} e^{-H(t_2)(\theta + \eta)}.$$

Figure 3.2 visualizes the relationship between joint survival function of $T^*_1, T^*_2$ and corresponding marginal survival functions on the upper wedge $0 \leq t_1 \leq t_2$, and gives the copula of the proposed model.

Kendall's $\tau$ is a popular global measure of dependence between bivariate survival times. One can show that the Kendall’s $\tau$ on the upper wedge $0 \leq t_1 \leq t_2$ is given by:

$$\tau = 4 \times \int_0^\infty \int_0^y S_1(x, y)f_1(x, y)dxdy - 1 = \frac{\theta \mu}{\eta + \theta + \theta \mu}. $$

Remarkably, this measure does not depend on the baseline hazard function, a consequence of the proportionality assumption.

The crossratio function ([16, 51]) is commonly used to measure the local dependence between bivariate survival times. For the upper wedge of $0 \leq t_1 \leq t_2$, the
Figure 3.2: Copula plot when $\mu = 2, 1, 0.5$, representing a positive, neutral, and negative association between nonterminal and terminal events
crossratio function $\gamma(t_1, t_2)$ is as follows:

$$
\gamma(t_1, t_2) = \frac{S(t_1, t_2)f(t_1, t_2)}{\{D_1S(t_1, t_2)\}\{D_2S(t_1, t_2)\}} = \frac{\lambda_2(t_2|T_1 = t_1)}{\lambda_2(t_2|T_1 > t_1)} = \frac{\lambda_1(t_1|T_2 = t_2)}{\lambda_1(t_1|T_2 > t_2)}
$$

$$
= 1 + \frac{\eta + \theta \bar{\mu}}{\eta + \theta} \frac{1}{e^{(H(t_2) - H(t_1))(\eta + \theta \bar{\mu})} - 1}.
$$

We notice that $\gamma(t_1, t_2)$ cannot be written in terms of $S(t_1, t_2)$ alone, and thus the corresponding copula does not belong to the Archimedean class ([51]).

### 3.2.2 Observed Data Structure and Counting Process Notation

Let $V^*$ be the censoring time independent of $(T_1^*, T_2^*)$ given $Z$; $\xi$ be the maximum follow-up time in the study. Suppose $(T_{1i}^*, T_{2i}^*, V_{i}^*, Z_i), i = 1, \ldots, n$ are $n$ independent and identically distributed replicates of $(T_1^*, T_2^*, V^*, Z)$, such that the observed semicompeting risks data for subject $i, i = 1, \ldots, n$, consists of $\{T_i, \tilde{T}_i, \delta_{ki}, \tilde{\delta}_{2i}, Z_i; k = 1, 2, 0 \leq t \leq \xi\}$, where $T_i = \min(T_{1i}^*, T_{2i}^*, V), \tilde{T}_i = \min(T_{2i}^*, V)$ with $V = \min(V^*, \xi)$, $\delta_{ki} = I(T_i = T_{ki}^*), k = 1, 2, \tilde{\delta}_{2i} = \delta_{1i}I(\tilde{T}_i = T_{2i}^*)$, where $I(\cdot)$ is the indicator function. Because $0 \leq T_i \leq \tilde{T}_i$, the observations are restricted to the upper wedge. Furthermore, if $\delta_{1i} = 0, \delta_{2i} = 1, \tilde{\delta}_{2i} = 0$, then $T_i = \tilde{T}_i = T_{2i}^*$ and $T_{1i}^* = \infty$ by definition.

In counting process notation, for subject $i$, let $N_{ki}^* = I(T_{ki}^* \leq t), k = 1, 2$, be the underlying counting processes for nonterminal ($k = 1$, type 1) and terminal ($k = 2$, type 2) events, $Y_{i}(t) = I(T_i \geq t)$ be the at-risk process for the first-occurring event regardless of event type, $\tilde{Y}_{i}(t) = I(\tilde{T}_i \geq t > T_i)$ be the at-risk process for the terminal event, $N_{ki}(t) = \int_0^t Y_{i}(x)dN_{k}^*(x) = \delta_{ki}N_{k}^*(t), (k = 1, 2)$, be the observed counting process of event type $k$ occurring first, and $\tilde{N}_{2i} = \delta_{1i}\tilde{\delta}_{2i}I(T_{2i}^* \leq t)$ be the observed counting process of terminal event occurring after nonterminal event. In addition, we assume $\Pr(T_{1}^* = T_{2}^*)=0$.

The observed semicompeting risks data $\{T_i, \tilde{T}_i, \delta_{1i}, \delta_{2i}, \tilde{\delta}_{2i}, Z_i; i = 1, \ldots, n\}$ are expressed through three counting processes $N_{1i}(\cdot), N_{2i}(\cdot), \tilde{N}_{2i}(\cdot)$, and two corresponding
at-risk processes $Y_i(\cdot), \tilde{Y}_i(\cdot)$ for subject $i$. Note that the counting processes $N_{1i}(\cdot)$ and $N_{2i}(\cdot)$ record the time and type of subject $i$’s first event, and $\tilde{N}_{2i}(\cdot)$ records the time of event type 2 (terminal). Furthermore, $\tilde{N}_{2i} = 0$ if second event could not be observed for subject $i$ either because event type 1 has been censored by event type 2 or independent censoring event.

### 3.3 Likelihood with Missing Nonterminal Event Status

When nonterminal event status could be missing, the observed semicompeting risks data become a mixture of subjects whose disease history is completely or partially observed. Throughout the paper, we distinguish the observed status of one’s nonterminal event from the missing status indicator showing whether one’s nonterminal event status can be observed. The former is an observed event status (i.e., $\delta_{1i}$); the latter is assumed to be a binary unobserved missing data variable following a missing at random mechanism ([43]), given covariates. Let $M_i$ be the indicator of whether one’s nonterminal event is missing, i.e., $M_i = 1$ if the nonterminal event status is unobservable, and $M_i = 0$ if it is observable. It is natural to assume $\Pr(M_i = 1|Z_i) = p(Z_i)$, where $p(Z_i)$ is following a logistic model ([45]), $p(Z_i) = p = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$. Covariates can be easily added to this formulation. Furthermore, observing non-terminal event $\delta_{1i} = 1$ implies $M_i = 0$, while $\delta_{1i} = 0$ indicates subject $i$ may either be: (i) $M_i = 1$, if the nonterminal event has occurred but is unobservable; or (ii) $M_i = 0$ or $M_i = 1$, if the nonterminal event has not occurred yet. Subjects whose nonterminal events are not observed represent a mixture of (i) subjects whose nonterminal event has occurred but is unobservable, and (ii) subjects whose nonterminal event truly has not occurred yet.

As a result, the likelihood contribution of subject $i$ will take one of the following
four forms:

- If both nonterminal and terminal events are observed sequentially, i.e., \( T_i < \tilde{T}_i, \delta_{1i} = 1, \delta_{2i} = 0, \tilde{\delta}_{2i} = 1 \):

\[
L_{1i}^* = \Pr(T_{1i}^* = T_i, T_{2i}^* = \tilde{T}_i, M_i = 0) = \Pr(M_i = 0)\Pr(T_{1i}^* = T_i, T_{2i}^* = \tilde{T}_i | M_i = 0)
= (1 - p) \times f(T_i, \tilde{T}_i) = (1 - p) \times h(T_i)h(\tilde{T}_i)\eta_i\theta_i\mu_i e^{-H(T_i)(\eta_i + \theta_i, \mu_i)} - H(\tilde{T}_i)\theta_i\mu_i,
\]

- If nonterminal event is observed, but terminal event is censored, i.e., \( T_i < \tilde{T}_i, \delta_{1i} = 1, \delta_{2i} = 0, \tilde{\delta}_{2i} = 0 \):

\[
L_{2i}^* = \Pr(T_{1i}^* = T_i, T_{2i}^* \geq \tilde{T}_i, M_i = 0) = \Pr(M_i = 0)\Pr(T_{1i}^* = T_i, T_{2i}^* \geq \tilde{T}_i | M_i = 0)
= (1 - p) \times D_1 S(t_1, t_2) | (T_i, \tilde{T}_i) = (1 - p) \times h(T_i)\eta_i e^{-H(T_i)(\eta_i + \theta_i, \mu_i)} - H(\tilde{T}_i)\theta_i\mu_i,
\]

- If nonterminal event is not observed, and terminal event is observed, i.e., \( T_i = \tilde{T}_i, \delta_{1i} = 0, \delta_{2i} = 1, \tilde{\delta}_{2i} = 0 \):

\[
L_{3i} = \Pr(T_{1i}^* < T_i, T_{2i}^* = T_i, M_i = 1) + \Pr(T_{1i}^* > T_i, T_{2i}^* = T_i)
= p \times D_2 S(t_1, t_2) | (0, T_i) + f_\infty(T_i)
= p \times h(T_i) \frac{\eta_i\theta_i\mu_i}{\eta_i + \theta_i\mu_i} \{e^{-H(T_i)(\eta_i + \theta_i, \mu_i)} - e^{-H(T_i)(\eta_i + \theta_i)}\} + h(T_i)\theta_i e^{-H(T_i)(\eta_i + \theta_i)},
\]

- If neither event is observed, i.e., \( T_i = \tilde{T}_i, \delta_{1i} = 0, \delta_{2i} = 0, \tilde{\delta}_{2i} = 0 \):

\[
L_{4i} = \Pr(T_{1i}^* \leq T_i, T_{2i}^* \geq T_i, M_i = 1) + \Pr(T_{1i}^* \geq T_i, T_{2i}^* \geq T_i)
= p \times \frac{\eta_i}{\eta_i + \theta_i\mu_i} \{e^{-H(T_i)(\eta_i + \theta_i, \mu_i)} - e^{-H(T_i)(\eta_i + \theta_i)}\} + e^{-H(T_i)(\eta_i + \theta_i)}.
\]

Note that we can alternatively express \( L_{1i}^* \) and \( L_{2i}^* \) as

\[
L_{1i}^* = \Pr(T_{1i}^* = T_i, T_{2i}^* = \tilde{T}_i, M_i = 0) = L_{1i} \times \tilde{L}_{1i},
\]

\[
L_{2i}^* = \Pr(T_{1i}^* = T_i, T_{2i}^* \geq \tilde{T}_i, M_i = 0) = L_{1i} \times \tilde{L}_{2i}.
\]
where

\[
\mathcal{L}_{1i} = \Pr(M_i = 0) \Pr(T_{1i}^* = T_i | M_i = 0) = (1 - p) \times h(T_i) \eta_i e^{-H(T_i)(\eta_i + \theta_i)}
\]

\[
\tilde{\mathcal{L}}_{1i} = \Pr(T_{2i}^* = \tilde{T}_i | T_{1i}^* = T_i, M_i = 0) = h(\tilde{T}_i)\theta_i \mu_i e^{-(H(\tilde{T}_i) - H(T_i))\theta_i \mu_i}
\]

\[
\tilde{\mathcal{L}}_{2i} = \Pr(T_{2i}^* \geq \tilde{T}_i | T_{1i}^* = T_i, M_i = 0) = e^{-(H(\tilde{T}_i) - H(T_i))\theta_i \mu_i},
\]

such that the contribution from nonterminal event is separated from terminal event.

Therefore, the contribution of subject \(i\) in the log-likelihood is

\[
\ell_i = \delta_{1i}\tilde{\delta}_{2i} \log(\mathcal{L}_{1i}^*) + \delta_{1i}(1 - \tilde{\delta}_{2i}) \log(\mathcal{L}_{2i}^*) + \delta_{2i} \log(\mathcal{L}_{3i}) + (1 - \delta_{1i} - \delta_{2i}) \log(\mathcal{L}_{4i})
\]

\[
= \{\delta_{1i} \log(\mathcal{L}_{1i}) + \delta_{2i} \log(\mathcal{L}_{3i}) + (1 - \delta_{1i} - \delta_{2i}) \log(\mathcal{L}_{4i})\} + \delta_{1i}\tilde{\delta}_{2i} \log(\tilde{\mathcal{L}}_{1i}) + (1 - \tilde{\delta}_{2i}) \log(\tilde{\mathcal{L}}_{2i})\}.
\]

(3.3) \[
\delta_{1i}\{\tilde{\delta}_{2i} \log(\tilde{\mathcal{L}}_{1i}) + (1 - \tilde{\delta}_{2i}) \log(\tilde{\mathcal{L}}_{2i})\}:
\]

Using notation \(\ell_{1i} = \delta_{1i} \log(\mathcal{L}_{1i}) + \delta_{2i} \log(\mathcal{L}_{3i}) + (1 - \delta_{1i} - \delta_{2i}) \log(\mathcal{L}_{4i})\), \(\ell_{2i} = \delta_{1i}\{\tilde{\delta}_{2i} \log(\tilde{\mathcal{L}}_{1i}) + (1 - \tilde{\delta}_{2i}) \log(\tilde{\mathcal{L}}_{2i})\}\), and \(\ell_1 = \sum_{i=1}^n \ell_{1i}, \ell_2 = \sum_{i=1}^n \ell_{2i}\), we can write the full log-likelihood as

\[
\ell = \sum_{i=1}^n \ell_i = \sum_{i=1}^n \{\ell_{1i} + \ell_{2i}\} = \ell_1 + \ell_2.
\]

The log-likelihood contribution of semicompeting risks data from subject \(i\) can be partitioned into two parts, \(\ell_{1i}\) and \(\ell_{2i}\), where \(\ell_{1i}\) is based on information in the time period \(0 \leq t \leq T_i\), with \(T_i\) being the time of first-occurring event for subject \(i\), while \(\ell_{2i}\) is based on the extended information in the time period \(T_i \leq t \leq \tilde{T}_i\), with \(\tilde{T}_i\) being the time of second-occurring event (terminal event), if any. When the first-occurring event is the terminal event or independent censoring, such that no extended observation is available, \(\ell_{2i} = 0\) by definition. Furthermore, \(\ell_1\) is the log-likelihood based on competing risks where only the first-occurring event could be observed for every subject, and \(\ell_2\) is the additional information available due to semicompeting nature of the risks.
3.4 Nonparametric Maximum Likelihood Estimation

3.4.1 Martingale Theory

Recall that for subject $i$, $N_{ki} = \delta_{ki}I(T^*_{ki} \leq t), (k = 1, 2)$ is the counting process of event type $k$ as the first-occurring event, and $\bar{N}_{i2} = \delta_{i1}\delta_{2i}I(T^*_{2i} \leq t)$ the counting process of event type 2 (terminal event) as observed after the type 1 event. In addition, $Y_i(t) = I(T_i \geq t)$ and $\bar{Y}_i(t) = I(\bar{T}_i \geq t > T_i)$ are the at-risk processes corresponding to $N_{ki}(t), k = 1, 2$ and $\bar{N}_{2i}(t)$, respectively. In addition, denote $\beta = (\beta\eta, \beta\theta, \beta\mu, \beta\rho)$ and the full parameter set $\Omega = (\beta, H(\cdot))$, where $\beta$ is finite-dimensional and $H(\cdot)$ is the infinite-dimensional parameter. Using the NPMLE approach, we treat $H(\cdot)$ as a nondecreasing step function with jumps only at the time points where events are observed [61, 70, 13]. Thus $\{dH\}$ is the collection of the jump sizes of $H$ at the observed event times.

The following martingales can be constructed based on observed counting processes with respect to filtration $\mathcal{F}_{t^-} = \sigma\{N_{1i}(x), N_{2i}(x), \bar{N}_{2i}(x), Y_i(x), \bar{Y}_i(x), Z_i : x \in [0, t), i = 1, \cdots, n\}$ as

$$dM_{ki}(t) = dN_{ki}(t) - \Theta_{ki}(t; \beta, H)dH_t, k = 1, 2,$$

$$d\bar{M}_{2i}(t) = d\bar{N}_{2i}(t) - \bar{\Theta}_{2i}(\beta)dH_t,$$

where

$$\Theta_{ki}(t; \beta, H) = \frac{(1 - p) \times \eta_p e^{-H(t)(\eta_i + \theta_i)}}{p \times \eta_0 e^{-H(t)(\eta_i + \theta_i)} + e^{-H(t)(\eta_i + \theta_i)}},$$

$$\Theta_{2i}(t; \beta, H) = \frac{\eta_0 e^{-H(t)(\eta_i + \theta_i)} - e^{-H(t)(\eta_i + \theta_i)} + \theta_i e^{-H(t)(\eta_i + \theta_i)}}{\eta_0 e^{-H(t)(\eta_i + \theta_i)} + e^{-H(t)(\eta_i + \theta_i)}},$$

$$\bar{\Theta}_{2i}(\beta) = \theta_i \mu_i,$$

where $\Theta_{ki}(t; \beta, H) = \Theta_{ki}(t^-; \beta, H)$ such that $\Theta_{ki}(t; \beta, H), k = 1, 2$ are predictable.
The compensators $\Theta_{ki}(t; \beta, H)$, $k = 1, 2$ and $\tilde{\Theta}_{2i}(\beta)$ can be derived through the following probabilistic argument,

$$E\{dN_{ki}(t)|\mathcal{F}_i(t-)\} = Y_i(t)Pr(dN_{ki}(t) = 1|Y_i(t) = 1) = Y_i(t)\Theta_{ki}(t; \beta, H)dH_t, \quad k = 1, 2;$$

$$E\{d\tilde{N}_{2i}(t)|\mathcal{F}_i(t-)\} = \tilde{Y}_i(t)Pr(d\tilde{N}_{1i}(t) = 1|\tilde{Y}_i(t) = 1) = \tilde{Y}_i(t)\tilde{\Theta}_{2i}(\beta)dH_t,$$

where

$$\Pr(dN_{1i}(t) = 1|Y_i(t) = 1) = \frac{\Pr(T^*_{1i} = t, T^*_{2i} \geq t, M_i = 0)}{\Pr(T^*_{1i} \leq t, T^*_{2i} \geq t, M_i = 1) + \Pr(T^*_{1i} \geq t, T^*_{2i} \geq t)},$$

$$\Pr(dN_{2i}(t) = 1|Y_i(t) = 1) = \frac{\Pr(T^*_{1i} < t, T^*_{2i} = t, M_i = 1) + \Pr(T^*_{1i} > t, T^*_{2i} = t)}{\Pr(T^*_{1i} \leq t, T^*_{2i} \geq t, M_i = 1) + \Pr(T^*_{1i} \geq t, T^*_{2i} \geq t)},$$

$$\Pr(d\tilde{N}_{1i}(t) = 1|\tilde{Y}_i(t) = 1) = \frac{\Pr(T^*_{1i} = t^*, T^*_{2i} = t, M_i = 0)}{\Pr(T^*_{1i} = t^*, T^*_{2i} \geq t, M_i = 0)}.$$

### 3.4.2 Score Function and Estimating Equation

Let $t_*$ be some observed event time, $dH_* = dH(t_*)$ be the jump size of the step function $H(\cdot)$, and denote the derivative with respect to $\{dH\}$ and $\beta$ of $\Theta_{ki}(t; \beta, H)$ and $\tilde{\Theta}_{2i}(\beta)$ as: $\Theta_{ki,H}(t; \beta, H) = \partial \Theta_{ki}(t; \beta, H)/\partial dH$, $\Theta_{ki,\beta}(t; \beta, H) = \partial \Theta_{ki}(t; \beta, H)/\partial \beta$, and $\tilde{\Theta}_{2i,\beta}(\beta) = \partial \tilde{\Theta}_{2i}(\beta)/\partial \beta$, where for any functional $f$

$$\frac{\partial f(H(\cdot))}{\partial dH(t)} = \frac{df(H(\cdot) + a \times 1(\cdot - t))}{da} \bigg|_{a=0},$$

where $1(x) = 1$ when $x \geq 0$ and 0 otherwise (an impulse function), is a directional functional derivative with respect to $dH(t)$. Note that this definition corresponds to taking a derivative with respect to a jump of $H$ at time $t$ if $H$ is a step-function.

It is important to notice that we can express $\phi_i(t; \beta, H) = \exp \left\{ - \int_0^t \sum_{k=1}^2 Y_i(x)\Theta_{ki}(x; \beta, H) dH_x \right\} = p \times \frac{\eta_k}{\eta_k + \theta_i \mu_i} \left\{ e^{-H(t)\theta_i \mu_i} - e^{-H(t)(\eta_k + \theta_i)} \right\} + e^{-H(t)(\eta_k + \theta_i)}.$
Therefore, the log-likelihood \( \ell = \ell_1 + \ell_2 \) can be re-expressed in counting process notation as

\begin{align}
\ell_1 &= \sum_{i=1}^{n} \sum_{k=1}^{2} \left\{ \int_{0}^{\xi} \left[ \log dH_x + \log \Theta_{ki}(x; \beta, H) \right] dN_{ki}(x) - Y_i(x)\Theta_{ki}(x; \beta, H)dH_x \right\} \\
\ell_2 &= \sum_{i=1}^{n} \left\{ \int_{0}^{\xi} \left[ \log dH_x + \log \tilde{\Theta}_{2i}(\beta) \right] d\tilde{N}_{2i}(x) - \tilde{Y}_i(x)\tilde{\Theta}_{2i}(\beta)dH_x \right\},
\end{align}

By taking derivative of the re-expressed log-likelihood with respect to \( \{dH\} \), we obtain the following score function for \( \{dH\} \) as:

\begin{align}
U_H &= \sum_{i=1}^{n} \int_{0}^{\xi} \left\{ \sum_{k=1}^{2} \left[ dN_{ki}(x) - Y_i(x)\omega_{ki}(x; \beta, H)\Theta_{ki}(x; \beta, H)dH_x \right] + d\tilde{N}_{2i}(x) - \tilde{Y}_i(x)\tilde{\Theta}_{2i}(\beta)dH_x \right\},
\end{align}

where

\[ \omega_{ki}(t; \beta, H) = 1 - \frac{\int_{\xi+}^t \psi_{ki}(x; \beta, H)dM_{ki}(x)}{\Theta_{ki}(t; \beta, H)}; \quad \psi_{ki}(t; \beta, H) = \frac{\Theta_{ki,H}(t; \beta, H)}{\Theta_{ki}(t; \beta, H)}. \]

The score function for \( \beta \) is

\begin{align}
U_\beta &= \sum_{i=1}^{n} \int_{0}^{\xi} \left\{ \sum_{k=1}^{2} \left[ \frac{\Theta_{ki,\beta}(x; \beta, H)}{\Theta_{ki}(x; \beta, H)} dN_{ki}(x) - Y_i(x)\Theta_{ki,\beta}(x; \beta, H)dH_x \right] \right\} + \sum_{i=1}^{n} \int_{0}^{\xi} \left\{ \left[ \frac{\tilde{\Theta}_{2i,\beta}(\beta)}{\tilde{\Theta}_{2i}(\beta)} d\tilde{N}_{2i}(x) - \tilde{Y}_i(x)\tilde{\Theta}_{2i,\beta}(\beta)dH_x \right] \right\},
\end{align}

Solutions to the above estimating equations (3.6) and (3.7) set to zero, gives an NPMLE of \( \{d\hat{H}\} \) and \( \hat{\beta} \).

**Remark 2.** The NPMLE is obtained by maximizing the log-likelihood \( \ell \), or equivalently, solving the system of score functions (3.6) and (3.7). Note that the proposed method is semiparametric and thus the dimension of the parameter \( \Omega \) is of the same
order as the sample size. We can write the NPMLE of $H$ based on $U_{H_i}$ as

$$
\hat{H}_t = \sum_{i=1}^n \int_0^t \sum_{k=1}^2 \frac{dN_{ki}(x) + d\tilde{N}_{2i}(x)}{S_0(x; \beta, H)},
$$

where $S_0(t; \beta, H) = \sum_{i=1}^n \left[ \sum_{k=1}^2 Y_i(t) \omega_{ki}(t; \beta, H) \Theta_{ki}(t; \beta, H) + \tilde{Y}_i(t) \tilde{\Theta}_{2i}(\beta) \right]$.

Note that $\hat{H}$ can viewed as a weighted Breslow-type estimator ([10]), and has a close connection to the estimator proposed by [13] for semiparametric transformation models. $\omega_{ki}$ has unit expectation based on martingale properties of score function. Therefore, maximization of the likelihood with respect to $H$ can be done by an iterative reweighting algorithm ([13]).

### 3.5 Asymptotic Properties

We use a combination of empirical process and martingale theory following a general line of [71, 72, 13, 14].

The following regularity conditions are needed:

1. The true $H$ is strictly increasing and differentiable. The true value of parameter vector $\beta$ are in the interior of a compact Euclidean space.

2. With probability one, the covariate process $Z(t)$ is left continuous with total bounded variation (BV) within $[0, \xi]$. Also, $Z(t)$ is linearly independent in the sense that, if there exist $a(t)$ and $c$ such that $a(t) + c^T Z(t) = 0$ with probability one, then $a(t) = 0$ and $c = 0$.

3. With probability one, $P(Y(\xi)|Z(t)) > 0$, $P(\delta_1 = \delta_2 = 0, T = \xi|Z(t)) > 0$, $P(\tilde{Y}(\xi)|Z(t)) > 0$, $P(\delta_1 = 1, \tilde{\delta}_2 = 0, \tilde{T} = \xi|Z(t)) > 0$. e.g., the at risk set $Y(t)$ and $\tilde{Y}(t)$ will not shrink to empty.

4. Hessian matrix $\mathcal{L}_n$ evaluated at the true values of $H$ and $\beta$ is positive definite, and converges to a deterministic matrix.
Based on the score functions (3.6) and (3.7), we can write the score equations for \( \Omega = (\beta, \{dH\}) \) in martingale form

\[
U_\beta = \sum_{i=1}^{n} \int_{0}^{\xi} \left\{ \sum_{k=1}^{2} \frac{\Theta_{ki,\beta}(x; \beta, H)}{\Theta_{ki}(x; \beta, H)} dM_{ki}(x) + \frac{\tilde{\Theta}_{2i,\beta}(\beta)}{\tilde{\Theta}_{2i}(\beta)} d\tilde{M}_{2i}(x) \right\},
\]

and

\[
U_{H_t} = \sum_{i=1}^{n} \int_{0}^{t} \left\{ \sum_{k=1}^{2} dM_{ki}(x) + \int_{x^+}^{\xi} \psi_{ki}(u; \beta, H) dM_{ki}(u) dH_x \right\} + d\tilde{M}_{2i}(x) \right\}.
\]

After exchanging the integrals, we have:

\[
U_{H_t} = \sum_{i=1}^{n} \left\{ \int_{0}^{t} \left[ \sum_{k=1}^{2} dM_{ki}(x) + \int_{0}^{t} \int_{x^+}^{\xi} \psi_{ki}(u; \beta, H) dM_{ki}(u) dH_x \right] + d\tilde{M}_{2i}(x) \right\}
\]

\[
= \sum_{i=1}^{n} \int_{0}^{t} \sum_{k=1}^{2} \left[ 1 + \int_{0}^{u} \psi_{ki}(u; \beta, H) dH_x dM_{ki}(u) \right] + \sum_{i=1}^{n} \int_{0}^{t} d\tilde{M}_{2i}(u)
\]

\[
= \sum_{i=1}^{n} \int_{0}^{\xi} \left[ \sum_{k=1}^{2} \varepsilon_{ki}(u; t; \beta, H) dM_{ki}(u) + I(u \leq t) d\tilde{M}_{2i}(u) \right],
\]

where \( \varepsilon_{ki}(u; t; \beta, H) = I(u \leq t) + \int_{0}^{u \wedge t} \psi_{ki}(u; \beta, H) dH_x \). As we show in the Appendix, the linear transform \( \int_{0}^{\xi} \varepsilon_{ki}(u; t; \beta, H) dM_{ki}(u) \) and \( \int_{0}^{\xi} I(u \leq t) d\tilde{M}_{2i}(x) \) is a martingale when \( \varepsilon_{ki}(u; t; \beta, H) \) and \( I(u \leq t) \) do not depend on \( t \) for \( u \leq t \) as is the case here. Therefore, the score functions \( U_\beta \) and \( U_{H_t} \) are both martingales under the true model.

In the following we present the consistency and weak convergence results for the proposed NPMLE \( \hat{\Omega} = (\hat{\beta}, \hat{H}_t) \) with details given in the Appendix.

**Theorem III.1.** Assuming regularity conditions hold, with probability one, \( \hat{\beta} \) converges to \( \beta^0 \), \( \hat{H}_t \) converges to \( H_t^0 \) uniformly in the interval \([0, \xi]\), where \( H_t^0, \beta^0 \) are the true values of \( H_t, \beta \).
Consider a linear functional

\[
(3.10) \quad n^{1/2} \left\{ \int_0^\xi a^T (\hat{\beta} - \beta^0) + b(t)^T (\hat{H}_t - H_t^0) \right\},
\]

where \( a \) is real vector, \( b(t) \) is a function with bounded total variation in \([0, \xi]\), and let \( B \) be the vector consisting of the values of \( b(t) \) evaluated at the observed failure times corresponding to the set \( \{dH\} \), and \( E^T = (a^T, B^T) \). We have

**Theorem III.2.** Assuming regularity conditions hold, \( n^{1/2} \{\hat{\beta} - \beta^0, \hat{H}_t - H_t^0\} \) converges weakly to a zero-mean Gaussian process. In addition, \( nE^T(I_n)^{-1}E \) converges in probability to the asymptotic variance-covariance function of the linear functional (3.10), where \( I_n \) is the negative Hessian matrix of the observed log-likelihood function with respect to \( \hat{\Omega} \).

For a differentiable functional \( F(\Omega) \) of \( \Omega \), based on the functional delta method ([3] Section II.8), \( n^{1/2} \{\hat{F}(\hat{\Omega}) - F(\Omega)\} \) converges weakly to a zero-mean Gaussian process with variance-covariance function \( \hat{F}(\Omega)^T (I^0)^{-1} \hat{F}(\Omega) \), where \( \hat{F}(\Omega) \) is the gradient of \( F(\Omega) \) with respect to \( \Omega \), \( I^0 \) can be consistently estimated by \( n^{-1} I_n \), with the explicit expression of \( I_n \) is derived in the Appendix.

**3.6 Numerical Examples**

**3.6.1 Simulation Studies**

Data on two event times \( T_1^* \) and \( T_2^* \) are simulated based on the marginal and conditional specifications. We consider the following true models: (i) all nonterminal events are observable (scenario I); or (ii) 30% subjects’ nonterminal events may be missing (scenario II). The simulation results are summarized in Tables 3.1 and 3.2, respectively. Both simulations assume a common baseline hazard \( h(t) = 0.1 \), with two covariates \( Z_1 \) and \( Z_2 \) included in regression models of \( T_1^* \) and \( T_2^* \), where \( Z_1 \) is
generated from a uniform distribution $U(0, 1)$, and $Z_2$ is generated from a binomial distribution with a success probability of 0.5. The true logarithmic hazard ratios for $T_1^*$ and $T_2^*$ are $\beta_{\eta 1} = 0.5, \beta_{\eta 2} = 1, \beta_{\theta 1} = -0.5, \beta_{\theta 2} = 0.5$, and $\beta_{\mu 0} = 1, \beta_{\mu 1} = 0.5, \beta_{\mu 2} = 0$. Under independent right censoring time from a uniform distribution $U(0, 10)$ and scenario I, 31.1% of subjects have both nonterminal and terminal events, 14.9% only have a nonterminal event and censored terminal event, 22.6% have only a terminal event, and 31.5% have neither nonterminal nor terminal events. In scenario II, on average 30% of subjects’ nonterminal events are missing, the proportions with different types of events being 22.1%, 10.9%, 31.2%, and 35.8%, respectively.

Under both scenarios, the simulated datasets were analyzed by the proposed method with and without accounting for missing nonterminal events (Model I and II, respectively), as well as by the Markov illness-death model based on partial likelihood ([36], p.270). The performance of all these methods was assessed under sample sizes of $n = 250$ and $n = 400$. The Markov model assumes a set of baseline hazards, one for each transition intensity, $\lambda_{jk}(t) = \lambda_{j0k}(t) \exp(\beta_{jk}Z_j)$, for $j < k, j = 0, 1, k = 1, 2$.

Compared with the Markov model, Model II is more parsimonious since it assumes proportional baseline hazards, such that (i) $\beta_{\eta}$ and $\beta_{\theta}$ have same interpretations as $\beta_{01}$ and $\beta_{02}$; (ii) $\beta_{\mu 0}$ in our proposed model can be viewed as the log hazard ratio between $\lambda_{120}$ and $\lambda_{020}$; (iii) $\beta_{\mu}$ can be viewed as the interaction between $Z$ and the occurrence of $T_1^*$, and $\beta_{\mu} = \beta_{12} - \beta_{02}$. For comparison purpose, $\beta_{12}$ in the Markov model is expressed in terms of $\beta_{\mu}$ based on the above relationship. The simulation results in Tables 3.1 and 3.2 are based on 1,000 replications.

When the observed data do not contain missing nonterminal events (scenario I, Table 3.1), as expected, we find Model II performs best. Slight small-sample bias of the NPMLE for regression coefficients $\beta$ reduces as sample size increases. The
Table 3.1: Simulation results for a model with two covariates; no missing nonterminal events except by right censoring

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta_{\eta_1} = 0.5$</th>
<th>$\beta_{\eta_2} = 1$</th>
<th>$\beta_{\eta_1} = -0.5$</th>
<th>$\beta_{\eta_2} = 0.5$</th>
<th>$\beta_{\mu_0} = 1$</th>
<th>$\beta_{\mu_1} = 0.5$</th>
<th>$\beta_{\mu_2} = 0$</th>
<th>$p = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 250$, Model I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>0.05</td>
<td>0.04</td>
<td>-0.10</td>
<td>-0.07</td>
<td>-0.03</td>
<td>0.10</td>
<td>0.09</td>
<td>0.05</td>
</tr>
<tr>
<td>ASE</td>
<td>0.30</td>
<td>0.19</td>
<td>0.46</td>
<td>0.30</td>
<td>0.37</td>
<td>0.63</td>
<td>0.40</td>
<td>0.05</td>
</tr>
<tr>
<td>ESE</td>
<td>0.30</td>
<td>0.19</td>
<td>0.45</td>
<td>0.31</td>
<td>0.38</td>
<td>0.63</td>
<td>0.41</td>
<td>0.04</td>
</tr>
<tr>
<td>CP(%)</td>
<td>94.5</td>
<td>94.2</td>
<td>95.2</td>
<td>93.7</td>
<td>95.1</td>
<td>94.2</td>
<td>95.2</td>
<td>96.2</td>
</tr>
<tr>
<td>$n = 250$, Model II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>-0.01</td>
<td>0.01</td>
<td>-0.02</td>
<td>-0.01</td>
<td>-0.01</td>
<td>0.00</td>
<td>0.02</td>
<td>-</td>
</tr>
<tr>
<td>ASE</td>
<td>0.29</td>
<td>0.18</td>
<td>0.39</td>
<td>0.26</td>
<td>0.35</td>
<td>0.57</td>
<td>0.36</td>
<td>-</td>
</tr>
<tr>
<td>ESE</td>
<td>0.31</td>
<td>0.19</td>
<td>0.41</td>
<td>0.27</td>
<td>0.37</td>
<td>0.61</td>
<td>0.37</td>
<td>-</td>
</tr>
<tr>
<td>CP(%)</td>
<td>93.7</td>
<td>93.4</td>
<td>93.7</td>
<td>93.3</td>
<td>94.0</td>
<td>93.5</td>
<td>93.7</td>
<td>-</td>
</tr>
<tr>
<td>$n = 250$, Markov Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>0.02</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.00</td>
<td>-</td>
<td>-0.01</td>
<td>0.01</td>
<td>-</td>
</tr>
<tr>
<td>ASE</td>
<td>0.33</td>
<td>0.20</td>
<td>0.49</td>
<td>0.29</td>
<td>-</td>
<td>0.64</td>
<td>0.38</td>
<td>-</td>
</tr>
<tr>
<td>ESE</td>
<td>0.34</td>
<td>0.20</td>
<td>0.49</td>
<td>0.30</td>
<td>-</td>
<td>0.65</td>
<td>0.40</td>
<td>-</td>
</tr>
<tr>
<td>CP(%)</td>
<td>94.8</td>
<td>94.8</td>
<td>94.6</td>
<td>94.7</td>
<td>-</td>
<td>94.6</td>
<td>94.1</td>
<td>-</td>
</tr>
<tr>
<td>$n = 400$, Model I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>0.05</td>
<td>0.03</td>
<td>-0.09</td>
<td>-0.06</td>
<td>-0.02</td>
<td>0.09</td>
<td>0.07</td>
<td>0.04</td>
</tr>
<tr>
<td>ASE</td>
<td>0.24</td>
<td>0.15</td>
<td>0.36</td>
<td>0.23</td>
<td>0.28</td>
<td>0.49</td>
<td>0.31</td>
<td>0.04</td>
</tr>
<tr>
<td>ESE</td>
<td>0.23</td>
<td>0.15</td>
<td>0.36</td>
<td>0.24</td>
<td>0.28</td>
<td>0.48</td>
<td>0.30</td>
<td>0.04</td>
</tr>
<tr>
<td>CP(%)</td>
<td>94.8</td>
<td>94.4</td>
<td>96.0</td>
<td>94.6</td>
<td>94.6</td>
<td>95.4</td>
<td>95.0</td>
<td>95.2</td>
</tr>
<tr>
<td>$n = 400$, Model II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>-0.01</td>
<td>-0.00</td>
<td>-0.02</td>
<td>-0.01</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>-</td>
</tr>
<tr>
<td>ASE</td>
<td>0.23</td>
<td>0.15</td>
<td>0.31</td>
<td>0.20</td>
<td>0.28</td>
<td>0.45</td>
<td>0.28</td>
<td>-</td>
</tr>
<tr>
<td>ESE</td>
<td>0.24</td>
<td>0.15</td>
<td>0.32</td>
<td>0.21</td>
<td>0.29</td>
<td>0.46</td>
<td>0.29</td>
<td>-</td>
</tr>
<tr>
<td>CP(%)</td>
<td>94.6</td>
<td>94.4</td>
<td>94.2</td>
<td>94.6</td>
<td>94.6</td>
<td>95.5</td>
<td>93.8</td>
<td>94.8</td>
</tr>
<tr>
<td>$n = 400$, Markov Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>-</td>
<td>-0.00</td>
<td>-0.01</td>
<td>-</td>
</tr>
<tr>
<td>ASE</td>
<td>0.26</td>
<td>0.16</td>
<td>0.38</td>
<td>0.22</td>
<td>-</td>
<td>0.50</td>
<td>0.30</td>
<td>-</td>
</tr>
<tr>
<td>ESE</td>
<td>0.26</td>
<td>0.16</td>
<td>0.37</td>
<td>0.23</td>
<td>-</td>
<td>0.48</td>
<td>0.29</td>
<td>-</td>
</tr>
<tr>
<td>CP(%)</td>
<td>94.8</td>
<td>94.2</td>
<td>96.0</td>
<td>94.2</td>
<td>-</td>
<td>96.2</td>
<td>95.9</td>
<td>-</td>
</tr>
</tbody>
</table>

ESE: empirical standard errors based on 1,000 estimates
ASE: average of estimated standard errors
CP: 95% Coverage Probability

estimated mean asymptotic standard errors obtained from the observed information matrix are quite close to the empirical standard errors, and the resultant 95% confidence intervals attain appropriate coverage probabilities. We also find that the method allowing for missing nonterminal events (Model I) also achieves almost unbiased estimates, and maintains the 95% coverage probability at nominal levels, especially with large sample size of $n = 400$. The model with no missing nonterminal events, $p = 0$, in Model I gives a reasonably small bias of 0.04-0.05, considering the
fact that $p = 0$ (logit($p$) = $-\infty$) is on the boundary of parameter space for this model. The estimates from the Markov model, while unbiased, are less efficient than results from Model II, since the more parsimonious model is correctly specified.

Table 3.2: Simulation results for a model with two covariates, missing nonterminal events present

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta_{\eta 1} = 0.5$</th>
<th>$\beta_{\eta 2} = 1$</th>
<th>$\beta_{\eta 1} = -0.5$</th>
<th>$\beta_{\eta 2} = 0.5$</th>
<th>$\beta_{\mu 0} = 1$</th>
<th>$\beta_{\mu 1} = 0.5$</th>
<th>$\beta_{\mu 2} = 0$</th>
<th>$p = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>-0.05</td>
<td>-0.02</td>
<td>0.02</td>
<td>-0.00</td>
<td>0.01</td>
<td>-0.04</td>
<td>0.02</td>
<td>-0.03</td>
</tr>
<tr>
<td>ASE</td>
<td>0.39</td>
<td>0.25</td>
<td>0.61</td>
<td>0.37</td>
<td>0.48</td>
<td>0.84</td>
<td>0.51</td>
<td>0.13</td>
</tr>
<tr>
<td>ESE</td>
<td>0.38</td>
<td>0.25</td>
<td>0.63</td>
<td>0.38</td>
<td>0.49</td>
<td>0.86</td>
<td>0.53</td>
<td>0.13</td>
</tr>
<tr>
<td>CP(%)</td>
<td>93.7</td>
<td>94.8</td>
<td>93.2</td>
<td>94.6</td>
<td>94.1</td>
<td>93.8</td>
<td>94.9</td>
<td>93.1</td>
</tr>
<tr>
<td>$n = 250$, Model I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>-0.34</td>
<td>-0.16</td>
<td>0.49</td>
<td>0.25</td>
<td>0.13</td>
<td>-0.46</td>
<td>-0.21</td>
<td>-</td>
</tr>
<tr>
<td>ASE</td>
<td>0.33</td>
<td>0.21</td>
<td>0.34</td>
<td>0.22</td>
<td>0.41</td>
<td>0.60</td>
<td>0.37</td>
<td>-</td>
</tr>
<tr>
<td>ESE</td>
<td>0.36</td>
<td>0.22</td>
<td>0.34</td>
<td>0.22</td>
<td>0.45</td>
<td>0.62</td>
<td>0.38</td>
<td>-</td>
</tr>
<tr>
<td>CP(%)</td>
<td>81.3</td>
<td>87.0</td>
<td>70.0</td>
<td>80.4</td>
<td>91.9</td>
<td>87.5</td>
<td>89.8</td>
<td>-</td>
</tr>
<tr>
<td>$n = 250$, Model II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>-0.07</td>
<td>-0.10</td>
<td>0.25</td>
<td>0.20</td>
<td>-</td>
<td>-0.23</td>
<td>-0.17</td>
<td>-</td>
</tr>
<tr>
<td>ASE</td>
<td>0.39</td>
<td>0.24</td>
<td>0.40</td>
<td>0.24</td>
<td>-</td>
<td>0.64</td>
<td>0.40</td>
<td>-</td>
</tr>
<tr>
<td>ESE</td>
<td>0.41</td>
<td>0.24</td>
<td>0.40</td>
<td>0.24</td>
<td>-</td>
<td>0.66</td>
<td>0.40</td>
<td>-</td>
</tr>
<tr>
<td>CP(%)</td>
<td>94.8</td>
<td>93.3</td>
<td>91.0</td>
<td>87.4</td>
<td>-</td>
<td>92.6</td>
<td>92.4</td>
<td>-</td>
</tr>
<tr>
<td>$n = 250$, Markov Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>-0.02</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>-0.02</td>
<td>0.00</td>
<td>-0.01</td>
</tr>
<tr>
<td>ASE</td>
<td>0.31</td>
<td>0.21</td>
<td>0.47</td>
<td>0.33</td>
<td>0.36</td>
<td>0.66</td>
<td>0.42</td>
<td>0.11</td>
</tr>
<tr>
<td>ESE</td>
<td>0.30</td>
<td>0.21</td>
<td>0.48</td>
<td>0.33</td>
<td>0.37</td>
<td>0.67</td>
<td>0.42</td>
<td>0.11</td>
</tr>
<tr>
<td>CP(%)</td>
<td>94.5</td>
<td>94.3</td>
<td>94.7</td>
<td>94.9</td>
<td>94.3</td>
<td>94.8</td>
<td>94.8</td>
<td>94.6</td>
</tr>
<tr>
<td>$n = 400$, Model I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>-0.33</td>
<td>-0.18</td>
<td>0.49</td>
<td>0.25</td>
<td>0.15</td>
<td>-0.51</td>
<td>-0.22</td>
<td>-</td>
</tr>
<tr>
<td>ASE</td>
<td>0.26</td>
<td>0.17</td>
<td>0.27</td>
<td>0.18</td>
<td>0.32</td>
<td>0.47</td>
<td>0.29</td>
<td>-</td>
</tr>
<tr>
<td>ESE</td>
<td>0.28</td>
<td>0.18</td>
<td>0.27</td>
<td>0.18</td>
<td>0.33</td>
<td>0.48</td>
<td>0.29</td>
<td>-</td>
</tr>
<tr>
<td>CP(%)</td>
<td>74.5</td>
<td>78.8</td>
<td>54.3</td>
<td>69.6</td>
<td>92.1</td>
<td>80.2</td>
<td>88.7</td>
<td>-</td>
</tr>
<tr>
<td>$n = 400$, Model II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>-0.06</td>
<td>-0.12</td>
<td>0.25</td>
<td>0.20</td>
<td>-</td>
<td>-0.27</td>
<td>-0.19</td>
<td>-</td>
</tr>
<tr>
<td>ASE</td>
<td>0.31</td>
<td>0.19</td>
<td>0.32</td>
<td>0.19</td>
<td>-</td>
<td>0.50</td>
<td>0.31</td>
<td>-</td>
</tr>
<tr>
<td>ESE</td>
<td>0.31</td>
<td>0.19</td>
<td>0.30</td>
<td>0.19</td>
<td>-</td>
<td>0.49</td>
<td>0.30</td>
<td>-</td>
</tr>
<tr>
<td>CP(%)</td>
<td>93.1</td>
<td>88.4</td>
<td>87.3</td>
<td>80.1</td>
<td>-</td>
<td>92.7</td>
<td>91.0</td>
<td>-</td>
</tr>
</tbody>
</table>

- ESE: empirical standard errors based on 1,000 estimates
- ASE: average of estimated standard errors
- CP: 95% Coverage Probability

When missing nonterminal events (30%) are present (scenario II), our proposed method incorporating the missing mechanism (Model I) is superior. As sample size increases, we find diminishing bias of the proposed estimators, good correspondence between asymptotic standard errors and empirical standard errors, and correct 95%
coverage probability at nominal levels. Neither method assuming complete data can match its performance, as expected. The Markov model allowing different baseline hazards for different transition intensities has slightly smaller bias than Model II due to its higher flexibility. Both methods ignoring missing data provide biased estimates and unreliable statistical inference.

3.6.2 Follicular Cell Lymphoma Study

We apply the proposed method to a follicular cell lymphoma dataset collected at the Princess Margaret Hospital, Toronto, between 1967 and 1996 ([53]). Follicular cell lymphoma is a common type of non-Hodgkin lymphoma (NHL), a slow-growing lymphoma arising from B-cells. It mainly affects older adults with few early warning signs, and is usually difficult to cure (unlike the Hodgkin’s lymphoma). It is characterized by a relatively high cancer relapse rate even with very good response to treatment.

In this observational study cohort, 517 patients with early stage disease (I or II) were treated with either radiation alone (RT, 80.0%) or with radiation and chemotherapy (RT+CMT, 20.0%). Patients who had complete treatment response are included in the analysis. The times to cancer relapse and death are available as outcome variables, and are measured in years from the date of cancer diagnosis. Other variables assessed at the time of diagnosis include patients’ age (mean 56.6 years, sd 14.0 years) and haemoglobin level (mean 139.1 g/l, sd 15.3 g/l). The median follow-up time is 8.9 years. Among the 354 (68.5%) and 163 (31.5%) patients with stage I and II cancers, respectively, 301 (85.0%) and 113 (69.3%) were treated by RT alone (p<.0001). Two-sample t-tests find the distributions of patients’ age are significantly different between treatment groups (p=0.032) and clinical stages (p=0.0002), while the distributions of haemoglobin level are similar between treatment groups.
Among the 517 patients, 162 (31.3%) had prior relapse before death, 74 (14.3%) died without a prior recorded relapse, 88 (17.0%) had relapse and were censored afterwards, and 193 (37.3%) were censored for both events. Like in many other cancer studies, it is plausible to expect that patients who died without recorded relapse or were censored for both events may have experienced a relapse that went missing.

Kaplan-Meier plots for survival distributions for times to relapse and death by treatment group are given in Figures 3.3 and 3.4. In Figure 3.3 we (inappropriately) treat death as independent censoring of relapse for descriptive purposes. Note that relapse affecting the risk of death would invalidate this estimator.

We first fit a model for time to relapse and time to death separately: (i) a Cox PH model for time to relapse under independent censoring assumption for death; and (ii) a Cox PH model on time to death with relapse as a time-dependent covariate. The results are summarized in Table 3.3. The baseline covariates included for regression analysis are treatment group, clinical stage, and patients’ age (centered at its sample mean 56.6 years and standardized by its sample standard deviation 14.0 years). We find the effect of patients’ haemoglobin level at baseline is always insignificant for all methods, therefore it is not included in the final model. After adjusting for other covariates, separate analysis finds that stage II, older age patients treated by RT had shorter time to relapse, and those diagnosed at older age also had shorter time to death without relapse. In addition, patients with relapse prior to time \( t \) are expected to have about 10 times (HR=exp(2.33), \( p < .0001 \)) the risk of dying at subsequent times as compared with a patient who has not experienced a relapse, and older patients may live longer than younger patients if both had relapse before
Figure 3.3: Kaplan-Meier plot for time to relapse by treatment group

(p=0.085). This finding suggests a strong positive association between relapse and death, as expected.

Our proposed illness-death models taking account of missing relapse was applied to the follicular cell lymphoma dataset (Table 3.3). We estimated that, among all patients who had experienced cancer relapse, 29.5% (se=3.6%, p<.0001) were not recorded as relapsed. With this important correction to the observed semicompeting risks data, we find that cancer relapse is positively associated with death, as the risk of death for a patients with prior relapse is about 5 times that of a patient without relapse (HR=exp(1.634), p<.0001). Compared with the other models, not
only our model finds similar effects of clinical stage, patient’s age, a stronger effect of treatment on time to relapse, but also concludes that stage I, younger patients receiving RT lived longer if they did not have relapse. Furthermore, a patient with stage I cancer and prior relapse is expected to die sooner (p=0.019). This result is not surprising, as many early-stage cancers become more aggressive and lethal once they relapse. In fact, our alternative analysis of time to death using a Cox PH model with relapse as a time-dependent covariate also suggests a similar but insignificant finding (p=0.470).

One important feature of our proposed joint model for semicompeting risks data
Table 3.3: Results of separate Cox PH analyses, and the proposed model accounting for possibly missing relapse

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Separate Analysis</th>
<th>Proposed Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
</tr>
<tr>
<td><strong>Time to relapse</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT vs RT+CMT</td>
<td>0.641</td>
<td>0.196</td>
</tr>
<tr>
<td>Stage I vs II</td>
<td>-0.463</td>
<td>0.139</td>
</tr>
<tr>
<td>Age</td>
<td>0.269</td>
<td>0.068</td>
</tr>
<tr>
<td><strong>Time to death without relapse</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT vs RT+CMT</td>
<td>0.219</td>
<td>0.351</td>
</tr>
<tr>
<td>Stage I vs II</td>
<td>-0.440</td>
<td>0.277</td>
</tr>
<tr>
<td>Age</td>
<td>1.098</td>
<td>0.147</td>
</tr>
<tr>
<td><strong>Time to death after relapse</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relapse</td>
<td>2.332</td>
<td>0.413</td>
</tr>
<tr>
<td>RT vs RT+CMT</td>
<td>-0.630</td>
<td>0.431</td>
</tr>
<tr>
<td>Stage I vs II</td>
<td>0.235</td>
<td>0.325</td>
</tr>
<tr>
<td>Age</td>
<td>-0.299</td>
<td>0.173</td>
</tr>
<tr>
<td>Prob. of missing relapse</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 3.5: Estimated Kendall’s $\tau$. Model I: accounting for possibly missing relapse; Model II: assuming no missing relapse except due to censoring

is a measure of dependence between times to relapse and death based on covariate-specific Kendall’s $\tau$ (Figure 3.5). The estimated Kendall’s $\tau$ from Model I and Model II show similar trends but are otherwise slightly different. Older patients, except for those stage II patients who received RT+CMT, tend to have stronger dependence
between relapse and death; stage II patients who received RT+CMT seem to have the strongest dependence among all patients.

In summary, our proposed method while in general agreement with separate analyses identified additional significant covariate effects. In particular, our proposed model supports a positive correlation between relapse and death, especially for stage I patients. Based on our simulation results in section 3.6.1, the nontrivial proportion of unobserved patients’ relapse status (29.5%) was estimated. Failing to recognize missing relapses may lead to diluted covariate effects and misleading biased estimates.

3.7 Discussion

We have developed an illness-death model for semicompeting risks that recognizes that an unknown proportion of nonterminal events may be missing at random (left censored) rather than right censored, if not observed prior to death or a censoring event. The inherent difficulty on top of the semicompeting risks problem is that, both the time to nonterminal event and its status may be unobserved in this situation. In addition to offering a general treatment for the missingness of nonterminal event, the proposed illness-death model incorporates positive and negative dependence between nonterminal and terminal events.

Dependent censoring issues were taken into account by the proposed model in a natural way.

Time to non-terminal event may be viewed as a mixture of exact, left and right censored data with partially missing status and dependent censoring. The PH assumption of common baseline hazard allowed a methodology similar to the one used with transformation models avoiding irregular convergence rates of current status data. The same assumption of proportionality between hazards for nonterminal and
terminal events yielded close-form expressions for the model and likelihood contributions facilitating an explicit evaluation of the properties of the model such as measures of local and global dependence. This assumption can be relaxed without changing the modeling principles by treating the baseline hazards as elements of a (non-proportional) transformation model. That is $h_0$ is considered as a transformation of $h_0^*$, invoking additional parameters in the transformation.

The conventional methodology for Markov multistate models, including illness-death models and progressive multistate models, are mainly based on partial likelihood. While these methods are flexible in analyzing multiple outcome data like semicompeting risks and recurrent event data, limited research has been conducted to address the situations when states are not always observable.

One advantage of our proposal is that the score functions retain a martingale structure, such that the martingale central limit theorems can be used to establish weak convergence. Empirical processes (a Glivenko-Cantelli argument) and functional differentiation are used to establish uniform law of large numbers implying consistency.

Relaxing the missing at random assumption does not appear to be difficult if an alternative model for the missing data mechanism is identified.
4.1 Introduction

In advanced cancer studies, the goal of anti-cancer therapeutic interventions is to prolong survival. Progression-related events that occur earlier, such as progression-free survival (PFS) or time to progression (TTP), and disease-free survival (DFS), are thus widely used as a primary or secondary endpoints in oncology clinical trials. All cancer endpoints are driven by the latent natural history of the disease and treatment effects on it. Several issues arise in statistical analysis and planning: (i) the definition of progression-related endpoints relies on some assumptions about the disease natural history; (ii) the surrogacy of progression-related endpoints for OS requires a model of the dependence between progression and death; (iii) a nontrivial proportion of patients’ progression times may be unobserved or missing.

The definition of progression-related endpoints reflects how we view the disease natural history. PFS or DFS is usually defined as the time from randomization to disease progression/recurrence or death ([24]). On the contrary, TTP only considers progression excluding death from the composite endpoint. The association between progression and survival is an important aspect of cancer studies. If the association is positive, an improvement in progression represents a clinical benefit that translates
into an improvement in survival ([24]). In addition, because progression is a precursor to death in cancer, it is natural to consider progression-related endpoints such as PFS or DFS as candidates for early surrogate endpoints for OS ([54, 11]).

Dependent on the perceived relationship between the progression event and death, a number of distinct models have been proposed.

1. *Semicompeting risks*: [21] considers the so-called semicompeting risks problem, where terminal event (death) may censor nonterminal event (progression), but not vice versa. This implies that the disease progression process is not fully reflected in the progression endpoint so that death from the disease may occur before the potential progression event. Examples include deaths due to metastases when the progression event is defined as local recurrence. Therefore, based on the definition of PFS/DFS and TTP, it is clear that semicompeting risks data analysis methods are suitable for joint modeling of TTP and death, but not PFS/DFS and death. The correlation between nonterminal and terminal events has been mainly modeled through copulas ([49]), such as [21, 63] in one-sample setting, and [52, 15] in regression setting. As pointed out by [64], the so-called illness-death model in the multistate model framework ([6, 36]) is another way to represent the semicompeting risks model.

2. *Recurrent event with terminal event*: The same situation could be considered as recurrent event (progression) and a terminal event (death) and treated by the recurrent events methodology [17, 67]. As in the semicompeting risks framework, these methods do not impose any order on progression and death allowing for deaths to occur without prior recorded progression.

3. *Recurrent events*: Another mechanism of progression and death is purely se-
sequential in the sense that it is assumed that the progression event always precedes death. Here both progression and death are considered as recurrent events. A number of methods for gap times in recurrent events setting have been proposed and studied, e.g., [55, 42, 17]. Alternatively, the right order of events can be imposed by a conditioning argument restricting the domain of a working unordered joint distribution of time to progression and death to the upper wedge.

In this paper we consider a situation when the true mechanism is believed to be of the recurrent event type (progression and death are always ordered), but the observed data having the appearance of semi-competing risks in that some death are recorded without progression, the latter representing missing data in such cases. This assumption is plausible in studies of advanced cancer such as in the example considered below where less than 2% of patients died without observed recurrence. We consider a progressive multistate model shown in Figure 4.1. When progression has not been detected or recorded, PFS/DFS is the same as survival. A shared-frailty model ([31, 38, 7, 69]) is used to describe the dependence between time to progression and death beyond the one induced by the ordering assumption. The model can be viewed as a generalization of the recent work by [18], and has a close connection to so-called modulated multistate models for recurrent events ([17]). We study global (Kendall’s τ) and local (cross-ratio function) dependence measures ([16, 51]) implied by the model. A semiparametric model specification is used, and the nonparametric maximum likelihood estimation (NPMLE) is used for statistical inference. We also rely on multistate model and recurrent events methodology [64, 67, 44]. We illustrate the finite-sample properties of proposed estimators through Monte Carlo simulation studies, and apply our method to a phase III clinical trial of adjuvant therapy for
colorectal cancer.

4.2 Notations, Model and Observations

4.2.1 Notation and Model

Let $T_1^*$ be the time to true progression from study entry, $T_2^*$ be the time to death from study entry. We consider a progressive multistate model illustrated in Figure 4.1, and define the joint distribution of $(T_1^*, T_2^*)$ through the following hazard
functions as

$$\lambda_1(t_1|Z, X(t)) = \lim_{\Delta \to 0} P(T_1^* \in [t_1, t_1 + \Delta]|T_1^* \geq t_1, X(t_1), Z)/\Delta = Z h_0(t_1) \mu(X(t_1)),$$

(4.1)

$$\lambda_2(t_2|Z, X(t_2), T_1^* = t_1) = \lim_{\Delta \to 0} P(T_2^* \in [t_2, t_2 + \Delta]|T_2^* \geq t_2, X(t_2), T_1^* = t_1, Z)/\Delta = I(t_2 \geq t_1) Z h_0^*(t_2) \eta(X(t_2)),$$

(4.2)

where $I(\cdot)$ is indicator function, $X$ are covariates, $\mu(X(t)) = e^{\beta_0 X(t)}$, $\eta(X(t)) = e^{\beta_1 X(t)}$, $Z$ is the subject-specific frailty term, and $h_0, h_0^*$ are baseline hazard functions. For brevity, we treat $X$ as time-independent, and keep in mind that it is straightforward to make it external time-dependent ([36] section 6.4). The model (4.1) and (4.2) implies that subjects are not at risk of death before progression $\lambda_2(t_2) = 0$ when $t_2 < t_1$.

With the model targeting advanced cancer studies where progression and death are manifestations of the same disease, it is natural to assume that there is a common hazard function $h(\cdot) = h_0^*(\cdot) = h_0(\cdot)$ that summarizes the pattern of disease development, the proportionality assumption. Interestingly, as we will see later, this assumption implies that the global association (Kendall’s $\tau$) between $T_1^*$ and $T_2^*$ does not depend on the baseline hazard. Also, with $Z$ fixed, the two events (progression and death) could be considered as a realization of a Markov-modulated non-homogeneous Poisson process (MMNHPP) with a non-parametric time transformation $H(t)$, the common cumulative hazard function. MMNHPP emerges when occurrence of an event in a Poisson process modifies the intensity for the subsequent events in a Markov way. In our case, occurrence of the progression event affects the hazard of death by a multiplicative factor $\eta$. This angle shows that the model can be easily extended to multiple progression events modeling stages of cancer progression.
Also, due to this assumption, the asymptotics of the model-based estimators is established by methods similar to the ones used for transformation models in survival analysis.

Following [50], we make the following additional assumptions for the joint model:

1. Conditional on $X, Z$, the distribution of censoring time $V^*$ does not depend on $T_1^*$ or $T_2^*$ (non-informative censoring).

2. Both progression and death processes have continuous intensities, so that the corresponding events cannot happen at the same time, $\Pr(T_1^* = T_2^*) = 0$.

From (4.1) and (4.2) the joint density of $(T_1^*, T_2^*)$ (conditional on frailty $X, Z$) becomes:

$$f(t_1, t_2 | Z, X) = Z^2 I(t_2 \geq t_1) h(t_1) h(t_2) \eta \mu \exp[-ZH(t_1) \mu - Z[H(t_2) - H(t_1)]\eta],$$

where $H(t) = \int_0^t h(x) dx$, $\eta \equiv \eta(X)$, $\mu \equiv \mu(X)$. The corresponding joint distribution of $(T_1^*, T_2^*)$ is

$$S(t_1, t_2 | Z, X) = \int_{t_2}^{\infty} \int_{t_1}^{\infty} f(t_1, t_2 | Z, X)$$

$$= \frac{1}{\mu - \eta} \left\{ \mu \exp[-ZH(t_1) \mu - Z[H(t_2) - H(t_1)]\eta] - \eta \exp[-ZH(t_2)\mu] \right\},$$

where $t_1 \leq t_2$ and $\mu \neq \eta$. In the particular case of singularity $\mu = \eta$, i.e. Markov modulation removed, the joint density and survival function becomes

$$f(t_1, t_2 | Z, X) = Z^2 I(t_2 \geq t_1) h(t_1) h(t_2) \eta \mu \exp[-ZH(t_2)\eta],$$

$$S(t_1, t_2 | Z, X) = \exp[-ZH(t_2)\eta] \left\{ 1 + Z\eta H(t_2) - Z\eta H(t_1) \right\}.$$

Kendall’s $\tau$ ([49]) is a popular global measure to assess the dependence between the bivariate survival data. We can show that the population version of Kendall’s $\tau$
(conditional on covariate $X$) between subject $i$ and $j$ is given by:

$$
\tau = 4 \times \mathbb{E} \left( \int_0^\infty \int_0^y S(x, y|Z, X) f(x, y|Z, X) \, dx \, dy \right) - 1
$$

$$= 4 \times \mathbb{E} \left[ \left( \frac{Z_i}{Z_i + Z_j} \right)^2 \times \left( 1 + \frac{Z_i \eta}{Z_i \mu + Z_j \eta} \right) \right] - 1.
$$

Therefore, the choice of baseline hazard does not influence $\tau$. The local measure of dependence though does depend on both distributions. The crossratio function ([16, 51]) $\gamma(t_1, t_2)$ is defined as follows:

$$
\gamma(t_1, t_2|Z, X) = \frac{S(t_1, t_2|Z, X) f(t_1, t_2|Z, X)}{\{D_1 S(t_1, t_2|Z, X)\} \{D_2 S(t_1, t_2|Z, X)\} } = \frac{\lambda_2(t_2|T_1 = t_1)}{\lambda_2(t_2|T_1 > t_1)} = \frac{\lambda_1(t_1|T_2 = t_2)}{\lambda_1(t_1|T_2 > t_2)}
$$

$$= \begin{cases} 
\exp[-ZH(t_1)\mu-Z[H(t_2)-H(t_1)]\eta]-\frac{\eta}{\mu} \exp[-ZH(t_2)\mu], & \mu \neq \eta, \\
1+E[Z(H(t_2)-H(t_1))]/Z[H(t_2)-H(t_1)], & \mu = \eta,
\end{cases}
$$

$0 \leq t_1 \leq t_2$. Note that $\gamma(t_1, t_2|Z, X) > 1$ when $\mu \geq \eta$, and $\gamma(t_1, t_2|Z, X) < 1$ when $\mu < \eta$. The marginal crossratio function becomes

$$
\gamma(t_1, t_2) = \begin{cases} 
\{L[H(t_1)\mu+[H(t_2)-H(t_1)]\eta]-\frac{\eta}{\mu} L[H(t_2)\mu]\} \{L(1)[H(t_1)\mu+[H(t_2)-H(t_1)]\eta]\} / \{L(2)[H(t_1)\mu+[H(t_2)-H(t_1)]\eta]\}, & \mu \neq \eta, \\
1+E[Z(H(t_2)-H(t_1))]/E[Z(H(t_2)-H(t_1))], & \mu = \eta,
\end{cases}
$$

where $L$ is the Laplace transform of $Z$, $L^{(k)}$ is the $k$th derivative of Laplace transform.

### 4.2.2 Observed Data Structure and Counting Process Notation

Let $V^*$ be the censoring time which is independent of $(T_1^*, T_2^*)$ given $X$; $\xi$ be the maximum follow-up time in the study. Throughout this paper, we distinguish the observed status of $T_1^*$ from whether $T_1^*$ can be observed. The status is a censoring index for $T_1^*$, while the latter is an indicator, $M$, of whether $T_1^*$ is not observable (missing for reason other than censoring). Let $M = 1$ if $T_1^*$ is not observable, and $M = 0$ if $T_1^*$ is observable. Missing at random (MAR, [43]) is assumed for $M$, i.e., $M$
depends only on observed covariates. A logistic regression model for the missingness probability is adopted, i.e.,

\[ \Pr(M = 1|X) = p(X; \alpha) = \frac{\exp(\alpha_0 + \alpha X)}{1 + \exp(\alpha_0 + \alpha X)}, \]

while other link functions could also be used as needed. Note that 1) observing \( T_1^* \) (progression) implies \( M = 0 \); 2) observing \( T_2^* \) but not \( T_1^* \) (death without recorded progression) implies \( M = 1 \); 3) observing neither \( T_1^* \) nor \( T_2^* \) (censored both progression and death) could be when \( M = 0 \) or \( M = 1 \).

Suppose \((T_{1i}^*, T_{2i}^*, V_i^*, X_i), i = 1, \ldots, n\) are \( n \) independent and identically distributed replicates of \((T_1^*, T_2^*, V^*, X)\), such that the observed data for subject \( i, i = 1, \ldots, n\), consists of \( \{T_i, \tilde{T}_i, \delta_{ki}, \tilde{\delta}_{2i}, X_i; k = 1, 2, 0 \leq t \leq \xi\} \), where \( T_i = \min(T_1^*, V)I(M = 0) + \min(T_2^*, V)I(M = 1) \), \( \tilde{T}_i = \min(T_2^*, V) \) with \( V = \min(V^*, \xi) \), \( \delta_{ki} = I(T_i = T_{ki}^*) \), \( k = 1, 2 \), \( \tilde{\delta}_{2i} = \delta_{1i}I(\tilde{T}_i = T_{2i}^*) \), where \( I(\cdot) \) is the indicator function, \( \xi \) is the maximum follow-up time. Notice that when \( \Pr(M = 1) = 0 \) our model assumes \( \Pr(\delta_2 = 1) = 0 \), however, the existence of deaths without recorded progression in the data implies \( \Pr(M = 1) > 0 \) and leads to \( \Pr(\delta_2 = 1) > 0 \). According to the definitions above we have that \( \delta_{1i} + \delta_{2i} \leq 1 \) as both are event indicators for the first-occurring event, and \( \tilde{\delta}_{2i} = 1 \) only if \( \delta_{1i} = 1 \). In other words, the observed data structure is the same as with semicompeting risks data ([21]), although the model behind the data is fundamentally different.

In counting process notation, for subject \( i \), let \( N_{ki}^* = I(T_{ki}^* \leq t), k = 1, 2 \), be the underlying counting processes for progression \((k = 1)\) and death \((k = 2)\) events, \( Y_i(t) = I(T_i \geq t) \) be the at-risk process for the observed first-occurring event regardless of event type, \( \tilde{Y}_i(t) = I(\tilde{T}_i \geq t > T_i) \) be the at-risk process for the death following observed progression event, \( N_{ki}(t) = \int_0^t Y_i(x) dN_k^*(x) = \delta_{ki} N_k^*(t), (k = 1, 2) \), be the observed counting processes of progression \((k = 1)\) or death \((k = 2)\) for the
observed first-occurring event, and \( \tilde{N}_{2i} = \delta_1 \delta_2 I(T_{2i}^* \leq t) \) be the observed counting process of death following observed progression event.

Based on the observed data \( \{T_i, \tilde{T}_i, \delta_{1i}, \delta_{2i}, \tilde{\delta}_{2i}, Z_i; i = 1, \ldots, n\} \), we thus observe three counting processes \( N_{1i}(\cdot), N_{2i}(\cdot), \tilde{N}_{2i}(\cdot) \), and two corresponding at-risk processes \( Y_i(\cdot), \tilde{Y}_i(\cdot) \) for subject \( i \). Note that the counting processes \( N_{1i}(\cdot) \) and \( N_{2i}(\cdot) \) record the time and type of subject \( i \)'s first event, and \( \tilde{N}_{2i}(\cdot) \) records the time of death following observed progression. Furthermore, \( \tilde{N}_{2i} \equiv 0 \) if second event could not be observed for subject \( i \), either because of missing progression or a censoring event.

4.3 Maximum Likelihood Estimation with Missing Progression

Let \( \Omega \) denote the full parameter set \( (\beta, \alpha, H) \) that involves a large number of parameters making direct maximization of the likelihood unattractive. We use Expectation-Maximization (EM) algorithm treating the missingness indicator \( M \) and the frailty \( Z \) as the missing data. To deal with the still high-dimensional nonlinear maximization over \( H(\cdot) \) at the M-Step we adopt the artificial mixture method of [61] for semiparametric transformation models.

4.3.1 Complete Data Likelihood with \( M \) and \( Z \) Observed

Pretending \( M \) and \( Z \) are observed, the complete data consist of \( (T_i, \tilde{T}_i, \delta_{1i}, \delta_{2i}, \tilde{\delta}_{2i}, X_i, M_i, Z_i, i = 1, \ldots, n) \). Each subject's likelihood contribution takes one of the following forms:

1. if \( T_i^* \) and \( T_{2i}^* \) both are observed, i.e., \( T_i < \tilde{T}_i, \delta_{1i} = 1, \delta_{2i} = 0, \tilde{\delta}_{2i} = 1, M_i = 0 \),

\[
\mathcal{L}_{1i}^c(\beta, H; M, Z) = f(T_i, \tilde{T}_i|Z_i)
= Z_i^2 h(T_i)h(\tilde{T}_i) \eta \mu \exp \left\{ -Z_i H(T_i) \mu_i - Z_i [H(\tilde{T}_i) - H(T_i)] \eta_i \right\}.
\]
2. if $T_1^*$ is observed and $T_2^*$ is censored, i.e., $T_i < \tilde{T}_i$, $\delta_{1i} = 1$, $\delta_{2i} = 0$, $\tilde{\delta}_{2i} = 0$, $M_i = 0$,

$$\mathcal{L}_{4i}^c(\beta; H; M; Z) = \int_{\tilde{T}_i}^{\infty} f(T_i; y|Z_i)dy$$

$$= Z_i h(\tilde{T}_i)\mu_i \exp \left\{ -Z_i H(\tilde{T}_i)\mu_i - Z_i[H(\tilde{T}_i) - H(T_i)]\eta_i \right\} .$$

3. if $T_1^*$ and $T_2^*$ are both censored, and $T_1^*$ is observable, i.e., $T_i = \tilde{T}_i$, $\delta_{1i} = 0$, $\delta_{2i} = 0$, $\tilde{\delta}_{2i} = 0$, $M_i = 0$,

$$\mathcal{L}_{3i}^c(\beta; H; M; Z) = \int_{\tilde{T}_i}^{\infty} \int_{\tilde{T}_i}^{\infty} f(x, y|Z_i)dxdy$$

$$= \exp[-Z_i H(T_i)\mu_i] .$$

4. if $T_2^*$ is observed and $T_1^*$ is unknown, i.e., $T_i = \tilde{T}_i$, $\delta_{1i} = 0$, $\delta_{2i} = 1$, $\tilde{\delta}_{2i} = 0$, $M_i = 1$,

$$\mathcal{L}_{4i}^c(\beta; H; M; Z) = \int_{0}^{T_i} f(x, \tilde{T}_i|Z_i)dx$$

$$= \begin{cases} 
Z_i h(\tilde{T}_i) \frac{m_\mu}{\mu - \eta} \left\{ \exp[-Z_i H(\tilde{T}_i)\eta_i] - \exp[-Z_i H(\tilde{T}_i)\mu_i] \right\} , & \mu_i \neq \eta_i, \\
Z_i^2 h(\tilde{T}_i) H(\tilde{T}_i)\eta_i\mu_i \exp[-Z_i H(\tilde{T}_i)\eta_i], & \mu_i = \eta_i 
\end{cases} .$$

5. if $T_1^*$ and $T_2^*$ are both censored, $T_1^*$ is unobservable and is not known whether it occurred or not, i.e., $T_i = \tilde{T}_i$, $\delta_{1i} = 0$, $\delta_{2i} = 0$, $\tilde{\delta}_{2i} = 0$, $M_i = 1$,

$$\mathcal{L}_{5i}^c(\beta; H; M; Z) = \int_{\tilde{T}_i}^{\infty} \int_{\tilde{T}_i}^{\infty} f(x, y|Z_i)dxdy + \int_{\tilde{T}_i}^{\infty} \int_{0}^{T_i} f(x, y|Z_i)dxdy$$

$$= \begin{cases} 
\frac{1}{\mu - \eta_i} \left\{ \mu_i \exp[-Z_i H(T_i)\eta_i] - \eta_i \exp[-Z_i H(T_i)\mu_i] \right\} , & \mu_i \neq \eta_i \\
\exp[-Z_i H(T_i)\eta_i][1 + Z_i H(T_i)\eta_i], & \mu_i = \eta_i 
\end{cases} .$$

The complete data log-likelihood for subject $i$ is

$$\ell_i^c(\beta; H; M; Z) = (1 - M_i) \left[ \delta_{1i}\tilde{\delta}_{2i} \log(\mathcal{L}_{1i}^c) + \delta_{1i}(1 - \tilde{\delta}_{2i}) \log(\mathcal{L}_{2i}^c) + (1 - \delta_{1i} - \delta_{2i}) \log(\mathcal{L}_{3i}^c) \right]$$

$$+ M_i \left[ \tilde{\delta}_{2i} \log(\mathcal{L}_{4i}^c) + (1 - \delta_{1i} - \delta_{2i}) \log(\mathcal{L}_{5i}^c) \right]$$

$$+ \{ M_i \log p(X_i; \alpha) + (1 - M_i) \log[1 - p(X_i; \alpha)] \} + \log f(Z_i) ,$$

(4.3)
where \( f(Z_i) \) denotes the density of frailty. Notice that \( \delta_{1i} = 1 \) implies \( M_i = 0 \), and \( \delta_{2i} = 1 \) implies \( M_i = 1 \). The complete data log-likelihood is thus

\[
\ell^c(\beta, H; M, Z) = \sum_{i=1}^{n} \ell^c_i(\beta, H; M, Z).
\]

Notice that if \( T^*_1 \) is always observable, the complete data likelihood only consists of \( \mathcal{L}^c_1, \mathcal{L}^c_2, \mathcal{L}^c_3 \).

4.3.2 Alternative Expression of Complete Data Likelihood

When complete data \((T_i, \tilde{T}_i, \delta_{1i}, \delta_{2i}, \tilde{\delta}_{2i}, X_i, M_i, Z_i), i = 1, \cdots, n\) are available, with the counting process notation introduced in section 4.2.2, we may construct the following martingales with respect to filtration \( \mathcal{F}_{-t} = \sigma\{N_{1i}(s), N_{2i}(s), \tilde{N}_{2i}(s), Y_i(s), \tilde{Y}_i(s), M_i, Z_i, X_i : u \in [0, t), i = 1, \cdots, n\} \)

\[
dM_{1i}(t) = dN_{1i}(t) - U_{1i}dH(t),
\]

\[
d\tilde{M}_{2i}(t) = d\tilde{N}_{2i}(t) - U_{2i}dH(t),
\]

\[
dM_{2i}(t) = dN_{2i}(t) - U_{3i}dH(t),
\]

where

\[
U_{1i} = \Pr(dN_{1i}(t) = 1|Y_i(t) = 1, M_i = 0) = Z_i\mu_i
\]

\[
U_{2i} = \Pr(d\tilde{N}_{2i}(t) = 1|\tilde{Y}_i(t) = 1, M_i = 0) = Z_i\eta_i
\]

\[
U_{3i}(t) = \Pr(dN_{2i}(t) = 1|Y_i(t) = 1, M_i = 1) = \begin{cases} 
Z_i \mu_i \eta_i \exp[-Z_iH(t)\eta_i]-\mu_i \eta_i \exp[-Z_iH(t)\mu_i] \mu_i \neq \eta_i, \\
Z_i \frac{Z_i H(t) \mu_i}{1 + Z_i H(t)}, \quad \mu_i = \eta_i
\end{cases}
\]

Therefore, we can rewrite the individual level complete-data likelihood contribu-
The complete data likelihood is thus rewritten as

\[
L_1^c = h(T_i)U_{1i}h(\tilde{T}_i)U_{2i}\exp\left[-\int_0^\xi Y_i(s)U_{1i}dH(s)\right]\exp\left[-\int_0^\xi \tilde{Y}_i(s)U_{2i}dH(s)\right],
\]

\[
L_2^c = h(T_i)U_{1i}\exp\left[-\int_0^\xi Y_i(s)U_{1i}dH(s)\right]\exp\left[-\int_0^\xi \tilde{Y}_i(s)U_{2i}dH(s)\right],
\]

\[
L_3^c = \exp\left[-\int_0^\xi Y_i(s)U_{1i}dH(s)\right],
\]

\[
L_4^c = h(T_i)U_{3i}(T_i)\exp\left[-\int_0^\xi Y_i(s)U_{3i}(s)dH(s)\right],
\]

\[
L_5^c = \exp\left[-\int_0^\xi Y_i(s)U_{3i}(s)dH(s)\right].
\]

We can view \(U_1, U_2, U_3\) as derived missing data that depend on \(Z\). Notice that \(U_1\) and \(U_2\) are time-invariant while \(U_3(t)\) is time-dependent.

**4.3.3 EM Algorithm**

The E-step is carried out as following: First, recall that \(\delta_{1i} = 1\) implies \(M_i = 0\), \(\delta_{2i} = 1\) implies \(M_i = 1\), and notice

\[
\mathbb{E}\left\{(1 - M_i)U_{1i}|\delta_{1i}, \delta_{2i}, T_i, \tilde{T}_i\right\} = \mathbb{E}\left\{(1 - M_i)\mathbb{E}\left(U_{1i}|\delta_{1i}, \delta_{2i}, T_i, \tilde{T}_i, M_i = 0\right)\right\}
\]

\[
= \left[1 - \mathbb{E}(M_i|\delta_{1i}, \delta_{2i}, T_i, \tilde{T}_i)\right]\mathbb{E}\left(U_{1i}|\delta_{1i} = 0, T_i = \tilde{T}_i, M_i = 0\right),
\]

\[
\mathbb{E}\left\{M_iU_{3i}|\delta_{1i}, \delta_{2i}, T_i, \tilde{T}_i\right\} = \mathbb{E}\left\{M_i\mathbb{E}\left(U_{3i}|\delta_{1i}, \delta_{2i}, T_i, \tilde{T}_i, M_i = 1\right)\right\}
\]

\[
= \mathbb{E}(M_i|\delta_{1i}, \delta_{2i}, T_i, \tilde{T}_i)\mathbb{E}\left(U_{3i}|\delta_{1i}, \delta_{2i}, T_i, \tilde{T}_i, M_i = 1\right).
\]
Second, denote

\[ \Theta^{c_1+c_2}_{1i} = E(U_{1i}|\delta_{1i} = c_1, \tilde{\delta}_{2i} = c_2, T_i, \tilde{T}_i, M_i = 0), \quad (c_1 \geq c_2) \]

\[ \Theta^{c}_{2i} = E(U_{2i}|\delta_{1i} = 1, \tilde{\delta}_{2i} = c, T_i, \tilde{T}_i, M_i = 0), \]

\[ \Theta^{0}_{3i} = E(U_{3i}|\delta_{1i} = 0, \tilde{\delta}_{2i} = c, T_i, \tilde{T}_i, M_i = 1). \]

It can be shown that

\[ \Theta^{c_1+c_2}_{1i} = \mu_i E(Z_i|\delta_{1i} = c_1, \tilde{\delta}_{2i} = c_2, T_i, \tilde{T}_i, M_i = 0) = \frac{\mathcal{L}^{(c_1+c_2+1)} E(T_i)^{\mu_i} + [H(T_i) - H(T_i)]^\eta_i}{\mathcal{L}^{(c_1+c_2)} E(T_i)^{\mu_i} + [H(T_i) - H(T_i)]^\eta_i}, \]

\[ \Theta^{c}_{2i} = \eta_i E(Z_i|\delta_{1i} = 1, \tilde{\delta}_{1i} = c, T_i, \tilde{T}_i, M_i = 0) = \frac{\mathcal{L}^{(c+2)} E(T_i)^{\mu_i} + [H(T_i) - H(T_i)]^\eta_i}{\mathcal{L}^{(c+1)} E(T_i)^{\mu_i} + [H(T_i) - H(T_i)]^\eta_i}, \]

\[ \Theta^{0}_{3i} = E \left\{ Z_i(\eta_i \exp [-Z_i H(t) \eta_i] - \mu_i \exp [-Z_i H(t) \mu_i])|\delta_{1i} = 0, \tilde{\delta}_{2i} = 1, T_i, \tilde{T}_i, M_i = 1 \right\} \]

\[ = \frac{\eta_i \mathcal{L}^{(2)}[H(T_i) \eta_i] - \mu_i \mathcal{L}^{(2)}[H(T_i) \mu_i]}{\mathcal{L}^{(1)}[H(T_i) \eta_i] - \mathcal{L}^{(1)}[H(T_i) \mu_i]}, \]

\[ \Theta^{0}_{3i} = E \left\{ Z_i(\mu_i \eta_i \exp [-Z_i H(t) \eta_i] - \mu_i \eta_i \exp [-Z_i H(t) \mu_i])|\delta_{1i} = 0, \tilde{\delta}_{2i} = 0, T_i, \tilde{T}_i, M_i = 1 \right\} \]

\[ = \frac{\mu_i \eta_i \mathcal{L}^{(1)}[H(T_i) \eta_i] - \mu_i \eta_i \mathcal{L}^{(1)}[H(T_i) \mu_i]}{\mu_i \eta_i \mathcal{L}[H(T_i) \eta_i] - \eta_i \mathcal{L}[H(T_i) \mu_i]}, \]

where \( \mathcal{L} \) is the Laplace transform of \( Z \), \( \mathcal{L}^{(k)} \) is the \( k \)th derivative of Laplace transform.

Third, based on censored observations, we have

\[ \pi_i = E_{M}(M_i|\delta_{1i} = 0, \tilde{\delta}_{2i} = 0, T_i = \tilde{T}_i) = \frac{p(X_i; \alpha) G_2(T_i)}{[1 - p(X_i; \alpha)] G_1(T_i) + p(X_i; \alpha) G_2(T_i)}, \]

where

\[ G_1(T_i) = E \left\{ \mathcal{L}^{5}_{3i}|\delta_{1i}, \tilde{\delta}_{2i}, T_i, \tilde{T}_i, M_i = 0 \right\} = E \left\{ \exp[- \int_0^\xi Y_i(s) U_{1i} dH(s)] \right\}, \]

\[ G_2(T_i) = E \left\{ \mathcal{L}^{5}_{3i}|\delta_{1i}, \tilde{\delta}_{2i}, T_i, \tilde{T}_i, M_i = 1 \right\} = E \left\{ \exp[- \int_0^\xi Y_i(s) U_{3i} dH(s)] \right\}. \]

In the M-step, after plugging in \( \Theta_{1i}, \Theta_{2i}, \Theta_{3i} \) and \( \pi_i \), the target function becomes
\[ \ell_{M1}(\beta, H) + \ell_{M2}(\alpha), \text{ where} \]
\[
\ell_{M1} = \sum_{i=1}^{n} (1 - \pi_i) \left[ \delta_{1i} \log h(T_i) - \int_{0}^{\xi} Y_i(s) \Theta_{1i} dH(s) + \delta_{1i} \tilde{\delta}_{2i} \log h(\tilde{T}_i) - \delta_{1i} \int_{0}^{\xi} \tilde{Y}_i(s) U_{2i} dH(s) \right]
\]
\[
+ \sum_{i=1}^{n} \pi_i \left[ \delta_{2i} \log h(\tilde{T}_i) - \int_{0}^{\xi} Y_i(s) \Theta_{3i} dH(s) \right],
\]
\[
\ell_{M2} = \sum_{i=1}^{n} \left\{ \pi_i \log p(X_i; \alpha) + (1 - \pi_i) \log[1 - p(X_i; \alpha)] \right\},
\]
where \( \ell_{M1} \) and \( \ell_{M2} \) can be maximized separately as their parameter space are distinct.

Let \( t_* \) be some observed event time, \( dH_* \equiv dH(t_*) \) be the jump size of the step function \( H(\cdot) \), we also have a Breslow-type estimate for \( H \) when maximizing \( \ell_{M1} \), i.e.,
\[
d\hat{H}_* = \sum_{i=1}^{n} \frac{dN_{1i}(t_*) + dN_{2i}(t_*) + d\tilde{N}_{2i}(t_*)}{\sum_{i=1}^{n} Y_i(t_*)[(1 - \pi_i) \Theta^\delta_{1i} + \pi_i \Theta^\delta_{2i}] + \tilde{Y}_i(t_*)'(1 - \pi_i) \Theta^\delta_{2i}}.
\]

In the gamma frailty case, such that \( Z_i \) is i.i.d. with mean 1 and variance \( \theta \), i.e.,
\[
f(Z) = \frac{Z^\theta - 1 \exp(-Z/\theta)}{\Gamma(1/\theta) \theta^{1/\theta}},
\]
we have
\[
\Theta^c_{1i} = \mu_i \left[ \frac{1}{\theta} + H(T_i) \mu_i + [H(T_i) - H(T_i)] \eta_i \right], \quad (c = 0, 1, 2)
\]
\[
\Theta^c_{2i} = \eta_i \left[ \frac{1}{\theta} + H(T_i) \mu_i + [H(T_i) - H(T_i)] \eta_i \right], \quad (c = 1, 2)
\]
\[
\Theta^0_{3i} = \left( 1 + \frac{\eta_i [1 + H(T_i) \eta_i]^{-\frac{1}{\theta} - 2} - \mu_i [1 + H(T_i) \mu_i]^{-\frac{1}{\theta} - 2}}{[1 + H(T_i) \eta_i]^{-\frac{1}{\theta} - 1} - \frac{1}{\theta} + H(T_i) \mu_i [1 + H(T_i) \mu_i]^{-\frac{1}{\theta} - 1}} \right),
\]
\[
\Theta^1_{3i} = \left( 1 + \frac{\eta_i [1 + H(T_i) \eta_i]^{-\frac{1}{\theta} - 1} - \eta_i \mu_i [1 + H(T_i) \mu_i]^{-\frac{1}{\theta} - 1}}{\mu_i [1 + H(T_i) \eta_i]^{-\frac{1}{\theta} - 1} - \frac{1}{\theta} + H(T_i) \mu_i [1 + H(T_i) \mu_i]^{-\frac{1}{\theta} - 1}} \right).
\]

4.4 Asymptotic Properties

4.4.1 Observed Data Likelihood

When only the observed data \((T_i, \tilde{T}_i, \delta_{1i}, \delta_{2i}, \tilde{\delta}_{2i}, X_i), i = 1, \cdots, n\) are available, recall that observing \( T_{1i}^* \) implies \( M = 0 \), and observing \( T_{2i}^* \) without \( T_{1i}^* \) implies \( M = 1 \),
so that the likelihood contribution of subject $i$ is:

- If $T_1^*$ and $T_2^*$ both are observed, i.e., $(T_i < \tilde{T}_i, \delta_{1i} = 1, \delta_{2i} = 0, \tilde{\delta}_{2i} = 1)$:

  $$
  \mathcal{L}_{1i}^o(\Omega) = (1 - p) \times f(T_i, \tilde{T}_i) = (1 - p) \times E[\mathcal{L}_{1i}^o(\beta, H; M, Z)].
  $$

- If $T_1^*$ is observed and $T_2^*$ is censored, i.e., $T_i < \tilde{T}_i, \delta_{1i} = 1, \delta_{2i} = 0, \tilde{\delta}_{2i} = 0$,

  $$
  \mathcal{L}_{2i}^o(\Omega) = (1 - p) \times \int_{\tilde{T}_i}^{\infty} f(T_i, y)dy = (1 - p) \times E[\mathcal{L}_{2i}^o(\beta, H; M, Z)].
  $$

- If $T_2^*$ is observed and $T_1^*$ is unknown, i.e., $T_i = \tilde{T}_i, \delta_{1i} = 1, \delta_{2i} = 1, \tilde{\delta}_{2i} = 0$,

  $$
  \mathcal{L}_{3i}^o(\Omega) = p \times \int_{0}^{\tilde{T}_i} f(x, \tilde{T}_i)dx = p \times E[\mathcal{L}_{3i}^o(\beta, H; M, Z)].
  $$

- If $T_1^*$ and $T_2^*$ are both censored, $T_1^*$ is not known whether it occurred or not, i.e., $T_i = \tilde{T}_i, \delta_{1i} = 0, \delta_{2i} = 0, \tilde{\delta}_{2i} = 0$,

  $$
  \mathcal{L}_{4i}^o(\Omega) = \int_{\tilde{T}_i}^{\infty} \int_{T_i}^{\infty} f(x, y)dxdy + p \times \int_{\tilde{T}_i}^{\infty} \int_{0}^{T_i} f(x, y)dxdy
  = (1 - p) \times E[\mathcal{L}_{4i}^o(\beta, H; M, Z)] + p \times E[\mathcal{L}_{5i}^o(\beta, H; M, Z)].
  $$

Note that we can alternatively express $\mathcal{L}_{1i}^o$ and $\mathcal{L}_{2i}^o$ as

$$
\mathcal{L}_{1i}^o(\Omega) = \mathcal{L}_{5i}^o(\Omega) \times \mathcal{L}_{6i}^o(\Omega), 
\mathcal{L}_{2i}^o(\Omega) = \mathcal{L}_{5i}^o(\Omega) \times \mathcal{L}_{7i}^o(\Omega)
$$

where

$$
\mathcal{L}_{5i}^o(\Omega) = (1 - p)f_1(T_i), 
\mathcal{L}_{6i}^o(\Omega) = f_2(\tilde{T}_i|T_i), 
\mathcal{L}_{7i}^o(\Omega) = S_2(\tilde{T}_i|T_i),
$$

such that the contribution from nonterminal event is separated from terminal event.

Therefore, the contribution of subject $i$ in the log-likelihood is

$$
\ell_i^o = \delta_{1i} \tilde{\delta}_{2i} \log(\mathcal{L}_{1i}^o) + \delta_{1i}(1 - \tilde{\delta}_{2i}) \log(\mathcal{L}_{2i}^o) + \delta_{2i} \log(\mathcal{L}_{3i}^o) + (1 - \delta_{1i} - \delta_{2i}) \log(\mathcal{L}_{4i}^o)
= \{ \delta_{1i} \log(\mathcal{L}_{5i}^o) + \delta_{2i} \log(\mathcal{L}_{6i}^o) + (1 - \delta_{1i} - \delta_{2i}) \log(\mathcal{L}_{7i}^o) \} + \delta_{1i} \{ \tilde{\delta}_{2i} \log(\mathcal{L}_{6i}^o) \} + (1 - \tilde{\delta}_{2i}) \log(\mathcal{L}_{7i}^o).
$$

(4.7)
Using the notation $\ell^{o}_{a,i} = \delta_{1i} \log(L^{o}_{5i}) + \delta_{2i} \log(L^{o}_{3i}) + (1 - \delta_{1i} - \delta_{2i}) \log(L^{o}_{4i})$, $\ell^{o}_{b,i} = \delta_{1i} \{\tilde{\delta}_{2i} \log(L^{o}_{6i}) + (1 - \tilde{\delta}_{2i}) \log(L^{o}_{7i})\}$, and $\ell^{o}_{a} = \sum_{i=1}^{n} \ell^{o}_{a,i}$, $\ell^{o}_{b} = \sum_{i=1}^{n} \ell^{o}_{b,i}$, we can write the full observed log-likelihood as

$$\ell^{o} = \sum_{i=1}^{n} \ell^{o}_{i} = \sum_{i=1}^{n} \{\ell^{o}_{a,i} + \ell^{o}_{b,i}\} = \ell^{o}_{a} + \ell^{o}_{b}.$$  

(4.8)

Following [15] we can partition the likelihood contribution of observed data from subject $i$ into two parts, $\ell^{o}_{a,i}$ and $\ell^{o}_{b,i}$, where $\ell^{o}_{a,i}$ is based on information in the time period $0 \leq t \leq T_{i}$, with $T_{i}$ being the time of observed first-occurring event for subject $i$, while $\ell^{o}_{b,i}$ is based on the extended information in the time period $T_{i} \leq t \leq \tilde{T}_{i}$, with $\tilde{T}_{i}$ being the time to death following progression, if any. When the observed first-occurring event is death or censoring, such that no extended observation is available, $\ell^{o}_{b,i} = 0$ by definition. Furthermore, $\ell^{o}_{a}$ is the log-likelihood based on observed competing risks data where only the first-occurring event is available, and $\ell^{o}_{b}$ corresponds to the additional information.

### 4.4.2 Martingale Structure

We can construct the following martingales based on observed counting processes with respect to filtration $\mathcal{F}^{o}_{t_{-}} = \sigma\{N_{1i}(x), N_{2i}(x), \tilde{N}_{2i}(x), Y_{i}(x), \tilde{Y}_{i}(x), Z_{i} : x \in [0, t), i = 1, \ldots, n\}$ as

$$dM_{ki}(t) = dN_{ki}(t) - \Upsilon_{ki}(t; \Omega)dH_{t}, k = 1, 2,$$

$$d\tilde{M}_{2i}(t) = d\tilde{N}_{2i}(t) - \tilde{\Upsilon}_{2i}(t; \Omega)dH_{t},$$

where

$$\Upsilon_{1i}(t; \Omega) = \frac{L^{o}_{5i}}{L^{o}_{4i}}; \ Upsilon_{2i}(t; \Omega) = \frac{L^{o}_{3i}}{L^{o}_{4i}}; \ \tilde{\Upsilon}_{2i}(t; \Omega) = \frac{L^{o}_{6i}}{L^{o}_{7i}}.$$

Note that we can express $L^{o}_{4}$ as

$$L^{o}_{4} = \phi_{i}(t; \Omega) = \exp \left\{ - \int_{0}^{t} \sum_{k=1}^{2} Y_{i}(x) \Upsilon_{ki}(x; \Omega)dH_{x} \right\}.$$
Therefore, the log-likelihood \( \ell^o = \ell^o_a + \ell^o_b \) can be re-expressed in counting process notation as

\[
\ell^o_a = \sum_{i=1}^{n} \sum_{k=1}^{2} \left\{ \int_{0}^{\xi} [\log dH_x + \log \Upsilon_{ki}(x; \Omega)] dN_{ki}(x) - Y_i(x) \Upsilon_{ki}(x; \Omega) dH_x \right\}
\]

\[
\ell^o_b = \sum_{i=1}^{n} \left\{ \int_{0}^{\xi} [\log dH_x + \log \bar{\Upsilon}_{2i}(t; \Omega)] d\bar{N}_{2i}(x) - \bar{Y}_i(x) \bar{\Upsilon}_{2i}(t; \Omega) dH_x \right\}
\]

(4.9) \hspace{1cm} (4.10)

Denote \( \gamma^T = (\beta^T, \alpha^T) \). It can be shown that the score functions for \( \Omega = (\gamma, \{dH\}) \)
can be written in martingale form

\[
U_\gamma = \sum_{i=1}^{n} \int_{0}^{\xi} \left\{ \sum_{k=1}^{2} \frac{\Upsilon_{ki,\beta}(x; \Omega)}{\Upsilon_{ki}(x; \Omega)} dM_{ki}(x) + \frac{\bar{\Upsilon}_{2i,\beta}(t; \Omega)}{\bar{\Upsilon}_{2i}(t; \Omega)} d\bar{M}_{2i}(x) \right\},
\]

(4.11)

and

\[
U_{H_t} = \sum_{i=1}^{n} \sum_{k=1}^{2} \int_{0}^{t} \left[ dM_{ki}(x) + \int_{x^+}^{\xi} \nu_{ki}(u; \Omega) dM_{ki}(u) dH_x \right] + \sum_{i=1}^{n} \int_{0}^{t} \left[ d\bar{M}_{2i}(x) + \int_{x^+}^{\xi} \bar{\nu}_{2i}(u; \Omega) d\bar{M}_{2i}(u) dH_x \right],
\]

where

\[
\omega_{ki}(t^*; \Omega) = 1 - \int_{t^*}^{\xi} \psi_{ki}(x; \Omega) dM_{ki}(x); \quad \psi_{ki}(t; \Omega) = \Upsilon_{ki,H}(t; \Omega)/\Upsilon_{ki}(t; \Omega);
\]

\[
\bar{\omega}_{2i}(t^*; \Omega) = 1 - \int_{t^*}^{\xi} \bar{\psi}_{2i}(x; \Omega) d\bar{M}_{2i}(x); \quad \bar{\psi}_{2i}(t; \Omega) = \bar{\Upsilon}_{2i,H}(t; \Omega)/\bar{\Upsilon}_{2i}(t; \Omega).
\]

The score \( U_{H_t} \) represents a local directional functional derivative with respect to \( H \) defined as follows. For any functional \( f \)

\[
\frac{\partial f(H(\cdot))}{\partial dH(t)} = \frac{df(H(\cdot) + a \times 1(\cdot - t))}{da} \bigg|_{a=0},
\]

where \( 1(x) = 1 \) when \( x \geq 0 \) and 0 otherwise (an impulse function). Note that this definition corresponds to taking a derivative with respect to a jump of \( H \) at time \( t \) if \( H \) is a step-function.
Exchanging integrals, we have:

\[
U_{H_t} = \sum_{i=1}^{n} \sum_{k=1}^{2} \int_0^{t} \left[ dM_{k_i}(x) + \int_{x}^{\xi} \psi_{k_i}(u; \Omega) dM_{k_i}(u) dH_x \right] \\
+ \sum_{i=1}^{n} \int_0^{t} \left[ d\tilde{M}_{2i}(x) + \int_{x}^{\xi} \tilde{\psi}_{2i}(u; \Omega) d\tilde{M}_{2i}(u) dH_x \right] \\
= \sum_{i=1}^{n} \int_0^{t} \left\{ \sum_{k=1}^{2} \left[ 1 + \int_0^{u} \psi_{k_i}(u; \Omega) dH_x dM_{k_i}(u) \right] + \int_{t}^{\xi} \int_0^{t} \psi_{k_i}(u; \Omega) dH_x dM_{k_i}(u) \right\} \\
+ \sum_{i=1}^{n} \int_0^{t} \left\{ \left[ 1 + \int_0^{u} \tilde{\psi}_{2i}(u; \Omega) dH_x d\tilde{M}_{2i}(u) \right] + \int_{t}^{\xi} \int_0^{t} \tilde{\psi}_{2i}(u; \Omega) dH_x d\tilde{M}_{2i}(u) \right\} \\
= \sum_{i=1}^{n} \int_0^{\xi} \left[ \sum_{k=1}^{2} \varepsilon_{k_i}(u, t; \Omega) dM_{k_i}(u) + \tilde{\varepsilon}_{2i}(u, t; \Omega) d\tilde{M}_{2i}(u) \right],
\]

where \( \varepsilon_{k_i}(u, t; \Omega) = I(u \leq t) + \int_0^{u \land t} \psi_{k_i}(u; \Omega) dH_x, (k = 1, 2) \) and \( \tilde{\varepsilon}_{2i}(u, t; \Omega) = I(u \leq t) + \int_0^{u \land t} \tilde{\psi}_{2i}(u; \Omega) dH_x \). As we show in the Appendix, martingale transform \( \int_0^{\xi} \varepsilon_{k_i}(u, t; \Omega) dM_{k_i}(u), (k = 1, 2) \) and \( \int_0^{\xi} \tilde{\varepsilon}_{2i}(u, t; \Omega) d\tilde{M}_{2i}(x) \) are martingales as \( \varepsilon_{k_i}(u, t; \Omega), (k = 1, 2) \) and \( \tilde{\varepsilon}_{2i}(u, t; \Omega) \) do not depend on \( t \) for \( u \leq t \). Therefore, the score functions \( U_{\gamma} \) and \( U_{H_t} \) are both martingales at the true parameters.

In what follows we present the consistency and weak convergence results for the proposed NPMLE \( \hat{\Omega} = (\hat{\beta}, \hat{\alpha}, \hat{H_t}) = (\hat{\gamma}, \hat{H_t}) \). The detailed proofs may be found in Appendix.

4.4.3 Asymptotic Results

We use an empirical process argument for the Uniform Law of Large Numbers (consistency) ([71, 72]) and a martingale argument for weak convergence ([13, 14]).

Assume the following regularity conditions:

1. The true \( H \) is strictly increasing differentiable. The true values of parameter vector \( \beta \) and \( \alpha \) are in the interior of a compact Euclidean space.
2. With probability one, the covariate process $X(t)$ is left continuous with total bounded variation within $[0, \xi]$. Also, $X(t)$ is linearly independent in the sense that, if there exist $a(t)$ and $c$ such that $a(t) + c^T X(t) = 0$ with probability one, then $a(t) = 0$ and $c = 0$.

3. With probability one, $P(Y(\xi)|X(t)) > 0$, $P(\delta_1 = \delta_2 = 0, T = \xi|X(t)) > 0$, $P(\tilde{Y}(\xi)|X(t)) > 0$, $P(\delta_1 = 1, \delta_2 = 0, \tilde{T} = \xi|X(t)) > 0$. e.g., the at-risk set $Y(t)$ and $\tilde{Y}(t)$ will not shrink to empty.

4. Hessian matrix $\mathcal{I}_n$ evaluated at the true values of $\beta$, $\alpha$ and $H$ is positive definite, and converges to a deterministic matrix.

**Theorem IV.1.** Under regularity conditions, with probability one, $\hat{\beta}$ converges to $\beta^0$, $\hat{\alpha}$ converges to $\alpha^0$, $\hat{H}_t$ converges to $H^0_t$ uniformly in the interval $[0, \xi]$, where $\beta^0, \alpha^0, H^0_t$ are the true values of $\beta, \alpha, H_t$.

Consider a linear functional

\begin{equation}
(4.13) \quad n^{1/2}\left\{ \int_0^\xi a^T(\hat{\beta} - \beta^0) + b^T(\hat{\alpha} - \alpha^0) + c(t)^T(\hat{H}_t - H^0_t) \right\},
\end{equation}

where $a, b$ are real vectors, $C(t)$ is a function with bounded total variation in $[0, \xi]$, and let $C$ be the vector consisting of the values of $C(t)$ evaluated at the observed failure times corresponding to the set $\{dH\}$, and $\mathcal{E}^T = (a^T, b^T, C^T)$.

**Theorem IV.2.** Under regularity conditions, $n^{1/2}\{\hat{\beta} - \beta^0, \hat{\alpha} - \alpha^0, \hat{H}_t - H^0_t\}$ converges weakly to a zero-mean Gaussian process. In addition, $n\mathcal{E}^T(\mathcal{I}_n)^{-1}\mathcal{E}$ converges in probability to the asymptotic variance-covariance function of the linear functional (4.13), where $\mathcal{I}_n$ is the negative Hessian matrix of the observed log-likelihood function with respect to $\hat{\Omega}$.
For a differentiable functional $F(\Omega)$ of $\Omega$, based on the functional delta method ([3] Section II.8), $n^{1/2}\{F(\hat{\Omega}) - F(\Omega)\}$ converges weakly to a zero-mean Gaussian process with variance-covariance function $\hat{F}(\Omega)^T (I^0)^{-1} \hat{F}(\Omega)$, where $\hat{F}(\Omega)$ is the gradient of $F(\Omega)$ with respect to $\Omega$, $I^0$ can be consistently estimated by $n^{-1}I_n$, with the explicit expression of $I_n$ is derived in Appendix.

**Remark 3.** We may also use the profile likelihood method ([48]) to estimate the covariance matrix of $\hat{\beta}$ and $\hat{\alpha}$. The estimator is the negative inverse of the second-order numerical differences of the profile log-likelihood function at $\hat{\beta}$ and $\hat{\alpha}$, respectively. This approach avoids inverting a potentially large matrix but fails to provide a variance estimator for $\hat{H}(\cdot)$.

### 4.5 Numerical Examples

#### 4.5.1 Simulation Studies

<table>
<thead>
<tr>
<th>Event proportions in different simulation scenarios</th>
<th>$\theta = 0.5$</th>
<th>$\theta = 1$</th>
<th>$\theta = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1^<em>$ and $T_2^</em>$ observed</td>
<td>50.5%</td>
<td>34.9%</td>
<td>37.4%</td>
</tr>
<tr>
<td>$T_1^<em>$ observed, $T_2^</em>$ censored</td>
<td>14.4%</td>
<td>21.0%</td>
<td>7.0%</td>
</tr>
<tr>
<td>$T_1^<em>$ missed, $T_2^</em>$ observed</td>
<td>18.8%</td>
<td>13.1%</td>
<td>37.3%</td>
</tr>
<tr>
<td>$T_1^<em>$ and $T_2^</em>$ censored</td>
<td>16.3%</td>
<td>31.0%</td>
<td>18.3%</td>
</tr>
</tbody>
</table>

Monte Carlo simulation studies are carried out to evaluate the finite-sample properties of proposed estimators. When frailty terms follow gamma distribution, we consider three scenarios that representing different levels of heterogeneity, namely the variance of gamma frailty is $\theta = 0.5, \theta = 1$ or $\theta = 2$. Two covariates are included in the regression model for progression, death, and missingness of progression. One is a continuous variable uniformly distributed on $(0, 1)$ ($Z_1$), and the other is a binary variable with success probability 0.5 ($Z_2$). Baseline hazard function $h(t) = 0.15 \sqrt{t/10}$, which follows an increasing Weibull distribution. The censoring
time is taken to follow a uniform distribution for each scenario, which results different proportions of events, as summarized in Table 4.1. These event proportions aim to represent a range of different situations in practice. For each simulation scenario, \( n = 250 \), \( n = 500 \) and \( n = 1,000 \) are considered, and all analysis results are based on 1,000 simulation replications.

### Table 4.2: Simulation results for gamma frailty variance \( \theta = 0.5 \)

<table>
<thead>
<tr>
<th></th>
<th>Progression</th>
<th>Death</th>
<th>Missingness</th>
<th>Frailty</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_{\mu_1} )</td>
<td>( \beta_{\mu_2} )</td>
<td>( \beta_{\eta_1} )</td>
<td>( \beta_{\eta_2} )</td>
<td>( \alpha_0 )</td>
<td>( \alpha_1 )</td>
</tr>
<tr>
<td>True</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1.5</td>
</tr>
<tr>
<td>Bias</td>
<td>-0.02</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>-0.04</td>
</tr>
<tr>
<td>ASE</td>
<td>0.39</td>
<td>0.23</td>
<td>0.30</td>
<td>0.20</td>
<td>0.40</td>
</tr>
<tr>
<td>ESE</td>
<td>0.40</td>
<td>0.23</td>
<td>0.31</td>
<td>0.21</td>
<td>0.42</td>
</tr>
<tr>
<td>CP</td>
<td>94.6%</td>
<td>94.2%</td>
<td>94.7%</td>
<td>94.2%</td>
<td>94.1%</td>
</tr>
<tr>
<td>( n = 250 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>-0.01</td>
<td>-0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>-0.03</td>
</tr>
<tr>
<td>ASE</td>
<td>0.28</td>
<td>0.16</td>
<td>0.21</td>
<td>0.14</td>
<td>0.28</td>
</tr>
<tr>
<td>ESE</td>
<td>0.28</td>
<td>0.16</td>
<td>0.22</td>
<td>0.14</td>
<td>0.29</td>
</tr>
<tr>
<td>CP</td>
<td>94.1%</td>
<td>94.2%</td>
<td>94.6%</td>
<td>95.0%</td>
<td>95.2%</td>
</tr>
<tr>
<td>( n = 500 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ASE</td>
<td>0.20</td>
<td>0.11</td>
<td>0.15</td>
<td>0.10</td>
<td>0.19</td>
</tr>
<tr>
<td>ESE</td>
<td>0.20</td>
<td>0.11</td>
<td>0.15</td>
<td>0.10</td>
<td>0.19</td>
</tr>
<tr>
<td>CP</td>
<td>94.4%</td>
<td>95.2%</td>
<td>95.4%</td>
<td>94.4%</td>
<td>95.0%</td>
</tr>
</tbody>
</table>

ESE: empirical standard errors based on 1,000 estimates  
ASE: average of estimated standard errors  
CP: 95% Coverage Probability

Tables 4.2, 4.3 and 4.4 summarize the simulation studies when \( \theta = 0.5 \), \( \theta = 1 \) and \( \theta = 2 \), respectively. At all sample sizes we consider, the estimates of regression coefficients for progression \((T_1^*)\), death \((T_2^*)\), and missingness of progression \((p)\), are almost unbiased. The average of estimated standard errors (ASE) and the empirical standard errors (ESE) based on Monte Carlo replicates are in agreement with each other, which indicates that the proposed variance estimators are reliable. The 95% coverage probability (CP) also remain at nominal levels, which suggests the normality of proposed estimators is preserved in finite-sample settings. We notice that the estimator for the frailty variance \( \hat{\theta} \) is slightly biased when sample size is small \((n = \)
Table 4.3: Simulation results for gamma frailty variance $\theta = 1$

<table>
<thead>
<tr>
<th>Progression</th>
<th>Death</th>
<th>Missingness</th>
<th>Frailty $\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{\mu_1}$</td>
<td>$\beta_{\mu_2}$</td>
<td>$\beta_{\eta_1}$</td>
<td>$\beta_{\eta_2}$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>True</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 250$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>-0.03</td>
<td>-0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>ASE</td>
<td>0.48</td>
<td>0.28</td>
<td>0.38</td>
</tr>
<tr>
<td>ESE</td>
<td>0.49</td>
<td>0.28</td>
<td>0.39</td>
</tr>
<tr>
<td>CP</td>
<td>94.1%</td>
<td>95.3%</td>
<td>94.2%</td>
</tr>
<tr>
<td>$n = 500$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>0.00</td>
<td>-0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>ASE</td>
<td>0.34</td>
<td>0.19</td>
<td>0.27</td>
</tr>
<tr>
<td>ESE</td>
<td>0.35</td>
<td>0.20</td>
<td>0.27</td>
</tr>
<tr>
<td>CP</td>
<td>94.8%</td>
<td>94.7%</td>
<td>95.5%</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ASE</td>
<td>0.24</td>
<td>0.14</td>
<td>0.19</td>
</tr>
<tr>
<td>ESE</td>
<td>0.24</td>
<td>0.14</td>
<td>0.20</td>
</tr>
<tr>
<td>CP</td>
<td>94.6%</td>
<td>94.5%</td>
<td>93.8%</td>
</tr>
</tbody>
</table>

ESE: empirical standard errors based on 1,000 estimates
ASE: average of estimated standard errors
CP: 95% Coverage Probability

Table 4.4: Simulation results for gamma frailty variance $\theta = 2$

<table>
<thead>
<tr>
<th>Progression</th>
<th>Death</th>
<th>Missingness</th>
<th>Frailty $\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{\mu_1}$</td>
<td>$\beta_{\mu_2}$</td>
<td>$\beta_{\eta_1}$</td>
<td>$\beta_{\eta_2}$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>True</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 250$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>-0.02</td>
<td>-0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>ASE</td>
<td>0.57</td>
<td>0.33</td>
<td>0.45</td>
</tr>
<tr>
<td>ESE</td>
<td>0.59</td>
<td>0.33</td>
<td>0.44</td>
</tr>
<tr>
<td>CP</td>
<td>94.6%</td>
<td>95.1%</td>
<td>95.9%</td>
</tr>
<tr>
<td>$n = 500$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>ASE</td>
<td>0.40</td>
<td>0.24</td>
<td>0.32</td>
</tr>
<tr>
<td>ESE</td>
<td>0.40</td>
<td>0.23</td>
<td>0.32</td>
</tr>
<tr>
<td>CP</td>
<td>95.1%</td>
<td>95.3%</td>
<td>95.0%</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ASE</td>
<td>0.28</td>
<td>0.17</td>
<td>0.22</td>
</tr>
<tr>
<td>ESE</td>
<td>0.28</td>
<td>0.16</td>
<td>0.23</td>
</tr>
<tr>
<td>CP</td>
<td>94.8%</td>
<td>95.6%</td>
<td>95.6%</td>
</tr>
</tbody>
</table>

ESE: empirical standard errors based on 1,000 estimates
ASE: average of estimated standard errors
CP: 95% Coverage Probability

250), and the bias diminishes as sample size increase. Such observation is expected and consistent with [50], as the unobserved frailty may not be accurately estimated
at small sample sizes. In addition, as the underlying variance of frailty increases, i.e., the simulated dataset becomes more heterogeneous, the standard errors of all covariate coefficients increase.

### 4.5.2 Stage C Colorectal Cancer Adjuvant Therapy

Each year more than 100,000 new cases of colon cancer are diagnosed in United States, and it remains the second leading cause of cancer-related death for decades. Back to the 80’s of last century, no established reliable screening approach for early diagnosis. Among patients who have regional nodal involvement that is clinically completely resected (Duke’s Stage C disease), more than half of them have residual cancer that leads to cancer recurrence and death within 5 years. Chemotherapy as adjuvant therapies have been thus studied to reduce cancer recurrence and prolong patients’ lives. The seminal randomized clinical trial of levamisole (Lev) and fluorouracil (5FU) for adjuvant therapy of resected colon cancers ([46]) first demonstrated the effectiveness of fluorouracil-based chemotherapy for resected colorectal cancers. This study was designed to determine the efficacy of Lev alone or 5FU+Lev as adjuvant therapy regimens comparing with observation in improving surgical cure rates in Duke’s stage C colon cancer. The endpoints of interests were cancer recurrence and death. Some important questions that was not addressed by the investigators include the association of cancer recurrence with death, and how the therapy benefits patients’ survival after cancer recurrence. In this analysis, we use the Lev+5FU and observation arm (Obs) of the trial.

Table 4.5 summarizes the number of cancer recurrences and deaths in each arm. The median times to recurrence and death are 4.9 and 5.5 years, respectively. In this study, a relatively small fraction of patients were died without prior recorded recurrence (4.8% in Obs and 5.9% in Lev+5FU). These patients’ recurrence are likely
Table 4.5: Summary of stage C colon cancer adjuvant therapy study

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Observation (N=315)</th>
<th>5FU+Lev (N=304)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no. of patients (%)</td>
<td>no. of patients (%)</td>
</tr>
<tr>
<td>Obstruction</td>
<td>63 (20.0%)</td>
<td>54 (17.8%)</td>
</tr>
<tr>
<td>Adherence to adjacent organs</td>
<td>47 (14.9%)</td>
<td>39 (12.8%)</td>
</tr>
<tr>
<td>≥21 days since surgery</td>
<td>91 (28.9%)</td>
<td>76 (25.0%)</td>
</tr>
<tr>
<td>&gt;4 Positive lymph nodes</td>
<td>87 (27.6%)</td>
<td>79 (26.0%)</td>
</tr>
<tr>
<td>Invade serosa</td>
<td>269 (85.4%)</td>
<td>262 (86.2%)</td>
</tr>
<tr>
<td>Age at Treatment, year</td>
<td>59.5 (12.0)</td>
<td>59.7 (12.3)</td>
</tr>
<tr>
<td>Recurrence and Death</td>
<td>153 (48.6%)</td>
<td>105 (34.5%)</td>
</tr>
<tr>
<td>Recurrence and Censored</td>
<td>22 (7.0%)</td>
<td>11 (3.6%)</td>
</tr>
<tr>
<td>Death w/o Recorded Recurrence</td>
<td>15 (4.8%)</td>
<td>18 (5.9%)</td>
</tr>
<tr>
<td>No Recurrence or Death</td>
<td>125 (39.7%)</td>
<td>170 (55.9%)</td>
</tr>
</tbody>
</table>

missed. Important prognostic factors included occurrence of obstruction of colon by tumor, adherence to adjacent organs, whether randomization occurred more than 21 days since surgery, whether more than 4 positive lymph nodes involved, whether tumor spreads to serosa or not, and age at surgery (centered at 61 years and scaled by 12.1 years, the median and standard deviation of age, respectively). These prognostic factors are also summarized in Table 4.5.

The disease-free survival (time to recurrence or death) and overall survival are summarized by Kaplan-Meier estimates in Figure 4.2 and 4.3. It is clear that Lev+5FU group delayed the time to recurrence as well as time to death as measured from randomization. Cox proportional hazard (PH) models are also applied to DFS and OS to estimate the treatment effect with adjustments of important prognostic factors (Table 4.6). The analysis results suggest that patients receiving Lev+5FU who have less than 4 positive lymph nodes involved have longer times to recurrence and death.

We also applied a shared gamma-frailty Cox PH model, where a common frailty $Z_i$ is shared between $T_{1i}$ and $T_{2i}$ for subject $i$. It is noted that this model ignores the ordering of progression and death that $T_1^* \leq T_2^*$. This model has been stud-
ied by [38, 59], among many others. We follow the approach developed by [58, 59] based on penalized likelihood, who also show that it’s equivalent to [38, 7] based on EM algorithm for gamma frailty with certain penalty terms. This approach has been implemented in `survival` package in R. As suggested by [58], the variance of the frailty variance may be obtained by resampling approaches such as bootstrap or jackknife when sample size is not sufficiently large, and thus the standard errors we report are based on jackknife method. As commented by [18], failing to address the ordering of progression and death could result in a poor fit of the joint distribution \((T_1^*, T_2^*)\), and lead to bias in estimating the associations, since probability mass would be assigned to the wedge \(T_1^* > T_2^*\) in estimation. Interestingly, based on this model, the treatment effect between Lev+5FU and observation, although

![Kaplan-Meier plot for recurrence-free survival by treatment group](image_url)
still favoring Lev+5FU, is not significantly different from null anymore. Meanwhile, older patients who had colon obstruction by tumor, adherence to adjacent organs, more than 4 positive lymph nodes involved, and received treatment more than 21 days after surgery, were more likely to experience cancer recurrence and death. The estimated variance of gamma frailty $\theta$ is close to a nontrivial 6.0 with standard error 0.60 ($P < 0.001$). Furthermore, while frailty models generally tend to provide covariate effects that are more likely away from null than marginal models ([7]), failing to address the ordering issue of progression and death may result in the insignificant treatment effect for both progression and death.

The analysis results based on the proposed model are summarized in Table 4.7. Lev+5FU is found effective in prolonging the time to recurrence ($\hat{\beta} = -0.60, P =$
Table 4.6: Analysis results of separate Cox marginal models and Cox frailty model

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Marginal DFS</th>
<th>Marginal OS</th>
<th>Cox Frailty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est.</td>
<td>SE</td>
<td>P-value</td>
</tr>
<tr>
<td>$Z_1$</td>
<td>-0.46</td>
<td>0.11</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>$Z_2$</td>
<td>-0.01</td>
<td>0.06</td>
<td>0.86</td>
</tr>
<tr>
<td>$Z_3$</td>
<td>0.14</td>
<td>0.14</td>
<td>0.33</td>
</tr>
<tr>
<td>$Z_4$</td>
<td>0.26</td>
<td>0.15</td>
<td>0.09</td>
</tr>
<tr>
<td>$Z_5$</td>
<td>0.25</td>
<td>0.12</td>
<td>0.04</td>
</tr>
<tr>
<td>$Z_6$</td>
<td>0.84</td>
<td>0.12</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>$\theta$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

$Z_1$: randomization arm: Lev+5FU vs Obs
$Z_2$: age (per 12.1 years)
$Z_3$: obstruction vs. not
$Z_4$: adherence to adjacent organs vs. not
$Z_5$: post-surgery duration long vs. short
$Z_6$: number of positive lymph nodes >4 vs. ≤4

Table 4.7: Analysis results of proposed model with unobserved recurrence

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Progression</th>
<th>Death</th>
<th>Missingness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est.</td>
<td>SE</td>
<td>P-value</td>
</tr>
<tr>
<td>$Z_1$</td>
<td>-0.60</td>
<td>0.24</td>
<td>0.01</td>
</tr>
<tr>
<td>$Z_2$</td>
<td>0.11</td>
<td>0.11</td>
<td>0.33</td>
</tr>
<tr>
<td>$Z_3$</td>
<td>0.94</td>
<td>0.37</td>
<td>0.01</td>
</tr>
<tr>
<td>$Z_4$</td>
<td>1.02</td>
<td>0.37</td>
<td>0.01</td>
</tr>
<tr>
<td>$Z_5$</td>
<td>0.74</td>
<td>0.28</td>
<td>0.01</td>
</tr>
<tr>
<td>$Z_6$</td>
<td>1.94</td>
<td>0.28</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Intercept</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\theta$</td>
<td>3.26</td>
<td>0.61</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

$Z_1$: randomization arm: Lev+5FU vs Obs
$Z_2$: age (per 12.1 years)
$Z_3$: obstruction vs. not
$Z_4$: adherence to adjacent organs vs. not
$Z_5$: post-surgery duration long vs. short
$Z_6$: number of positive lymph nodes >4 vs. ≤4

0.013), but not for the time to death ($\hat{\beta} = 0.06, P = 0.74$). The frailty parameter was estimated to be $\hat{\theta} = 3.26$, with standard errors 0.61 ($P < 0.001$). In addition, patients with colon obstruction by tumor, adherence to adjacent organs, and more than 4 positive lymph nodes involved were more likely to observe recurrence and death. Receiving treatment more than 21 days after surgery is found only significantly increase the likelihood of recurrence but not death. There is also some evidence of missing progressions, increasingly so in older patients. Combining all the above results, stage C colon cancer patients were found very heterogenous in terms of
cancer progression, and the survival benefit of receiving Lev+5FU, if any, was mainly through prolonging cancer recurrence but not afterwards. This finding was first hypothesized by [47], and subsequently studied by [42] based on a marginal gap time model without adjustments of prognostic factors. Our model provided a novel insight into this issue, namely the heterogeneity among cancer patients, including their responses to Lev+5FU treatment.

4.6 Discussion

In this paper, we propose a progressive multistate model to jointly model progression and death in advanced cancer studies accounting for a possible missing data mechanism affecting progression events. In advanced disease progression and death are stages of the same development and we adopted an ordering restriction and a common baseline hazard function to address this. As pointed out by [18], failing to impose the ordering may lead to severe bias in estimation of joint distribution. Our approach addresses potential complications arising with a variety of progression-related endpoints such as TTP and PFS/DFS, and incorporates all available information from both completely and partially observed subjects to make inference about the joint distribution of progression and death. The dependence between progression and death is modeled through covariates, a shared frailty, and the order restriction, making the description informative and flexible. Unlike [18] where the baseline hazard function is parametric or piece-wise constant, the semiparametric NPMLE machinery is involved treating the baseline hazard summarizing disease development dynamic pattern as fully unspecified and data-driven. Kendall’s τ was found to be independent of the baseline hazard. This property allows manipulation of local dependence at a fixed global dependence level.
The proportionality assumption between progression and death given covariates and fixed frailty can be relaxed if needed without modifying the general line of reasoning. Specifically, Markov modulation of the hazard of death by the occurrence of progression can be made more complex to include a parametrically specified transformation.
This dissertation presents a unified approach to utilize information on indirectly or partially observable disease natural history process based on observed events governed by this process. A common disease progression process that generates both the disease natural history events and the observed survival time consists of the complete data multivariate survival structure that statistical models built upon. This approach is particularly suitable for the task of integration of the observed pieces of information and elucidation of the complex associations between them. The approach also offers a way to pose justifiable and flexible assumptions that make the model identifiable. Basing the model on a series of nested PH model and a modulation mechanism of correlation between successive events, this approach bypasses the technical difficulties arising in de-convolution literature in the classical progressive multistate models for latent disease natural history, [68, 2, 56], and avoids the potential nonidentifiability when treating some of the observed events as a nuisance. Throughout the dissertation, the view of the unobservable disease natural history process as missing data greatly facilitates the model formulation and the development of estimation algorithms and the asymptotic results. Various missing data mechanisms and assumptions of competition vs. succession for the series of progres-
sion events and the survival time make up the variety of related models considered in this thesis. Marginalizing over the missing data induces a transformation model enabling utilization of the associated algorithmic and theoretical methodology for statistical estimation and inference.

The current framework may be extended in several directions:

- More complex observed phenotype and missing data mechanism.

  The observed data may be more complicated representing multivariate marks on the survival time. On the latent level serial and parallel developments of disease progression process can be hypothesized and the corresponding hypotheses on the latent model structure can be tested based on the observed data. A classical example for the complex multistate model arises in bone marrow transplantation ([39]). Depending on the observation schemes, any one of the intermediate states that is either completely or partially missing would be interesting to investigate. Another scenario could be a model for cancer stage and grade progression, in which the stage progression and grade progression may develop sequentially or represent a selection from the fixed grade at tumor initiation that never changes within the subject. Furthermore, depending on the observed data, the stage and grade progressions may be either completely or partially missing. Last but not least, the missingness mechanism may not necessarily be at random, and approaches such as pattern mixture ([43]) may be introduced to handle nonignorable missingness.

- Screening mechanism.

  In cancer prevention settings, the point of cancer diagnosis is random and greatly affected by early detection measures such as screening. One may consider a more detailed multistate model that would incorporate the screening effects in the
detection of cancer that would naturally lead to a less advanced stage and grade at diagnosis without necessarily leading to a better prognosis. Incorporating treatment effects in this causal context would be of great interest.

- Data synthesis.

The missing data viewpoint may also be used to synthesize datasets from different sources. For example, one may combine clinical trial datasets where detailed information of death and recurrence is available with vast cancer registry datasets that do not have recurrence information, thus representing an intermediate event missing by design. Under some general assumptions, one may thus assume that subjects in registry datasets, whose progression information is completely missing, still share the same or similar disease natural history as those in clinical trials. The question of whether clinical trials datasets are representative of the US population may call for additional modeling efforts. However the promise of greatly improved statistical power based on the combined datasets would motivate further research in this area.

- Causal inference.

The event history framework offers a complete model of possible trajectories linking multiple observed events. Therefore, it is plausible to dissect a direct effect of treatment on the survival time from the mediator effect exerted through the latent progression process (and then through the process on survival). A systematic causal approach of the kind [1] may help to better elucidate the treatment mechanism and design better treatments.

- Surrogacy evaluation.

The framework proposed in this dissertation essentially provides the joint distri-
bution of the observed multiple events in presence of missingness. Meanwhile, the nonterminal event that represents the disease natural history of interest, is always of great interest and serves as a surrogate endpoint for the terminal event. Therefore, it is possible to develop a more systematic approach to evaluate the value of such surrogacy based on the joint distribution.

- Efficiency improvement for observed data.
  By treating the progression-related events as auxiliary variables ([23, 30]), one may subsequently apply the resultant joint distribution of multiple events to improve the estimation efficiency of the model for the terminal event.

- Incorporation of time-dependent variables such as biomarkers.
  The dependence between multiple events and disease natural history is introduced through a modulation mechanism that leads itself naturally to incorporation of time-dependent covariates. We will then have a better chance of understanding the disease progression mechanism and many other aspects of the disease monitoring and treatment. Besides, biomarkers could be considered informative of a continuous disease progression process.
APPENDICES
APPENDIX A

Important Results for Asymptotic Properties

A.1 Property of Transform $\int_{0}^{\xi} \varepsilon(u, t; \Omega) dM(u)$

Let $V(t) = \int_{0}^{\xi} \varepsilon(u, t; \Omega) dM(u)$, where $\varepsilon$ is a predictable function such that it only depends on $u$ when $u < t$. We have

$$dV(t) = \int_{0}^{\xi} \varepsilon(u, t + dt) dM(u) - \varepsilon(u, t) dM(u) = \int_{0}^{\xi} \varepsilon'(u, t) dM(u),$$

where $\varepsilon'(u, t)$ is the partial derivative of $\varepsilon(u, t)$ with respect to $t$. Taking an expectation conditional on filtration,

$$E\{dV(t)|\mathcal{F}_{t-}\} = \int_{0}^{\xi} E\{\varepsilon'(u, t) dt dM(u)|\mathcal{F}_{t-}\} = dt \int_{0}^{\xi} \varepsilon'(u, t) E\{dM(u)|\mathcal{F}_{t-}\} = dt \int_{0}^{t} \varepsilon'(u, t) dM(u),$$

since $E\{dM(u)|\mathcal{F}_{t-}\} = I(u < t) dM(u)$. Finally, $E\{dV(t)|\mathcal{F}_{t-}\} = 0$ if $\varepsilon'(u, t) = 0$ when $u < t$. Therefore, $V(t)$ is a martingale.
APPENDIX B

Asymptotic Results for Chapter II

B.1 Asymptotic Results for Estimating Equation Based Estimator

B.1.1 Regularity Conditions

Following [22] (p289-p290), we assume the following regularity conditions to prove
the asymptotics of \( \hat{\beta} \) and \( \hat{H}(\cdot) \):

1. The true \( H \) is strictly increasing differentiable. The true values of parameter
   vector \( \beta \) are in the interior of a compact Euclidean space.

2. With probability one, the covariate process \( Z(t) \) is left continuous with total
   bounded variation within \([0, \tau]\). Also, \( Z(t) \) is linearly independent in the sense
   that, if there exist \( a(t) \) and \( c \) such that \( a(t) + c^T Z(t) = 0 \) with probability one,
   then \( a(t) = 0 \) and \( c = 0 \).

3. With probability one, \( P(Y(\tau)|Z(t)) > 0, P(\delta_1 = \delta_2 = 0, T = \tau|Z(t)) > 0, \)
   \( P(\bar{Y}(\tau)|Z(t)) > 0, P(\delta_1 = 1, \delta_2 = 0, \bar{T} = \tau|Z(t)) > 0 \). e.g., the at-risk set \( Y(t) \)
   and \( \bar{Y}(t) \) will not shrink to empty.

4. For \( S_0(\cdot), S_0^*(\cdot), \hat{S}_1(\cdot), \hat{S}_1^*(\cdot), \bar{S}_2(\cdot) \), there exists a neighborhood \( \mathcal{B} \) of \( \beta^0 \) and, re-
   spectively, scalar, vector, and matrix functions of \( s_0(\cdot), s_0^*(\cdot), s_1(\cdot), s_1^*(\cdot), s_2(\cdot) \) defined on
\( \mathcal{B} \times [0, \tau] \) such that as as \( n \to \infty \)

\[
\sup_{x \in [0,\tau], \beta \in \mathcal{B}} ||S_j(x; \beta, H) - s_j(x; \beta, H)|| \to 0, \quad \sup_{x \in [0,\tau], \beta \in \mathcal{B}} ||S_j^*(x; \beta, H) - s_j^*(x; \beta, H)|| \to 0,
\]

\[
\sup_{x \in [0,\tau], \beta \in \mathcal{B}} ||\tilde{S}_2(x; \beta, H) - \tilde{s}_2(x; \beta, H)|| \to 0, \quad j = 0, 1
\]

5. \( \Sigma_\ast \) and \( \Sigma^* \) are positive definite.

\textbf{B.1.2 Proof of Theorem II.1}

The proofs accomplished in 5 steps.

\textbf{Step 1} Show uniform consistency of \( \hat{H}(t; \beta^0) \) using Theorem IV.1.1. [3] or Lenglart’s inequality.

\textbf{Step 2} Show \( a_n \left[ \hat{H}_t(\beta^0) - H^0_t \right] \xrightarrow{D} W(t) \), where \( W(t) = \kappa(t; \beta^0, H^0) \int_0^t \kappa^{-1}(x; \beta^0, H^0) dV(x) \), and \( V(t) \) is a Gaussian martingale process with variance \( \sigma^2(t) = \int_0^t s_0^{-1}(x; \beta^0, H^0) dH^0_0 \).

This follows from

\[
n^{1/2} \left[ \hat{H}_t(\beta) - H^0_t \right] = n^{1/2} \int_0^t \frac{dM(x)}{S_0(x; \beta^0, H)} + n^{1/2} \int_0^t \frac{S_0(x; \beta^0, H^0) - S_0(x; \beta^0, \hat{H})}{S_0(x; \beta^0, H)} dH^0_0,
\]

and

\[
n^{1/2} \int_0^t \frac{dM(x)}{S_0(x; \beta^0, H)} \xrightarrow{D} V(t), \quad W(t) = V(t) - \int_0^t \frac{s_0^*(x; \beta^0, H^0)}{s_0(x; \beta^0, H^0)} W(x) dH^0_0 + o_p(1).
\]

\textbf{Step 3} Using a technique as in step 2, we can show the Jacobian operator

\[
\hat{H}_t(\beta) := \frac{d}{d\beta} \hat{H}_t(\beta) = -\kappa(t; \beta, H^0) \int_0^t \kappa^{-1}(x; \beta, H^0) \frac{s_1(x; \beta, H^0)}{s_0(x; \beta, H^0)} dH^0_0 + o_p(1).
\]

\textbf{Step 4} Here we derive asymptotic distribution of \( U_{pr}(\beta) \).

Using Taylor expansion and results from steps 1-3, we can show

\[
n^{-1/2} U_{pr}(\beta^0) = n^{-1/2} \sum_{i=1}^n \sum_{s=0}^1 \int_0^t \left[ \hat{Q}_s(x; \beta^0, \hat{H}) - Q(x; \beta^0, \hat{H}) \right] dM_{si}(x) + o_p(1),
\]

such that by Martingale Central Limit Theorem and under regularity conditions, we have \( n^{-1/2} U_{pr}(\beta^0) \xrightarrow{D} N(0, \Sigma^*) \) accordingly.
Step 5 Here we derive asymptotic properties of $\mathcal{I}_{pr}(\beta) = \mathcal{I}_{pr}(\tau, \beta) = -\frac{d}{d\beta} U_{pr}(\beta)$.

Using Taylor expansion and results from steps 1-3, we can show

$$-n^{-1} \mathcal{I}_{pr}(\beta)|_{\beta = \beta^0} = -n^{-1} \sum_{i=1}^{n} \sum_{s=1}^{1} \int_{0}^{\tau} \left[ \frac{\tilde{\Theta}_{si}(x; \beta, \hat{H})}{\Theta_{si}(x; \beta, \hat{H})} - \tilde{S}_2(x; \beta^0, \hat{H}) \right] dM_{si}(x)$$

$$+ n^{-1} \sum_{i=1}^{n} \sum_{s=1}^{1} \int_{0}^{\tau} \left[ \frac{\tilde{\Theta}^2_{si}(x; \beta, \hat{H})}{\Theta^2_{si}(x; \beta, \hat{H})} - \tilde{S}^2_1(x; \beta^0, \hat{H}) \right] dM_{si}(x)$$

$$+ n^{-1} \int_{0}^{\tau} \frac{\tilde{S}_2(x; \beta, \hat{H}) \tilde{S}_0(x; \beta^0, \hat{H}) - \tilde{S}^2_1(x; \beta^0, \hat{H})}{\tilde{S}_0(x; \beta^0, \hat{H})} dH^0_x + o_p(1).$$

The first three terms all converge to 0 in probability by Lenglart’s inequality, and $-n^{-1} \mathcal{I}_{pr}(\beta)|_{\beta = \beta^0} \overset{P}{\to} \Sigma_*$ by the Weak Law of Large Numbers (WLLN).

B.1.3 Additional Notation

Let $t$ be some observed event time, $dH(t)$ be the jump of the step function $H(\cdot)$, and denote the functional derivative of $\Theta_{si}(t; \beta, H)$ with respect to $\{dH\}$ as $\Theta_{si,H}(t; \beta, H) = \partial \Theta_{si}(t; \beta, H)/\partial dH$. We define the following quantities

$$S^*_0(t; \beta, \hat{H}) = n^{-1} \sum_{i,s} Y_i(t) \Theta_{si,H}(t; \beta, \hat{H}); \quad \tilde{S}^*_1(t; \beta, \hat{H}) = n^{-1} \sum_{i,s} Y_i(t) \frac{\tilde{\Theta}_{si}(t; \beta, \hat{H})}{\Theta_{si}(t; \beta, \hat{H})} \Theta_{si,H}(t; \beta, \hat{H});$$

$$\tilde{S}_2(t; \beta, \hat{H}) = n^{-1} \sum_{i,s} Y_i(t) \frac{\tilde{\Theta}^2_{si}(t; \beta, \hat{H})}{\Theta^2_{si}(t; \beta, \hat{H})} \Theta_{si,H}(t; \beta, \hat{H}); \quad \tilde{S}^*_2(t; \beta, \hat{H}) = n^{-1} \sum_{i,s} Y_i(t) \frac{\tilde{\Theta}^2_{si}(t; \beta, \hat{H})}{\Theta^2_{si}(t; \beta, \hat{H})} \Theta_{si,H}(t; \beta, \hat{H});$$

$$\tilde{\Theta}_{si}(x; \beta, \hat{H}) = \frac{d^2}{d\beta \partial \beta^0} \Theta_{si}(x; \beta, \hat{H}); \quad \tilde{S}^*_2(t; \beta, \hat{H}) = n^{-1} \sum_{i,s} Y_i(t) \frac{\tilde{\Theta}_{si}(t; \beta, \hat{H})}{\Theta_{si}(t; \beta, \hat{H})} \Theta_{si,H}(t; \beta, \hat{H});$$

$$\tilde{S}_2(t; \beta, \hat{H}) = n^{-1} \sum_{i,s} Y_i(t) \tilde{\Theta}_{si}(t; \beta, \hat{H}); \quad K(t; \beta, \hat{H}) = \exp(- \int_{0}^{t} \frac{S^*_0(x; \beta, \hat{H})}{S_0(x; \beta, \hat{H})} d\hat{H}_x).$$
\[ L(t; \beta, \hat{H}) = \frac{\dot{S}_1(t; \beta, \hat{H})S_0(t; \beta, \hat{H})}{S_0(t; \beta, \hat{H})} - \ddot{S}_1(t; \beta, \hat{H}); \]
\[ \ddot{L}(t; \beta, \hat{H}) = \frac{\ddot{S}_2(t; \beta, \hat{H})S_0^*(t; \beta, H^0) - \dddot{S}_2(t; \beta, \hat{H})S_0(t; \beta, H^0)}{S_0^*(t; \beta, H^0)}; \]
\[ \ddot{L}(t; \beta, \hat{H}) = \frac{\dddot{S}_2^2(t; \beta, \hat{H})S_0^*(t; \beta, H^0) - \ddot{S}_2^2(t; \beta, \hat{H})S_0^*(t; \beta, H^0)}{S_0^2(t; \beta, H^0)}; \]
\[ Q(t; \beta, \hat{H}) = \frac{\dot{S}_1(t; \beta, \hat{H})}{S_0(t; \beta, \hat{H})} - \left[ \int_t^\tau \dot{L}(u; \beta, \hat{H})K(u; \beta, \hat{H})d\hat{H}_u \right] / \left[ K(t; \beta, \hat{H})S_0(t; \beta, \hat{H}) \right]. \]

**B.2 Asymptotic Results for Nonparametric Maximum Likelihood Estimator**

**B.2.1 Regularity Conditions**

In addition to regularity conditions of the previous section assume

1. Hessian matrix \( \tilde{L}_n \) evaluated at the true values of \( H \) and \( \beta \) is positive definite, and converges to a deterministic matrix

Let \( \| \cdot \|_{l^\infty[0,\tau]} \) denote the supremum norm in \([0, \tau]\), and \( \| w \|_{BV[0,\tau]} \) the total variation of \( w(t) \) in \([0, \tau]\). Also define \( Q = \{ w(t) : \| w \|_{BV[0,\tau]} \leq 1 \} \). Such that \( \tilde{H}_t \) may be regarded as a bounded linear functional in \( l^\infty(Q) \), and \( \{ \tilde{\beta} - \beta^0, \tilde{H}_t - H^0_t \} \) a random element in the metric space \( \mathcal{R}^p \times l^\infty(Q) \), where \( p \) is the dimension of \( \beta^0 \). We denote \( \mathcal{H} \) as the compact convex set in the metric space \( \mathcal{R}^p \times l^\infty(Q) \) in which \( \Omega^0 \) is contained.

**B.2.2 Proof of Theorem II.2**

To establish the consistency result, i.e., \( \| \tilde{H}_t - H^0_t \|_{l^\infty(Q)} \to 0 \) and \( |\tilde{\beta} - \beta^0| \to 0 \), in addition to conditions 1-3, we want to verify the following conditions:

1. identifiability condition: suppose that for any sequence \( \Omega_n \in \mathcal{H} \), the compact convex set in the metric space \( \mathcal{R}^p \times l^\infty(Q) \), \( \liminf_{n \to \infty} \ell(\Omega_n) \geq \ell(\Omega^0) \) implies \( \| \Omega_n - \Omega^0 \| \to 0 \).
2. uniform convergence condition: for any sequence $\Omega \in \mathcal{H}$ we have uniform convergence, i.e.,

$$
\sup_{\Omega \in \mathcal{H}} |\ell_n(\Omega) - \ell(\Omega)| \xrightarrow{p} 0.
$$

If so, then since $\ell_n(\tilde{\Omega}) = \sup_{\Omega \in \mathcal{H}} \ell_n(\Omega) + o_p(1)$, then based on Theorem 2.12 in [40], we have $\|\tilde{\Omega} - \Omega^0\| \xrightarrow{p} 0$. We verify these conditions in the following steps:

**Step 1** convexity and unique maximum of the likelihood function $\ell$. Denote by $\Omega^*$ the true value of $\Omega$ in the corresponding true model. Notice that the model may be characterized through corresponding hazard functions as

$$
d\Lambda^*_t = dH_t\Theta_s(\Omega), \quad s = 0, 1,
$$

which are all functionals that depend on the processes $H(\cdot), z(\cdot)$ on $[0, t]$. Furthermore, let $F^*_t$ be the cumulative incidence function for the outcome $S = s$, and $R_t$ be the survival function in presence of censoring, respectively. Note that $dF^*_t = R_t d\Lambda^*_t, s = 0, 1$. Therefore, we can rewrite the true likelihood as

$$
\ell(\Omega, \Omega^*) = E \int_0^T \sum_{s=0}^1 \left[ \log d\Lambda^*_x dF^*_x^{**} - R^*_t d\Lambda^*_t \right],
$$

where $R^*$ and $F^*$ denote the corresponding true quantities respectively, and expectation is taken with respect to the covariate process $Z_t$.

Now let us consider the negative true Kullback-Leibler distance, i.e.,

$$
D = \ell(\Omega, \Omega^*) - \ell(\Omega^*, \Omega^*).
$$

We have

$$
D = E \int_0^T \sum_{s=0}^1 \left[ \log \frac{d\Lambda^*_x}{d\Lambda^{**}_x} - \left( \frac{d\Lambda^*_x}{d\Lambda^{**}_x} - 1 \right) \right] dF^*_t = E \int_0^T \sum_{s=0}^1 \psi(\frac{d\Lambda^*_x}{d\Lambda^{**}_x}) dF^*_t,
$$
where \( v(x) = \log x - (x - 1) \), a non-positive convex function for any \( x \) except \( x = 1 \), which is the unique maximizer of \( v(x) \). Therefore, \( D \) has a unique maximum when \( d\Lambda_t^s = d\Lambda_t^{s*}, s = 0, 1 \) uniformly.

**Step 2** identifiability condition.

Since \( \Lambda \) is assumed to be a differentiable functional of \( H \), so is the likelihood function \( \ell(\Omega) \). Step 1 suggests that \( \Omega^* = \arg\max_{\Omega \in \mathcal{H}} \ell(\Omega) \) is unique, i.e., the model \( \ell(\Omega, \Omega^*) \) is identifiable so that \( \Lambda = \Lambda^* \) uniformly over \( \Omega \) implies \( \Omega = \Omega^* \). Therefore, based on Lemma 14.3 of [40], we have \( \liminf_{n \to \infty} \ell(\Omega_n) \geq \ell(\Omega^*) \), i.e., the identifiability condition is satisfied.

**Step 3** uniform convergence condition.

We assume that \( \Omega \) is in the class of functions of bounded variation with integrable envelope, which in particular implies that \( H \) is bounded. Therefore, \( \mathcal{H} \) belongs to a Glivenko-Cantelli class, whose \( \varepsilon \)-entropy with bracketing number is bounded by \( A/\varepsilon \), where \( A \) is some constant. Then by the assumption of continuity of the functionals \( \Lambda \) and \( \ell(\Omega) \), and the integrability of the envelope of \( \Omega \), the integrand in \( \ell(\Omega) \) is also Glivenko-Cantelli by the preservation theorems. Therefore, we may apply the uniform law of large numbers for the empirical process counterpart of \( D \) and \( \ell \) as

\[
D_n = \ell_n(\Omega, \Omega^*) - \ell_n(\Omega^*, \Omega^*)
\]

where

\[
\ell_n(\Omega, \Omega^*) = n^{-1} \sum_{i=1}^{n} \sum_{s=0}^{1} \left\{ \int_0^\tau \left[ \log dH_x + \log \Theta_{si}(x; \Omega) \right] dN_{si}(x) - Y_i(x) \Theta_{si}(x; \Omega) dH_x \right\},
\]
such that

\[
\sup_{\Omega \in \mathcal{H}} |D_n(\Omega) - D(\Omega)| \overset{p}{\to} 0
\]

\[
\sup_{\Omega \in \mathcal{H}} |\ell_n(\Omega) - \ell(\Omega)| \overset{p}{\to} 0.
\]

This completes the verification of uniform convergence condition.

**B.2.3 Proof of Theorem II.3**

We prove Theorem II.3 largely based on the martingale representation of score functions (2.17) and (2.16). Let \( U(\Omega) = (U_{\beta}, U_{H_t}) \) be the score, and the proposed NPMLE \( \tilde{\Omega} \) solves the score equation \( U(\Omega) = 0 \). By Theorem 3.3.1 of [62]

\[
n^{1/2}(\tilde{\Omega} - \Omega) = n^{-1/2}(\tilde{I}^0)^{-1}U(\Omega^0) + o_p(1),
\]

where \( \Omega^0 = (\beta^0, H_t^0) \), and \( \tilde{I}^0 \) is the expected single-observation information matrix evaluated at \( \Omega^0 \) and assumed to be positive definite. Based on the martingale representation of \( U(\Omega) \), and the fact that \( N_{1i}(\cdot) \) and \( N_{2i}(\cdot) (i = 1, \ldots, n) \) are orthogonal, it follows that \( n^{-1/2}U(\Omega) \) converges weakly to a zero-mean Gaussian process with its variance-covariance function characterized by \( \sigma_H^2(s, t; \beta^0, H^0) \), \( \sigma_\beta^2(\beta^0) \) and \( \sigma_{H,\beta}^2(t; \beta^0, H^0) \) as derived below, and is identical to \( \tilde{I}^0 \).

The predictable variation process for score equation \( U_{H_t} \) (equation 2.17) is

\[
n^{-1} \sum_{i=1}^n \int_0^\tau \sum_{s=0}^1 \varepsilon^2_{si}(u, t; \beta, H)Y_i(u)\Theta_{si}(u; \beta, H) dH_u,
\]

which converges to a zero-mean Gaussian process with covariate function as \( n \to \infty \)

\[
\sigma_H^2(s, t; \beta^0, H^0) = \int_0^\tau \sum_{s=0}^1 \varepsilon_s(u, s; \beta, H)\varepsilon_s(u, t; \beta, H)P(T \geq u)\Theta_s(u; \beta, H) dH_u,
\]

for \( t, s \in [0, \tau] \).
Similarly, \( n^{-1/2}U_\beta \) is a martingale that converges to a zero-mean Gaussian process with covariance \( \sigma^2_\beta \) as \( n \to \infty \)
\[
\sigma^2_\beta(\beta^0) = \int_0^\tau \sum_{s=0}^1 \Theta^2_{s,\beta}(u; \beta, H) \Theta_s(u; \beta, H) P(T \geq u) dH_u,
\]
and \( n^{-1/2}U(H_t^0, \beta^0) \), for some \( t \), is a martingale and converges as \( n \to \infty \) to a zero-mean Gaussian process with deterministic covariance function
\[
\sigma^2_{H,\beta}(t; \beta_0, H) = \int_0^\tau \sum_{s=0}^1 \varepsilon_s(u; t; \beta, H) \Theta_s(u; \beta, H) P(T \geq u) dH_u.
\]

The variance-covariance function may be consistently estimated by \( n^{-1} \tilde{I}_n \), the observed information matrix evaluated at \( \tilde{H}, \tilde{\beta} \).

### B.2.4 Observed Information Matrix

The information matrix for the parameter set \( \tilde{\Omega} = \{d\tilde{H}, \tilde{\beta}\} \) can be obtained explicitly as follows. Suppose failure times occur at \( t_1 < \cdots < t_l < \cdots < t_j < \cdots < t_J \).

\[
I_{dH_j,dH_j} = \sum_{i=1}^n \sum_{s=0}^1 \left\{ \frac{dN_{si}(t_j)}{dH_j^2} + Y_i(t_j) \left[ \Theta_{si,H}(t_j; \beta, H) + \int_{t_j}^{\tau} \Theta_{si,HH}(x; \beta, H) dH_x \right] \right\} - \\
\sum_{i=1}^n \sum_{s=0}^1 Y_i(t_j) \int_{t_j}^{\tau} \left[ \frac{\Theta_{si,HH}(x; \beta, H)}{\Theta_{si}(x; \beta, H)} - \frac{\Theta^2_{si,H}(x; \beta, H)}{\Theta^2_{si}(x; \beta, H)} \right] dN_{si}(x),
\]

\[
I_{dH_j,dH_i} = \sum_{i=1}^n \sum_{s=0}^1 Y_i(t_i) \left[ \Theta_{si,H}(t_i; \beta, H) I(t_i < t_j) + \int_{t_i \vee t_j}^{\tau} \Theta_{si,HH}(x; \beta, H) dH_x \right] - \\
\sum_{i=1}^n \sum_{s=0}^1 Y_i(t_i) \int_{t_i \vee t_j}^{\tau} \left\{ \left[ \frac{\Theta_{si,HH}(x; \beta, H)}{\Theta_{si}(x; \beta, H)} - \frac{\Theta^2_{si,H}(x; \beta, H)}{\Theta^2_{si}(x; \beta, H)} \right] dN_{si}(x) \right\},
\]

\[
I_{\beta dH_j} = \sum_{i=1}^n \sum_{s=0}^1 Y_i(t_j) \left[ \Theta_{si,\beta}(t_j; \beta, H) - \int_{t_j}^{\tau} \Theta_{si,H\beta}(x; \beta, H) dH_x \right] - \\
\sum_{i=1}^n \sum_{s=0}^1 Y_i(t_j) \int_{t_j}^{\tau} \left[ \frac{\Theta_{si,\beta}(x; \beta, H)}{\Theta_{si}(x; \beta, H)} - \frac{\Theta_{si,H}(x; \beta, H)}{\Theta^2_{si}(x; \beta, H)} \right] dN_{si}(x),
\]

\[
I_{\beta} = \sum_{i=1}^n \sum_{s=0}^1 \int_0^\tau Y_i(x) \Theta_{si,\beta}(x; \beta, H) dH_x - \left[ \frac{\Theta_{si,\beta}(x; \beta, H)}{\Theta_{si}(x; \beta, H)} - \frac{\Theta^2_{si,\beta}(x; \beta, H)}{\Theta^2_{si}(x; \beta, H)} \right] dN_{si}(x),
\]
where \( \Theta_{si,H}(t; \beta, H) = Y_i(t^*) \partial \Theta_{si,H}(t; \beta, H) / \partial dH^* \), \( \Theta_{si,\beta}(t; \beta, H) = \partial \Theta_{si,\beta}(t; \beta, H) / \partial \beta \), and \( \Theta_{si,H\beta}(t; \beta, H) = Y_i(t^*) \partial \Theta_{si,\beta}(t; \beta, H) / \partial dH^* \).
APPENDIX C

Asymptotic Results for Chapter III

This part provides the technical details of the asymptotic properties. Let \( \| \cdot \|_{l^\infty[0,\xi]} \) denote the supremum norm in \([0,\xi]\), and \( \|w\|_{BV[0,\xi]} \) the total variation of \( w(t) \) in \([0,\xi]\). Also define \( Q = \{ w(t) : \|w\|_{BV[0,\xi]} \leq 1 \} \). Such that \( \hat{H}_t \) may be regarded as a bounded linear functional in \( l^\infty(Q) \), and \( \{ \hat{\beta} - \beta^0, \hat{H}_t - H^0_t \} \) a random element in the metric space \( \mathcal{R}^p \times l^\infty(Q) \), where \( p \) is the dimension of \( \beta^0 \). We denote \( \mathcal{H} \) as the compact convex set in the metric space \( \mathcal{R}^p \times l^\infty(Q) \) in which \( \Omega^0 \) contains.

C.1 Proof of Theorem III.1

To establish the consistency result, i.e., \( \| \hat{H}_t - H^0_t \|_{l^\infty(Q)} \overset{p}{\to} 0 \) and \( |\hat{\beta} - \beta^0| \overset{p}{\to} 0 \), in addition to conditions 1-3, we want to verify the following conditions:

1. identifiability condition: suppose that for any sequence \( \Omega_n \in \mathcal{H} \), the compact convex set in the metric space \( \mathcal{R}^p \times l^\infty(Q) \), \( \liminf_{n \to \infty} \ell(\Omega_n) \geq \ell(\Omega^0) \) implies \( \|\Omega_n - \Omega^0\| \overset{p}{\to} 0 \).

2. uniform convergence condition: for any \( \Omega \in \mathcal{H} \) we have uniform convergence, i.e.,

\[
\sup_{\Omega \in \mathcal{H}} |\ell_n(\Omega) - \ell(\Omega)| \overset{p}{\to} 0.
\]
If so, then since \( \ell_n(\hat{\Omega}) = \sup_{\Omega \in \mathcal{H}} \ell_n(\Omega) + o_p(1) \), then based on Theorem 2.12 in [40], we have \( \|\hat{\Omega} - \Omega^0\| \overset{p}{\to} 0 \). We verify these conditions in the following steps:

**Step 1** convexity and unique maximum of the likelihood function \( \ell \).

Denote \( \Omega^* \) is the “true” value of \( \Omega \) in the corresponding “true” model for the semicompeting risks problem. Notice that the model may be characterized through corresponding hazard functions as

\[
d\Lambda^1_t = dH_t\Theta_1(\Omega); \quad d\Lambda^2_t = dH_t\Theta_2(\Omega); \quad d\tilde{\Lambda}^2_t = dH_t\tilde{\Theta}_2(\Omega),
\]

which are all functionals that depend on the processes \( H(\cdot), z(\cdot) \) on \([0, t]\). Furthermore, let \( F^1_t \) and \( F^2_t \) be the cumulative incidence function for the observed failures for the first occurrence of event type 1 and 2, \( \tilde{F}^2_t \) be the cumulative incidence function for event type 2 with prior type 1 event observed, \( R_t \) and \( \tilde{R}_t \) be the survival function for the first and second occurring events in presence of censoring, respectively. Note that \( dF^k_t = R_t d\Lambda^k_t, k = 1, 2, \) and \( d\tilde{F}_t = \tilde{R}_t d\tilde{\Lambda}^2_t. \)

Therefore, we can write the “true” likelihood as

\[
\ell(\Omega, \Omega^*) = E \int_0^\xi \left\{ \sum_{k=1}^2 \left[ \log d\Lambda^k_t dF^k_t - R^*_t d\Lambda^k_t \right] + \log d\tilde{\Lambda}^2_t d\tilde{F}^2_t - R^*_t d\tilde{\Lambda}^2_t \right\},
\]

where \( R^* \) and \( F^* \) denote the corresponding true quantities respectively, and expectation is taken with respect to the covariate process \( Z_t \).

Now let us consider the negative “true” Kullback-Leibler distance, i.e.,

\[
D = \ell(\Omega, \Omega^*) - \ell(\Omega^*, \Omega^*),
\]

such that for a semicompeting risks problem, we have

\[
D = E \int_0^\xi \left\{ \sum_{k=1}^2 \left[ \log \left( \frac{d\Lambda^k_x}{d\Lambda^{k*}_x} \right) - \left( \frac{d\Lambda^k_x}{d\Lambda^{k*}_x} - 1 \right) \right] dF^k_t + \left[ \log \left( \frac{d\tilde{\Lambda}^2_x}{d\tilde{\Lambda}^{2*}_x} \right) - \left( \frac{d\tilde{\Lambda}^2_x}{d\tilde{\Lambda}^{2*}_x} - 1 \right) \right] d\tilde{F}^2_t \right\}
\]

\[
= E \int_0^\xi \left\{ \sum_{k=1}^2 \left( \log \frac{d\Lambda^k_x}{d\Lambda^{k*}_x} + v(\frac{d\Lambda^k_x}{d\Lambda^{k*}_x})dF^k_t + \frac{d\Lambda^k_x}{d\Lambda^{k*}_x}d\tilde{F}^2_t \right) \right\},
\]
where \( v(x) = \log x - (x - 1) \), a non-positive convex function for any \( x \) with a unique maximum of 0 at \( x = 1 \). Therefore, \( D \) has a unique maximum when \( d\Lambda_t^k = d\Lambda_t'^k, k = 1, 2 \) and \( d\tilde{\Lambda}_t^2 = d\tilde{\Lambda}_t'^2 \) uniformly. Under an identifiable model this translates into the unique maximum of \( D \) at \( \Omega^* \).

**Step 2** identifiability condition.

Since \( \Lambda \)s are assumed to be continuous and differentiable functionals of \( H \), so is the likelihood function \( \ell(\Omega) \). Step 1 implies that \( \Omega^* = \arg\max_{\Omega \in \mathcal{H}} \ell(\Omega) \) is unique. We assume the model \( \ell(\Omega, \Omega^*) \) is identifiable in the sense that \( \Lambda = \Lambda^* \) uniformly over \( \Omega \) implies \( \Omega = \Omega^* \) uniformly. Therefore, based on Lemma 14.3 of [40], we have \( \liminf_{n \to \infty} \ell(\Omega_n) \geq \ell(\Omega^*) \), i.e., the identifiability condition is satisfied.

**Step 3** uniform convergence condition.

Condition 1 implies that \( \Omega \) is in the class of functions of bounded variation with integrable envelope, which in turn implies that \( H_t \) is bounded. Therefore, \( \mathcal{H} \) is a Glivenko-Cantelli class, whose \( \varepsilon \)-entropy with bracketing number is bounded by \( A/\varepsilon \), where \( A \) is some constant. Then by the assumption of continuity of the functionals \( \Lambda \) and \( \ell(\Omega) \), and the integrability of the envelope of \( \Omega \), the integrand in \( \ell(\Omega) \) is also Glivenko-Cantelli by the preservation theorems. Therefore, we may apply the uniform law of large numbers for the empirical process counterpart of \( D \) and \( \ell \) as

\[
D_n = \ell_n(\Omega, \Omega^*) - \ell_n(\Omega^*, \Omega^*)
\]
and \( \ell_n(\Omega, \Omega^*) = \ell_{1,n}(\Omega, \Omega^*) + \ell_{2,n}(\Omega, \Omega^*) \), where

\[
\ell_{1,n}(\Omega, \Omega^*) = n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{2} \left\{ \int_{0}^{\xi} [\log dH_x + \log \Theta_{ki}(x; \Omega)] dN_{ki}(x) - Y_{i}(x)\Theta_{ki}(x; \Omega) dH_x \right\}
\]

\[
\ell_{2,n}(\Omega, \Omega^*) = n^{-1} \sum_{i=1}^{n} \left\{ \int_{0}^{\xi} [\log dH_x + \log \tilde{\Theta}_{2i}(x; \Omega)] d\tilde{N}_{2i}(x) - \tilde{Y}_{i}(x)\tilde{\Theta}_{2i}(x; \Omega) dH_x \right\},
\]

such that

\[
\sup_{\Omega \in \mathcal{H}} |D_n(\Omega) - D(\Omega)| \xrightarrow{P} 0
\]

\[
\sup_{\Omega \in \mathcal{H}} |\ell_n(\Omega) - \ell(\Omega)| \xrightarrow{P} 0.
\]

This completes the verification of uniform convergence condition.

### C.2 Proof of Theorem III.2

We prove Theorem III.2 largely based on the martingale representation of score functions (3.9) and (3.8). Let \( U(\Omega) = (U_{\beta}, U_{Ht}) \). The proposed NPMLE \( \hat{\Omega} \) solves the score equation \( U(\hat{\Omega}) = 0 \). By Theorem 3.3.1 of [62]

\[
n^{1/2}(\hat{\Omega} - \Omega) = n^{-1/2}(I^0)^{-1}U(\Omega^0) + o_p(1),
\]

where \( \Omega^0 = (\beta^0, H_t^0) \), and \( I^0 \) is the expected single-observation information matrix evaluated at \( \Omega^0 \) and assumed to be positive definite. Based on the martingale representation of \( U(\Omega) \), and the fact that \( N_{1i}(\cdot), N_{2i}(\cdot) \) and \( \tilde{N}_{2i}(\cdot) \) \((i = 1, \cdots, n)\) are orthogonal, we can show that \( n^{-1/2}U(\Omega) \) converges weakly to a zero-mean Gaussian process with variance-covariance function characterized by \( \sigma_{H}^2(s,t; \beta^0, H^0) \), \( \sigma_{\beta}^2(\beta^0) \) and \( \sigma_{H,\beta}^2(t; \beta^0, H^0) \) as derived below, and is identical to \( I^0 \).

The predictable variation process for the score process \( U_{Ht} \) (equation 3.9) is

\[
n^{-1} \sum_{i=1}^{n} \int_{0}^{\xi} \left[ \sum_{k=1}^{2} \varepsilon_{ki}^2(u, t; \beta, H)Y_i(u)\Theta_{ki}(u; \beta, H) dH_u + I(u \leq t)\tilde{Y}_i(u)\tilde{\Theta}_{2i}(\beta) dH_u \right],
\]
which converges to a zero-mean Gaussian process with covariate function as $n \to \infty$

\[
\sigma_H^2(s, t; \beta^0, H^0) = \int_0^\xi \sum_{k=1}^2 \varepsilon_k(u, s; \beta, H)\varepsilon_k(u, t; \beta, H)P(T \geq u)\Theta_k(u; \beta, H)dH_u + I(u \leq s)I(u \leq t)P(T^* \geq u)\tilde{\Theta}_2(\beta)dH_u,
\]

for $t, s \in [0, \xi]$.

Similarly, $n^{-1/2}U_\beta$ is a martingale and converges to a zero-mean Gaussian process with covariance $\sigma_\beta^2$ as $n \to \infty$

\[
\sigma_\beta^2(\beta^0) = \int_0^\xi \left\{ \sum_{k=1}^2 \Theta_{k,\beta}^2(u; \beta, H) \frac{\Theta_k(u; \beta, H)}{\Theta_2(\beta)} P(T \geq u) dH_u + \frac{\tilde{\Theta}_{2,\beta}^2(\beta)}{\Theta_2(\beta)} P(T^* \geq u) dH_u \right\},
\]

and $n^{-1/2}U(H^0_t, \beta^0)$, for some $t$, is a martingale and converges to a zero-mean Gaussian process with deterministic covariance function as $n \to \infty$

\[
\sigma_{H,\beta}^2(t; \beta_0, H) = \int_0^\xi \sum_{k=1}^2 \varepsilon_k(u, t; \beta, H)\Theta_{k,\beta}(u; \beta, H)P(T \geq u)dH_u + I(u \leq t)\tilde{\Theta}_{2,\beta}(\beta)P(T^* \geq u)dH_u,
\]

The variance-covariance function may be consistently estimated by $n^{-1}{\mathcal{I}}_n$, the observed information matrix evaluated at $\hat{H}, \hat{\beta}$.

**C.3 Information Matrix**

The observed information matrix $\mathcal{I}_n$ for the parameter set $\hat{\Omega} = (\hat{\beta}, \{dH\})$ can be obtained explicitly as follows (where failure times are at $t_1 < \cdots < t_i < \cdots < t_j < \xi$):
\[ I_{dH_j} = \sum_{i=1}^{n} \left\{ \sum_{k=1}^{2} dN_{ki}(t_j) + d\tilde{N}_{2i}(t_j) \right\} - \sum_{i=1}^{n} \frac{Y_i(t_j)}{2} \sum_{k=1}^{2} \frac{dN_{ki}(x)}{dH_j^2} \]

\[ I_{dH_j} = \sum_{i=1}^{n} \sum_{k=1}^{2} Y_i(t_j) \Theta_{ki,HH}(t_j; \Omega) I(t_i < t_j) - \sum_{i=1}^{n} \sum_{k=1}^{2} \Theta_{ki,HH}(t_j; \Omega) dN_{ki}(x) \]

\[ I_{\beta dH_j} = \sum_{i=1}^{n} \sum_{k=1}^{2} \left\{ Y_i(t_j) \Theta_{ki,\beta}(t_j; \Omega) + \tilde{Y}_2i(t_j) \tilde{\Theta}_{2i,\beta}(\beta) \right\} - \sum_{i=1}^{n} \sum_{k=1}^{2} \Theta_{ki,\beta}(t_j; \Omega) dN_{ki}(x) + \Theta_{ki,\beta}(t_j; \Omega) dH_x \]

\[ I_{\beta \beta} = \sum_{i=1}^{n} \int_{0}^{\xi} \sum_{k=1}^{2} \int_{t_j^+}^{t_j^-} \left\{ \frac{\Theta_{ki,\beta}(t_j; \Omega)}{\Theta_{ki,H}(t_j; \Omega)} - \frac{\Theta_{ki,\beta}(t_j; \Omega)}{\Theta_{ki,H}(t_j; \Omega)} \right\} dN_{ki}(x) + \frac{\Theta_{ki,\beta}(\beta)}{\Theta_{2i}(\beta)} - \frac{\tilde{\Theta}_{2i,\beta}(\beta)}{\tilde{\Theta}_{2i}(\beta)} \]

where \( \Theta_{ki,HH}(t; \Omega) = Y_i(t_*) \partial \Theta_{ki,H}(t; \Omega)/\partial dH_x, \Theta_{ki,\beta}(t; \Omega) = \partial \Theta_{ki,\beta}(t; \Omega)/\partial \beta, \) and \( \Theta_{ki,\beta}(t; \Omega) = Y_i(t_*) \partial \Theta_{ki,\beta}(t; \Omega)/\partial dH_x. \)
APPENDIX D

Asymptotic Results for Chapter IV

This part provides the technical details of the asymptotic properties. Let \( \| \cdot \|_{l^\infty[0,\xi]} \) denote the supremum norm in \([0,\xi]\), and \( \| w \|_{BV[0,\xi]} \) the total variation of \( w(t) \) in \([0,\xi]\). Also define \( \mathcal{Q} = \{ w(t) : \| w \|_{BV[0,\xi]} \leq 1 \} \). Such that \( \hat{H}_t \) may be regarded as a bounded linear functional in \( l^\infty[\mathcal{Q}] \), and \( \{ \hat{\beta} - \beta_0, \hat{H}_t - H_0^t \} \) a random element in the metric space \( \mathcal{R}^p \times l^\infty(\mathcal{Q}) \), where \( p \) is the dimension of \( \beta_0 \). We denote \( \mathcal{H} \) as the compact convex set in the metric space \( \mathcal{R}^p \times l^\infty(\mathcal{Q}) \) in which \( \Omega_0 \) contains.

D.1 Proof of Theorem IV.1

To establish the consistency result, i.e., \( \| \hat{H}_t - H_0^t \|_{l^\infty(\mathcal{Q})} \overset{p}{\to} 0 \) and \( |\hat{\beta} - \beta_0| \overset{p}{\to} 0 \), in addition to conditions 1-3, we want to verify the following conditions:

1. identifiability condition: suppose that for any sequence \( \Omega_n \in \mathcal{H} \), the compact convex set in the metric space \( \mathcal{R}^p \times l^\infty(\mathcal{Q}) \), \( \liminf_{n \to \infty} \ell(\Omega_n) \geq \ell(\Omega_0) \) implies \( \| \Omega_n - \Omega_0 \| \overset{P}{\to} 0 \).

2. uniform convergence condition:

\[
\sup_{\Omega \in \mathcal{H}} |\ell_n(\Omega) - \ell(\Omega)| \overset{P}{\to} 0.
\]

If so, then since \( \ell_n(\hat{\Omega}) = \sup_{\Omega \in \mathcal{H}} \ell_n(\Omega) + o_p(1) \), then based on Theorem 2.12 in [40], we have \( \| \hat{\Omega} - \Omega_0 \| \overset{P}{\to} 0 \). We verify these conditions in the following steps:
Step 1 convexity and unique maximum of the likelihood function $\ell$.

Denote by $\Omega^*$ the “true” value of $\Omega$. Notice that the model may be characterized through corresponding hazard functions as

$$d\Lambda_1^t = dH_t^1(\Omega); \quad d\Lambda_2^t = dH_t^2(\Omega); \quad d\tilde{\Lambda}_2^t = dH_t^2(\Omega),$$

which are all functionals that depend on the processes $H(\cdot), z(\cdot)$ on $[0,t]$. Furthermore, let $F^1_t$ and $F^2_t$ be the cumulative incidence function for the observed failures for the first occurrence of event type 1 and 2, $\tilde{F}^2_t$ be the cumulative incidence function for event type 2 (death) with prior type 1 event (progression) observed, $R_t$ and $\tilde{R}_t$ be the survival function for the first and second occurring events in presence of censoring, respectively. Note that $dF^k_t = R_t d\Lambda^k_t, k = 1, 2$, and $d\tilde{F}_t = \tilde{R}_t d\tilde{\Lambda}^2_t$. Therefore, we can rewrite the “true” likelihood as

$$\ell(\Omega, \Omega^*) = E \int_0^\xi \left\{ \sum_{k=1}^2 \left[ \log d\Lambda^k_t dF^k_t - R^* d\Lambda^k_t \right] + \log d\tilde{\Lambda}^2_t d\tilde{F}^2_t - R^* d\tilde{\Lambda}^2_t \right\},$$

where $R^*$ and $F^*$ denote the corresponding true quantities respectively, and expectation is taken with respect to the covariate process $X_t$.

Now let us consider the negative “true” Kullback-Leibler distance, i.e.,

$$D = \ell(\Omega, \Omega^*) - \ell(\Omega^*, \Omega^*),$$

such that for a semicompeting risks problem, we have

$$D = E \int_0^\xi \left\{ \sum_{k=1}^2 \left[ \log \frac{d\Lambda^k_t}{d\Lambda^k_{\Omega^*}} - \left( \frac{d\Lambda^k_t}{d\Lambda^k_{\Omega^*}} - 1 \right) \right] dF^k_t + \left[ \log \frac{d\tilde{\Lambda}^2_t}{d\tilde{\Lambda}^2_{\Omega^*}} - \left( \frac{d\tilde{\Lambda}^2_t}{d\tilde{\Lambda}^2_{\Omega^*}} - 1 \right) \right] d\tilde{F}^2_t \right\}$$

$$= E \int_0^\xi \left\{ \sum_{k=1}^2 v \left( \frac{d\Lambda^k_t}{d\Lambda^k_{\Omega^*}} \right) dF^k_t + v \left( \frac{d\tilde{\Lambda}^2_t}{d\tilde{\Lambda}^2_{\Omega^*}} \right) d\tilde{F}^2_t \right\},$$

where $v(x) = \log x - (x - 1)$, a non-positive convex function having the unique maximizer $x = 1$ at $v = 0$. Therefore, $D$ has a unique maximum when $d\Lambda^k_t = d\Lambda^k_{\Omega^*}, k = 1, 2$ and $d\tilde{\Lambda}^2_t = d\tilde{\Lambda}^2_{\Omega^*}$ uniformly.
Step 2 identifiability condition.

Since $\Lambda$s are assumed to be continuous and differentiable functionals of $H$, so is the likelihood function $\ell(\Omega)$. Step 1 implies that $\Omega^* = \arg\max_{\Omega \in \mathcal{H}} \ell(\Omega)$ is unique. We assume the model $\ell(\Omega, \Omega^*)$ is identifiable in the sense that $\Lambda = \Lambda^*$ uniformly over $\Omega$ implies $\Omega = \Omega^*$ uniformly. Therefore, based on Lemma 14.3 of [40], we have $\liminf_{n \to \infty} \ell(\Omega_n) \geq \ell(\Omega^*)$, i.e., the identifiability condition is satisfied.

Step 3 uniform convergence condition.

Condition 1 implies that $\Omega$ is in the class of functions of bounded variation with integrable envelope, which in turn implies that $H_t$ is bounded. Therefore, $\mathcal{H}$ is a Glivenko-Cantelli class, whose $\varepsilon$-entropy with bracketing number is bounded by $A/\varepsilon$, where $A$ is some constant. Then by the assumption of continuity of the functionals $\Lambda$ and $\ell(\Omega)$, and the integrability of the envelope of $\Omega$, the integrand in $\ell(\Omega)$ is also Glivenko-Cantelli by the preservation theorems. Therefore, we may apply the uniform law of large numbers for the empirical process counterpart of $D$ and $\ell$ as

$$D_n = \ell_n(\Omega, \Omega^*) - \ell_n(\Omega^*, \Omega^*)$$

and $\ell_n(\Omega, \Omega^*) = \ell_{1,n}(\Omega, \Omega^*) + \ell_{2,n}(\Omega, \Omega^*)$, where

$$\ell_{1,n}(\Omega, \Omega^*) = n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{2} \left\{ \int_0^\xi [\log dH_x + \log \Upsilon_{ki}(x; \Omega)]dN_{ki}(x) - Y_i(x)\Upsilon_{ki}(x; \Omega)dH_x \right\}$$

$$\ell_{2,n}(\Omega, \Omega^*) = n^{-1} \sum_{i=1}^{n} \left\{ \int_0^\xi [\log dH_x + \log \tilde{\Upsilon}_{2i}(x; \Omega)]d\tilde{N}_{2i}(x) - \tilde{Y}_i(x)\tilde{\Upsilon}_{2i}(x; \Omega)dH_x \right\}.$$
such that

\[
\sup_{\Omega \in \mathcal{H}} |D_n(\Omega) - D(\Omega)| \xrightarrow{p} 0
\]

\[
\sup_{\Omega \in \mathcal{H}} |\ell_n(\Omega) - \ell(\Omega)| \xrightarrow{p} 0.
\]

This completes the verification of uniform convergence condition.

D.2 Proof of Theorem IV.2

We prove Theorem IV.2 largely based on the martingale representation of the score functions (4.12) and (4.11). Denote \( \gamma^T = (\beta^T, \alpha^T) \), and suppose \( U(\Omega) = (U_\gamma, U_{H_t}) \), the proposed NPMLE \( \hat{\Omega} \), solves the score equation \( U(\Omega) = 0 \). By Theorem 3.3.1 of [62]

\[
n^{1/2}(\hat{\Omega} - \Omega) = n^{-1/2}(I^0)^{-1}U(\Omega^0) + o_p(1),
\]

where \( \Omega^0 = (\gamma^0, H^0_t) \), and \( I^0 \) is the expected single-observation information matrix evaluated at \( \Omega^0 \) and assumed to be positive definite. Based on the martingale representation of \( U(\Omega) \), and the fact that \( N_{1i}(\cdot), N_{2i}(\cdot) \) and \( \tilde{N}_{2i}(\cdot) \) \((i = 1, \cdots, n)\) are orthogonal, we can show that \( n^{-1/2}U(\Omega) \) converges weakly to a zero-mean Gaussian process with its variance-covariance function characterized by \( \sigma_H^2(s, t; \gamma^0, H^0) \), \( \sigma_\gamma^2(\gamma^0) \), and \( \sigma_{H, \gamma}^2(t; \gamma^0, H^0) \) as derived below, and is identical to \( I^0 \).

The predictable variation process for score equation \( U_{H_t} \) (equation 4.12) is

\[
n^{-1} \sum_{i=1}^n \int_0^\xi \left[ \sum_{k=1}^2 \varepsilon_{ki}(u, t; \gamma, H) Y_i(u) \Upsilon_{ki}(u; \gamma, H) dH_u + \tilde{\varepsilon}_{2i}(u, t; \gamma, H) \tilde{Y}_i(u) \tilde{\Upsilon}_{2i}(\gamma) dH_u \right],
\]

which converges to a zero-mean Gaussian process with covariates function as \( n \to \infty \)

\[
\sigma_H^2(s, t; \gamma^0, H^0) = \int_0^\xi \sum_{k=1}^2 \varepsilon_k(u, s; \gamma, H) \varepsilon_k(u, t; \gamma, H) P(T \geq u) \Upsilon_k(u; \gamma, H) dH_u \\
+ \tilde{\varepsilon}_2(u, s; \gamma, H) \tilde{\varepsilon}_2(u, t; \gamma, H) P(T^* \geq u) \tilde{\Upsilon}_2(u; \gamma, H) dH_u,
\]
Similarly, \( n^{-1/2}U_\gamma \) is a martingale and converges to a zero-mean Gaussian process with covariance \( \sigma_\gamma^2 \) as \( n \to \infty \)
\[
\sigma_\gamma^2(\gamma^0) = \int_0^\xi \left\{ \sum_{k=1}^2 \frac{Y_k^2(u; \gamma, H)}{\gamma_k(u; \gamma, H)} P(T \geq u) dH_u + \frac{\tilde{\gamma}_2(u; \gamma, H)}{\gamma_2(u; \gamma, H)} P(T^* \geq u) dH_u \right\},
\]
and \( n^{-1/2}U(H^0_t, \gamma^0) \), for some \( t \), is a martingale and converges to a zero-mean Gaussian process with deterministic covariate function as \( n \to \infty \)
\[
\sigma_{H,\gamma}^2(t; \gamma_0, H) = \int_0^\xi \sum_{k=1}^2 \varepsilon_k(u, t; \gamma, H)Y_k(u; \gamma, H) P(T \geq u) dH_u
\]
\[+ \tilde{\varepsilon}_2(u, t; \gamma, H)\tilde{Y}_2(u; \gamma, H) P(T^* \geq u) dH_u,
\]

The variance-covariance function may be consistently estimated by \( n^{-1}\mathcal{I}_n \), the observed information matrix evaluated at \( \hat{H}, \hat{\gamma} \).

### D.3 Information Matrix

The observed information matrix \( \mathcal{I}_n \) for the parameter set \( \hat{\Omega} = (\hat{\gamma}, \{dH\}) \) can be obtained explicitly as follows. Let \( t_1 < \cdots < t_l < \cdots < t_j < \cdots < t_J \) be the distinct points where failure times occur.

\[
\mathcal{I}_{dH_j, dH_j} = \sum_{i=1}^n \left\{ \sum_{k=1}^2 \frac{dN_{ki}(t_j) + d\tilde{N}_{2i}(t_j)}{dH_j^2} + \sum_{k=1}^2 Y_i(t_j)Y_{ki,H}(t_j; \Omega) + \tilde{Y}_i(t_j)\tilde{Y}_{2i,H}(t_j; \Omega) \right\}
\]
\[+ \sum_{i=1}^n \left\{ \int_{t^+_j}^\xi \left[ \frac{Y_{ki,H}(x; \Omega)}{\gamma_{ki}(x; \Omega)} - \frac{\gamma_{ki}^{\otimes 2}(x; \Omega)}{\gamma_{ki}^2(x; \Omega)} \right] dN_{ki}(x) - Y_{ki,H}(x; \Omega)dH_x \right\}
\]
\[- \sum_{i=1}^n \left\{ \int_{t^+_j}^\xi \left[ \frac{\tilde{Y}_{2i,H}(x; \Omega)}{\gamma_{2i}(x; \Omega)} - \frac{\tilde{Q}_{2i}^{\otimes 2}(x; \Omega)}{\gamma_{2i}^2(x; \Omega)} \right] d\tilde{N}_{2i}(x) - \tilde{Y}_{2i,H}(x; \Omega)dH_x \right\},
\]

\[
\mathcal{I}_{dH_j, dH_l} = \sum_{i=1}^n \left[ \sum_{k=1}^2 Y_i(t_j)Y_{ki,H}(t_j; \Omega) I(t_l < t_j) + \tilde{Y}_i(t_j)\tilde{Y}_{2i,H}(t_j; \Omega) I(t_l < t_j) \right]
\]
\[- \sum_{i=1}^n \left\{ \int_{t^+_j \lor t^+_l}^\xi \left[ \frac{Y_{ki,H}(x; \Omega)}{\gamma_{ki}(x; \Omega)} - \frac{\gamma_{ki}^{\otimes 2}(x; \Omega)}{\gamma_{ki}^2(x; \Omega)} \right] dN_{ki}(x) \right\} - Y_{ki,H}(x; \Omega)dH_x
\]
\[- \sum_{i=1}^n \left\{ \int_{t^+_j \lor t^+_l}^\xi \left[ \frac{\tilde{Y}_{2i,H}(x; \Omega)}{\gamma_{2i}(x; \Omega)} - \frac{\tilde{Q}_{2i}^{\otimes 2}(x; \Omega)}{\gamma_{2i}^2(x; \Omega)} \right] d\tilde{N}_{2i}(x) \right\} - \tilde{Y}_{2i,H}(x; \Omega)dH_x,
\]
\[ \mathcal{I}_{\gamma dH_j} = \sum_{i=1}^{n} \left\{ \sum_{k=1}^{2} Y_i(t_j) Y_{ki,\gamma}(t_j; \Omega) + \bar{Y}_{2i}(t_j) \bar{Y}_{2i,\gamma}(u; \Omega) \right\} \\
- \sum_{i=1}^{n} \sum_{k=1}^{2} Y_i(t_j) \int_{t_j}^{\xi} \left\{ \left[ \frac{\gamma_{ki,H\gamma}(x; \Omega)}{\gamma_{ki}(x; \Omega)} - \frac{\gamma_{ki,H}(x; \Omega) \gamma_{ki,\gamma}(x; \Omega)}{\gamma_{ki}^2(x; \Omega)} \right] dN_{ki}(x) - \gamma_{ki,H\gamma}(x; \Omega) dH_x \right\} \\
- \sum_{i=1}^{n} \bar{Y}_i(t_j) \int_{t_j}^{\xi} \left\{ \left[ \frac{\bar{Y}_{2i,H\gamma}(x; \Omega)}{\bar{Y}_{2i}(x; \Omega)} - \frac{\bar{Y}_{2i,H}(x; \Omega) \bar{Y}_{2i,\gamma}(x; \Omega)}{\bar{Y}_{2i}^2(x; \Omega)} \right] d\bar{N}_{2i}(x) - \bar{Y}_{2i,H\gamma}(x; \Omega) dH_x \right\}, \]

\[ \mathcal{I}_{\gamma\gamma} = \sum_{i=1}^{n} \int_{0}^{\xi} \left\{ \sum_{k=1}^{2} Y_i(x) Y_{ki,\gamma\gamma}(x; \Omega) dH_x + \bar{Y}_i(x) \bar{Y}_{2i,\gamma\gamma}(x; \Omega) dH_x \right\} \\
- \sum_{i=1}^{n} \int_{0}^{\xi} \left\{ \sum_{k=1}^{2} \left[ \frac{\gamma_{ki,\gamma\gamma}(x; \Omega)}{\gamma_{ki}(x; \Omega)} - \frac{\gamma_{ki,\gamma}^2(x; \Omega)}{\gamma_{ki}^2(x; \Omega)} \right] dN_{ki}(x) + \left[ \frac{\bar{Y}_{2i,\gamma\gamma}(x; \Omega)}{\bar{Y}_{2i}(x; \Omega)} - \frac{\bar{Y}_{2i,\gamma}^2(x; \Omega)}{\bar{Y}_{2i}^2(x; \Omega)} \right] d\bar{N}_{2i}(x) \right\}, \]

where \( \gamma_{ki,HH}(t; \Omega) = Y_i(t_*) \partial \gamma_{ki,H}(t; \Omega) / \partial dH_* \), \( \gamma_{ki,\gamma\gamma}(t; \Omega) = \partial \gamma_{ki,\gamma}(t; \Omega) / \partial \gamma \), and

\( \gamma_{ki,H\gamma}(t; \Omega) = Y_i(t_*) \partial \gamma_{ki,H\gamma}(t; \Omega) / \partial dH_* \).
BIBLIOGRAPHY


