# SEMIPARAMETRIC REGRESSION UNDER LEFT-TRUNCATED AND INTERVAL-CENSORED COMPETING RISKS DATA AND MISSING CAUSE OF FAILURE

Jun Park

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Biostatistics,
Indiana University
April 2020

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

Giorgos Bakoyannis, Ph.D., Co-Chair

Constantin T. Yiannoutsos, Ph.D., Co-Chair

Doctoral Committee

Ying Zhang, Ph.D.

January 7, 2020

Sujuan Gao, Ph.D.

Yiqing Song, M.D., Sc.D.

# DEDICATION

I dedicate this to my beloved family.

# ACKNOWLEDGMENT

I gratefully acknowledge that my research was financially supported by the National Institute of Allergy and Infectious Diseases (NIAID), Eunice Kennedy Shriver National Institute of Child Health & Human Development (NICHD), National Institute on Drug Abuse (NIDA), National Cancer Institute (NCI), the National Institute of Mental Health (NIMH), in accordance with the regulatory requirements of the National Institutes of Health under Award Number U01AI069911 East Africa IeDEA Consortium, and the President's Emergency Plan for AIDS Relief (PEPFAR) through USAID under the terms of Cooperative Agreement No. AID-623-A-12-0001 it is made possible through joint support of the United States Agency for International Development (USAID). The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health, USAID, or the United States Government. Data collection for the Indianapolis-Ibadan Dementia Project was supported by National Institutes of Health grant R01 AG09956.

Over the past few years, I have received remarkable support from many individuals. My journey toward completing a doctoral degree would not have been possible without the support of those individuals.

First and foremost, it is my honor to be the first PhD student of my advisor Dr. Giorgos Bakoyannis. I am deeply indebted to him for his patient guidance, consistent encouragement and mentorship. He has been such a great mentor, colleague, and friend. I also would like to express my sincere appreciation to my co-advisor Dr. Constantin Yiannoutsos for giving me the opportunity to be involved in the Inter-

national epidemiology Databases to Evaluate AIDS research project. I am richer for having their time and energy. I am fortunate to have these two as my mentors.

I would also like to thank my committee members, Dr. Ying Zhang, Dr. Sujuan Gao, and Dr. Yiqing Song for their constructive criticism, guidance, and valuable advice for my research project. Their insight into my research helped me improve this dissertation. I have truly enjoyed the opportunity to work with and learn from these marvelous people.

Besides my dissertation committee members, I thank Dr. Barry Katz, Beverly Musick, Cynthia Maurer, Ann Lyon, George Nitsos, and Jarod Watt. They offered me untiring support whether I have problems or not over the years. Many thanks to my wonderful friends for making great memories together.

I cannot forget to give the biggest thanks to my family for unconditional love and support. I am especially grateful to my parents for their dedication and belief in my abilities. Special thanks should also go to my sister and brother's in law for unparalleled support and giving me a great pleasure to meet my sweet niece. I also thank my aunts for unwavering support. I would not have been able to achieve my goal without my whole family. Lastly, I would like to thank my loving girlfriend who has been my greatest support and my strongest motivation. I am grateful for the joy that she has brought into my life.

Jun Park

# SEMIPARAMETRIC REGRESSION UNDER LEFT-TRUNCATED AND INTERVAL-CENSORED COMPETING RISKS DATA AND MISSING CAUSE OF FAILURE

Observational studies and clinical trials with time-to-event data frequently involve multiple event types, known as competing risks. The cumulative incidence function (CIF) is a particularly useful parameter as it explicitly quantifies clinical prognosis. Common issues in competing risks data analysis on the CIF include interval censoring, missing event types, and left truncation. Interval censoring occurs when the event time is not observed but is only known to lie between two observation times, such as clinic visits. Left truncation, also known as delayed entry, is the phenomenon where certain participants enter the study after the onset of disease under study. These individuals with an event prior to their potential study entry time are not included in the analysis and this can induce selection bias. In order to address unmet needs in appropriate methods and software for competing risks data analysis, this thesis focuses the following development of application and methods. First, we develop a convenient and flexible tool, the R package intccr, that performs semiparametric regression analysis on the CIF for interval-censored competing risks data. Second, we adopt the augmented inverse probability weighting method to deal with both interval censoring and missing event types. We show that the resulting estimates are consistent and double robust. We illustrate this method using data from the East-African International Epidemiology Databases to Evaluate AIDS (IeDEA

EA) where a significant portion of the event types is missing. Last, we develop an estimation method for semiparametric analysis on the CIF for competing risks data subject to both interval censoring and left truncation. This method is applied to the Indianapolis-Ibadan Dementia Project to identify prognostic factors of dementia in elder adults. Overall, the methods developed here are incorporated in the R package intccr.

Giorgos Bakoyannis, Ph.D., Co-Chair

Constantin T. Yiannoutsos, Ph.D., Co-Chair

TABLE OF CONTENTS

List of Tables ..... xi

List of Figures ..... xiii

Chapter 1 Introduction ..... 1

Chapter 2 Semiparametric competing risks regression under interval censoring using the R package intccr ..... 5

2.1 Introduction ..... 6

2.2 Methodology ..... 9

2.2.1 Notation ..... 9

2.2.2 Estimation ..... 10

2.3 Simulation ..... 13

2.4 Example ..... 24

2.5 Discussion ..... 34

Chapter 3 Semiparametric regression on cumulative incidence function with interval-censored competing risks data and missing cause of failure ..... 37

3.1 Introduction ..... 38

3.2 Methods ..... 43

3.2.1 Data and model ..... 43

3.2.2 Semiparametric estimation ..... 46

3.2.3 Properties of the proposed estimator ..... 49

3.3 Simulation studies ..... 55

3.4 Analysis of HIV data ..... 69

3.5 Illustration of the R function ciregic_aipw ..... 72

3.6 Discussion ..... 79

Chapter 4 Competing risks regression analysis on the cumulative incidence function for left-truncated and interval-censored data ..... 82

4.1 Introduction ..... 83

4.2 Methods ..... 87

4.2.1 Data and models ..... 87

4.2.2 Semiparametric sieve estimation ..... 89

4.2.3 Practical implementation of the method ..... 93

4.3 Simulation studies ..... 95

4.4 Analysis of dementia data ..... 102

4.5 The R function ciregic_lt ..... 107

4.6 Conclusions ..... 114

Chapter 5 Discussion ..... 117

References ..... 121

Curriculum Vitae

LIST OF TABLES

Table 2.1 Arguments of the function ciregic ..... 17

Table 2.2 Monte Carlo simulation results based on 1,000 replications .... 21

Table 2.3 Computation time (seconds) based on Monte Carlo simulation using 1,000 replications ..... 22

Table 2.4 The arguments of the function predict ..... 32

Table 3.1 Monte Carlo simulation with 1,000 replications when ($\xi_4 = 0$) 58

Table 3.2 Monte Carlo simulation with 1,000 replications when ($\xi_4 = -0.5$) ..... 59

Table 3.3 Monte Carlo simulation with 1,000 replications when ($\xi_4 = -0.1$) ..... 60

Table 3.4 Monte Carlo simulation with 1,000 replications when ($\xi_4 = 0.1$) 61

Table 3.5 Monte Carlo simulation with 1,000 replications when ($\xi_4 = 0.5$) 62

Table 3.6 Descriptive characteristics of the study sample ..... 70

Table 3.7 Covariate effects on the CIF of disengagement from care and death based on the naïve complete case analysis and the proposed AIPW approach ..... 71

Table 3.8 Variables in the data set simdata_aipw ..... 74

Table 3.9 Argument of the function ciregic_aipw ..... 76

Table 4.1 Simulation results of comparison of the proposed method with naïve method for 50% left truncation ..... 98

Table 4.2 Simulation results of comparison of the proposed method with naïve method for 100% left truncation ..... 99
Table 4.3 Descriptive characteristics of the study sample ..... 105
Table 4.4 Covariate effects on the CIF of dementia and death based on the naïve analysis and the proposed method ..... 106
Table 4.5 Variables in the data set longdata_lt ..... 108
Table 4.6 Argument of the function ciregic_lt ..... 109

LIST OF FIGURES

Figure 2.1 Data resphaping scheme of the function dataprep ..... 16

Figure 2.2 The predicted baseline cumulative incidence functions resulted from a simulation study with sample sizes of 200, 400, and 800 ..... 23

Figure 2.3 The predicted cumulative incidence functions for females aged 20 to 50 years, with CD4 count 120 cells/$\mu l$ at ART initiation ..... 29

Figure 3.1 The predicted baseline cumulative incidence functions resulted from a simulation study with sample sizes of 200 and 400 when $\xi_4 = 0$ 64

Figure 3.2 The predicted baseline cumulative incidence functions resulted from a simulation study with sample sizes of 200 and 400 when $\xi_4 = -0.5$ ..... 65

Figure 3.3 The predicted baseline cumulative incidence functions resulted from a simulation study with sample sizes of 200 and 400 when $\xi_4 = -0.1$ ..... 66

Figure 3.4 The predicted baseline cumulative incidence functions resulted from a simulation study with sample sizes of 200 and 400 when $\xi_4 = 0.1$ 67

Figure 3.5 The predicted baseline cumulative incidence functions resulted from a simulation study with sample sizes of 200 and 400 when $\xi_4 = 0.5$ 68

Figure 3.6 The predicted cumulative incidence functions from the naïve analysis for a 30-year-old male patient ..... 72

Figure 3.7 The predicted cumulative incidence functions from the proposed AIPW method for a 30-year-old male patient ..... 73

Figure 3.8 The estimated baseline cumulative incidence function ..... 78

Figure 4.1 The predicted baseline cumulative incidence functions resulted from a simulation study with sample sizes of 250, 500, and 1,000 under a 50% left truncation ..... 100

Figure 4.2 The predicted baseline cumulative incidence functions resulted from a simulation study with sample sizes of 250, 500, and 1,000 under a 100% left truncation ..... 101

Figure 4.3 The predicted covariate-specific cumulative incidence function of dementia and death for individuals without alcohol use and smoking at baseline ..... 104

Figure 4.4 The estimated baseline cumulative incidence functions ..... 113

# Chapter 1

## Introduction

Clinical decision-making benefits from prognostic models. Clinicians often employ clinical characteristics to make treatment plans, and public health professionals use risk prediction of a disease of interest for policy discussions (Wolbers et al., 2009). Identifying prognostic factors relies on individuals' absolute risk of a disease or event of interest. However, various circumstances in cohort studies hinder to estimate prognostic models, such as competing risks, interval censoring, left truncation, and missing event types.

Competing risks data are common in cohort studies and clinical trials, where study participants are at risk of multiple mutually exclusive events or the scientific interest focuses on the first occurring event among multiple endpoints (Kalbfleisch & Prentice, 2011; Putter et al., 2007; Bakoyannis & Touloumi, 2012). The basic identifiable quantities in competing risks data framework are the cumulative incidence function (CIF) and the cause-specific hazard (CSH) function (Kalbfleisch & Prentice, 2011; Putter et al., 2007; Bakoyannis & Touloumi, 2012; Koller et al., 2012; Andersen et al., 2012). The CIF describes the cumulative probability of a particular event type occurring by a certain time in the presence of remaining event types, while the CSH function estimates the instantaneous occurrence rate of a specific event type in the presence of the others. It is important to note that the CIF explicitly quantifies clinical prognosis and is useful for prediction purposes (Koller et al., 2012; Bakoyannis

et al., 2017). In this dissertation, it is focused on making inferences about the CIF to build prognostic models.

Truncation and censoring are special features that make incomplete information of time in applications of survival analysis. An incident sampling generates right-censored data if a subset of individuals in the study do not experience the event of interest until the end of the study (Wolfson et al., 2019). On the other hand, a prevalent sampling produces left-truncated data because the recruitment of the study is screened by certain criteria such as a recruitment of individuals with a particular disease but not the event of interest (M.-C. Wang et al., 1993; Y.-J. Cheng et al., 2019). Interval censoring is a phenomenon, where the actual event time is not precisely observed but is only known to lie between two observation times such as clinic visits in periodic follow-up studies. Left truncation known as delayed entry is a peculiar phenomenon where certain participants enter the study after the onset of disease under study. These participants with an event prior to their potential study entry time are not included in the analysis and this can induce selection bias. In addition to interval censoring and left truncation, the event types for some individuals are missing due to the usual non-response or by the study design. This introduces an additional complexity in the analysis of the CIF. In order to address unmet needs in appropriate methods and software for competing risks data analysis, it is of necessity to develop application and methods.

In Chapter 2, the aim is to develop a convenient and flexible tool in the R environment. The R package intccr performs semiparametric regression analysis on the CIF for interval-censored competing risks data (Park, Bakoyannis, & Yiannoutsos, 2019). The B-spline-based semiparametric regression methodology proposed

by Bakoyannis et al. (2017), that provides semiparametrically efficient estimator of regression coefficients, is implemented. The package supports a large class of semiparametric odds rate transformation models, including the proportional odds and the Fine-Gray subdistribution hazards model as special cases. A comprehensive analysis is demonstrated by using data obtained from a human immunodeficiency virus (HIV) cohort study in sub-Saharan Africa.

In Chapter 3, an augmented inverse probability weighted sieve maximum likelihood estimator is proposed for the analysis of interval-censored competing risks data in the presence of missing event types. Weaker missing at random assumption is imposed to the estimator by incorporating auxiliary variables that are potentially associated with the probability of missingness. The proposed estimator offers double robustness that the estimator is consistent even if either the model for the probability of missingness or the model for the probability of the event type is misspecified. The proposed method is illustrated with data obtained from the HIV study in sub-Saharan Africa including HIV care and treatment clinics affiliated with the Academic Model Providing Access to Healthcare program in Kenya, Uganda, and Tanzania. The proposed methodology is subsumed under the existing R package intccr (Park et al., 2019).

In Chapter 4, the B-spline-based sieve maximum likelihood method proposed by Bakoyannis et al. (2017) is extended to left-truncated and interval-censored competing risks data. The proposed method is applied to the longitudinal data obtained from the Indianapolis-Ibadan Dementia Project to build a prognostic model for elderly African Americans in Indianapolis. The proposed methodology is included in

the existing R package intccr (Park et al., 2019). Also, a simple tutorial is available to provide the introduction how the proposed methodology is used in practice.

In Chapter 5, the implications of this dissertation are discussed followed by future research and limitations in line with the aims of the entire dissertation.

# Chapter 2

## Semiparametric competing risks regression under interval censoring using the R package intccr

Competing risks data are frequently interval-censored in real-world applications. This means that the exact event time is not precisely observed but is only known to lie between two time points such as clinic visits. This type of data requires special handling because the actual event times are unknown. An easy-to-use open-source statistical software is developed to deal with this problem. An approach to perform semiparametric regression analysis of the CIF with interval-censored competing risks data is the sieve maximum likelihood method based on B-splines. An important feature of this approach is that it does not impose restrictive parametric assumptions. Also, this methodology provides semiparametrically efficient estimates. Implementation of this methodology can be easily performed using new R package intccr. This R package performs semiparametric regression analysis of the CIF based on interval-censored competing risks data. It supports a large class of models including the proportional odds and the Fine–Gray proportional subdistribution hazards model as special cases. It also provides the estimated CIFs for a particular combination of covariate values. The package intccr also provides some data management functionality to handle data sets which are in a long format involving multiple lines of data per subject as well as the Wald test for overall model or the cause-specific model. The R package intccr comes up with a convenient and flexible software for the analysis of the CIF based on interval-censored competing risks data.

## 2.1 Introduction

Competing risk data are time-to-event data where multiple event types exist. The term "competing risks" also includes situations where the scientific interest is focused on the first occurring event (Putter et al., 2007; Bakoyannis & Touloumi, 2012). In a motivating example, taken from a Human Immunodeficiency Virus (HIV) care and treatment program in sub-Saharan Africa, patients were at risk of death while receiving antiretroviral treatment (ART) and while in care or of becoming lost to care. This latter situation is important because patients who are not retained in care are less likely to receive ART, can infect others in the community and have worse prognosis themselves. In such studies, the interest typically lies on the first event that patients experience, whether this is death or loss to HIV care. The main estimands from such competing risks data are the cause-specific hazard function and the CIF. The cause-specific hazard function represents the instantaneous failure rate from a specific event in the presence of the other events, while the CIF represents the cumulative probability of an event in the presence of the others. The analysis of the CIF is of interest because it is the key quantity for studying the risk of occurrence of various events. The CIF is used for studying disease prognosis, for evaluating interventions in populations and for prediction and implementation science purposes (Koller et al., 2012; Bakoyannis et al., 2017). In the case of right-censored competing risk data, the packages cmprsk and prodlim can be used to estimate the CIF by using nonparametric method, based on the Aalen–Johansen estimator (Gray, 2017; Gerds, 2017; Aalen & Johansen, 1978). The function cif in the package compeir estimates the CIF by using parametric model for each competing risk (Grambauer & Neudecker,

2011). For regression analysis of the CIF, the packages cmprsk, kmi, survival (with the function survfit), and riskRegression can be used to fit the Fine–Gray proportional subdistribution hazards model (Gray, 2017; Allignol, 2017; Therneau & Lumley, 2016; Gerds et al., 2017; Fine & Gray, 1999). The package timereg provides semiparametric estimators for a whole class of models that includes the Fine–Gray model as a special case (Scheike et al., 2017, 2008; Scheike & Zhang, 2011). Additionally, the package cmprskQR performs quantile regression analysis of subdistribution functions (Dlugosz et al., 2016; Peng & Fine, 2009).

A frequent problem in many clinical studies is that the event time is not precisely observed but is only known to lie between two examination times, such as clinic visits (Sun, 2006; Chen et al., 2012; Y. Zhang et al., 2010; Bakoyannis et al., 2017). This phenomenon is known as interval censoring in survival and competing risks analysis. In the motivating example, the working definition of loss to care was three months without a clinic visit. This cutoff was chosen by the clinical investigators because, typically, HIV patients receive ART supplies for up to three months at each clinic visit. The analytical problem is that the exact time of disengagement from HIV care, among patients who have not returned for their next visit, is only known to lie within the three-month interval following the last clinic visit. Similarly, the exact time to death is not known as the data set contains only the death reporting date which is usually after the actual death date. Therefore, the actual death date lies between the last clinic visit of the patient and the death reporting date. Although interval-censored competing risk data arise frequently in a variety of clinical and medical research settings, only two R packages exist for the analysis of such data. The first is the package MLEcens that applies the height mapping algorithm and the

support reduction algorithm by Maathuis (2005) and Groeneboom et al. (2008) to compute the nonparametric maximum likelihood estimate of the CIF with bivariate interval-censored data (Maathuis, 2013). The second is the package MIICD that includes the function MIICD.crreg (Delord, 2017). This package implements the multiple imputation approach proposed by Pan (2000) to estimate the regression coefficients and the baseline CIF based on the Fine–Gray proportional subdistribution hazards model (Fine & Gray, 1999). However, the package MLEcens does not involve covariates, and the package MIICD (Delord, 2017) uses Rubin's variance estimator, which is well known to be biased when the imputation model and the analysis models are uncongenial (Maathuis, 2013; Delord, 2017; Meng, 1994). Moreover, the latter package only fits the Fine–Gray proportional subdistribution hazards model, and the corresponding regression coefficient estimators do not attain the semiparametric efficiency bound (L. Mao & Lin, 2017).

The package intccr attempts to deal with the aforementioned issues by implementing the semiparametric regression methodology proposed by Bakoyannis, Yu, & Yiannoutsos (2017) for the analysis of interval-censored competing risk data (Park et al., 2019). It is important to note that the latter methodology provides semiparametric efficient regression coefficient estimates (Bakoyannis et al., 2017). The function ciregic contained in the intccr package fits semiparametric regression models for the CIF that belong to the large class of generalized odds rate transformation models with interval-censored competing risk data (Jeong & Fine, 2006; Dabrowska & Doksum, 1988; Fine, 2001; Scharfstein et al., 1998). This class includes the Fine–Gray proportional subdistribution hazards model and the proportional odds model as special cases. The function ciregic produces a simple and familiar table of the sum-

marized results. Also, the package intccr provides an option for parallel computing that can provide a faster bootstrap estimation of the variance-covariance matrix for the estimated regression coefficients.

## 2.2 Methodology

### 2.2.1 Notation

Let $T$ be the actual unobserved event time and let $C = j \in \{1, 2, \ldots, k\}$ be the observed event or event type. Currently, the package intccr allows for two competing risks, i.e. $C \in \{1, 2\}$. Let $[a, b]$ denote the examination time interval with $0 < a < b < \infty$. For $i = 1, \ldots, n$, the $l_i$ distinct examination times of the $i$th study participant are denoted by $a \leq E_{i,1} < E_{i,2} < \cdots < E_{i,l_i} \leq b$. Also, the last examination time prior to the event is denoted as $V_i$ and the first examination time after the event as $U_i$. Based on this notation, the event time of the $i$th study participant is contained in $(V_i, U_i]$. If the $i$th study participant's event time is left-censored then $(V_i, U_i] = (0, E_{i,1}]$, if it is right-censored then $(V_i, U_i] = (E_{i,l_i}, \infty]$, and if it is interval-censored between the examination times $E_{i,l_i-1}$ and $E_{i,l_i}$, then $(V_i, U_i] = (E_{i,l_i-1}, E_{i,l_i}]$. Now, let $\Delta_i^{(1)} = I(V_i < T_i \leq U_i)$ be the indicator function that the $i$th study participant is interval-censored, and let $\Delta_{ij}^{(1)} = \Delta_i^{(1)} I(C_i = j)$ be the interval censoring and the $j$th event indicator function. Similarly, let $\Delta_i^{(2)} = I(0 < T_i \leq U_i)$ denote that the indicator function that $i$th study participant is left-censored, and let $\Delta_{ij}^{(2)} = \Delta_i^{(2)} I(C_i = j)$ be the left censoring and the $j$th event indicator function.

The event indicator from any event types is defined as $\Delta_i = \sum_{j=1}^{2} \left(\Delta_{ij}^{(1)} + \Delta_{ij}^{(2)}\right)$. Obviously, $\Delta_i = 0$ indicates that the $i$th study participant is right-censored. Finally, let $Z \in \mathbb{R}^d$ be a vector of covariates of interest. The observed data for the $i$th study participant are thus $D_i = \left(V_i, U_i, C_i, \Delta_{ij}^{(1)}, \Delta_{ij}^{(2)}, Z_i\right)$. The cause-specific CIF for the $j$th event is expressed by

$$F_j(t; z) = P(T \leq t, C = j | Z = z)$$

for $j = 1, 2$.

### 2.2.2 Estimation methodology

With the assumptions that $(E_{i,1}, E_{i,2}, \ldots, E_{i,l_i}) \perp (T_i, C_i)$ conditional on $Z_i$ and that the observation time distribution does not contain the parameters of interest (non-informative interval censoring), the likelihood function is

$$L(\theta; D) \propto \prod_{i=1}^{n} \left[ \left[ \prod_{j=1}^{2} \left\{ F_j(U_i; Z_i, \theta_j) - F_j(V_i; Z_i, \theta_j) \right\}^{\Delta_{ij}^{(1)}} \right] \left[ \prod_{j=1}^{2} \left\{ F_j(U_i; Z_i, \theta_j) \right\}^{\Delta_{ij}^{(2)}} \right] \times \left\{ 1 - \sum_{j=1}^{2} F_j(V_i; Z_i, \theta_j) \right\}^{1-\Delta_i} \right] \tag{2.1}$$

where $\theta = \left(\theta_1^T, \theta_2^T\right)^T$ are the unknown parameters to be estimated. The CIFs can be modeled by using a member of the class of semiparametric transformation models. (Fine & Gray, 1999; Jeong & Fine, 2006; L. Mao et al., 2017) The general form is

given as

$$g_j\big\{F_j(t;z)\big\} = \phi_j(t) + \beta^T z$$

for $j = 1, 2$, where $g_j(\cdot)$ is a known increasing link function and $\phi_j(\cdot)$ is an unspecified increasing and invertible smooth function (infinite-dimensional parameter) which is related to the $j$th baseline CIF. In this case, $\theta_j = \left(\phi_j, \beta_j^T\right)^T$. A special subset of the class of semiparametric transformation models is the class of generalized odds transformation models which is defined as

$$g_j(F_j;\alpha_j) = \begin{cases} \log\left\{-\log\left(1-F_j\right)\right\} & \text{if } \alpha_j = 0 \\ \\ \log\left\{\dfrac{(1-F_j)^{-\alpha_j}-1}{\alpha_j}\right\} & \text{if } \alpha_j \in (0,\infty) \end{cases}$$

The Fine–Gray proportional subdistribution hazards model is a special case of this class of models with $\alpha_j = 0$, and so is the proportional odds model with $\alpha_j = 1$ (Fine & Gray, 1999; Eriksson et al., 2015). An effective approach to deal with maximum likelihood estimation problems that involve infinite-dimensional parameters is the sieve maximum likelihood approach (X. Shen & Wong, 1994). This approach avoids some theoretical problems related to likelihood maximization over infinite-dimensional parameter spaces and, also, provides computational efficiency gains (Y. Zhang et al., 2010; X. Shen & Wong, 1994). Bakoyannis et al. (2017) used a sieve maximum likelihood estimation approach based on B-splines. The corresponding sieve parameter

space is given by

$$\mathcal{M}_n(\gamma_j, N_j, m_j) = \left\{ \phi : \phi(t;\gamma_j) = \sum_{s=1}^{N_j+m_j} \gamma_{j,s} B_{s,m_j}(t), \right. \\ \left. \gamma \in \mathbb{R}^{N_j+m_j}, \gamma_{j,1} < \cdots < \gamma_{j,N_j+m_j} \right\} \tag{2.2}$$

where $N_j$ and $m_j$ are the number of internal knots and the order of the B-spline for the $j$th event type, and $\left\{\gamma_{j,1}, \ldots, \gamma_{j,N_j+m_j}\right\}$ is the set of B-spline coefficients. For more details about the optimal choice of the number of knots see Chapter 2.5 and Section 2.1 in Bakoyannis et al. (2017). Maximizing the likelihood function in Equation (2.1) with respect to the regression coefficients over a regular Euclidean space and the unspecified functions $\phi_1$ and $\phi_2$ over the B-spline sieve space provides the sieve maximum likelihood estimates $\left(\hat{\phi}_1, \hat{\beta}_1^T, \hat{\phi}_2, \hat{\beta}_2^T\right)^T$. The consistency for $\left(\hat{\phi}_1, \hat{\beta}_1^T, \hat{\phi}_2, \hat{\beta}_2^T\right)^T$, and the asymptotic normality and semiparametric efficiency of $\left(\hat{\beta}_1^T, \hat{\beta}_2^T\right)^T$, have been established by Bakoyannis et al. (2017).

The function ciregic in the package intccr performs the proposed method with nonlinear inequality constraints, using the package alabama, to impose the monotonicity constraints involved in Equation (2.2), which follow from the natural monotonicity of the CIF. In addition, the function ciregic utilizes the package alabama to impose the non-linear inequality constraint

$$\max_z \left\{ \sum_{j=1}^{2} F_j(b; z, \theta_j) \right\} < 1,$$

since the sum of the two CIFs is a probability and, as such, it is naturally bounded by 1.

## 2.3 Basic use of the package and simulation study

The version information of R and the platform of operating system(OS) used in this Chapter are as follows:

```
R> c(R.version$platform, R.version$version.string)
[1] "x86_64-w64-mingw32"    "R version 3.5.2 (2018-12-20)"
```

Under 64-bit version of Windows 10 OS, Monte Carlo simulation and data analysis were performed. With the assumption that the user has the most recent version of R installed, the most recent version of the package intccr has to be installed on the user's OS and loaded as follows:

```
R> install.package("intccr")
R> library(intccr)
R> packageVersion("intccr")
[1] '1.1.1'
```

The package intccr provides two simulated data sets. The first data set is longdata which is a long data format, and the second data set is simdata which is a ready-to-use data format. The data set longdata consists of 200 individuals with 5 variables, where id represents individuals' identification number, t represents the clinic visit or event evaluation times, c represents the event or censoring indicator, and z1 and z2 are binary and continuous covariates respectively. Note that c has to be 0, 1, or 2, with 0 indicating that the event was not observed throughout the total follow-up period (right censoring). The first 10 observations of longdata are listed below.

```
R> head(longdata, n = 10)
   id          t  c z1          z2
1   1 0.86224187 0  0 -2.29032656
2   1 1.20644148 0  0 -2.29032656
3   1 1.73209303 0  0 -2.29032656
4   1 1.73539999 0  0 -2.29032656
5   1 1.96647129 0  0 -2.29032656
6   1 2.12675792 0  0 -2.29032656
7   1 2.46613799 2  0 -2.29032656
8   2 0.05551998 0  1  0.00261902
9   2 0.17492399 0  1  0.00261902
10  2 0.18091429 0  1  0.00261902
```

To analyze the data set longdata in the function ciregic, the data must be reshaped to a suitable format. The package intccr provides the function dataprep to reshape data from a long format to a suitable format that is required by the function ciregic.

```
R> newdata <- dataprep(data = longdata, ID = id, time = t, event = c,
                       Z = c(z1, z2))
```

The first 10 observations of newdata are given by

```
R> head(newdata, n = 10)
   id         v         u c z1          z2
1   1 2.1267579 2.4661380 2  0 -2.29032656
2   2 0.1809143 0.3769367 1  1  0.00261902
```

```
3    3 2.9436552         Inf  0   1 −1.68379376
4    4 2.4305333         Inf  0   1 −0.90535264
5    5 0.5731781 1.2847889  2   0   0.22854677
6    6 0.0000000 0.3777047  1   0 −0.51449544
7    7 0.0000000 1.4617243  1   1 −1.42043786
8    8 0.0000000 0.4781881  2   1 −0.47006673
9    9 0.1068374 0.9656031  2   0 −0.19349437
10 10 0.3917861 1.0805153  1   0 −0.81510083

R> table(newdata$c)
 0  1  2
29 76 95
```

There are two competing events: the first (c = 1) and the second (c = 2) event type. Right-censored observations are indicated by c = 0. There are 76 observations with the first event type, 95 observations with the second, and 29 observations are right-censored. To elucidate the underlying mechanisms of the function dataprep, Figure 2.1 shows how longdata is reshaped into newdata via the use of the function dataprep. In longdata, three individuals with id = 1, id = 2, and id = 5 had 7, 4, and 3 time records respectively. These individuals experienced one of the event types between their last two time records.

This infers that the event times of those individuals were interval-censored. The function dataprep detected a type of events that an individual experienced and the corresponding time interval. In addition, the function dataprep returned v as the last observation prior to the event and u as the first observation after the event in

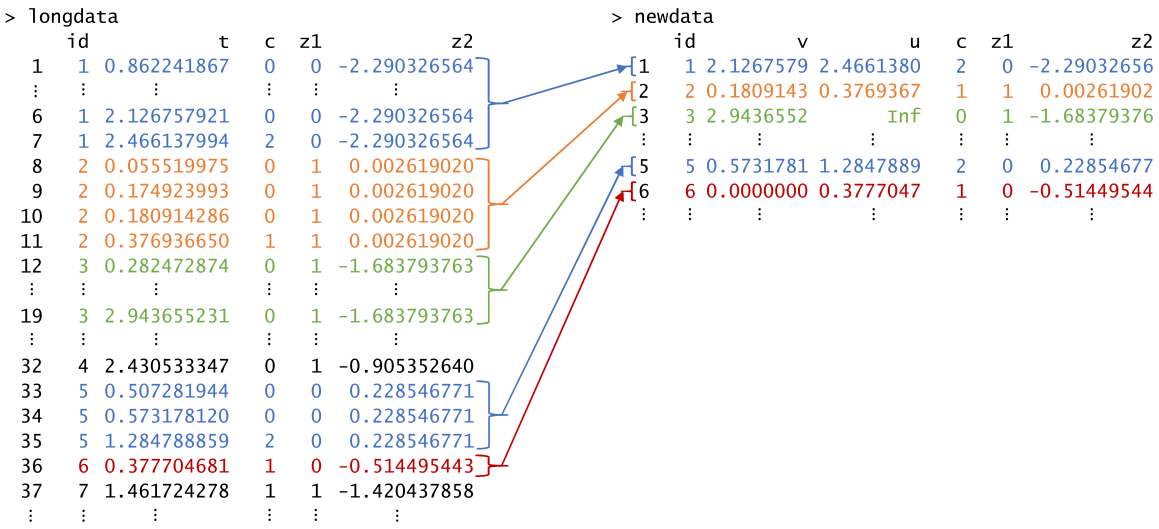



Figure 2.1: Data resphaping scheme of the function dataprep

newdata. The individual with id = 3 who have 8 time records in longdata did not experience any events. The function dataprep returned v = 2.9426552, which is the last time record of the individual with id = 3, as the last observation prior to the event and u = Inf as the first observation after the event in newdata because the individual with id = 3 was right-censored. For the individual with id = 6, the only one time record was observed with the first event type. Therefore, the last observation prior to the event was v = 0 and the first observation after the event was u = 0.3777047 in the newdata because the individual with id = 6 was left-censored. The summaries of the covariates z1 and z2 in newdata are listed below.

```
R> table(newdata$z1)

  0   1

122  78

R> summary(newdata$z2)

     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.

−2.64245 −0.61216 0.02428  0.02383   0.70391   2.86069
```

Table 2.1: Arguments of the function ciregic

| Arguments | Description |
|---|---|
| formula | a formula object relating survival object Surv2(v, u, event) to a set of covariates |
| data | an input data frame |
| alpha | parameters that define the link functions from class of generalized odds-rate transformation models |
| k | a tuning parameter that controls the number of knots in B-spline |
| nboot | a number of bootstrap samples for estimating variances and covariances of an estimated regression coefficients |
| do.par | a logical constant for using parallel computing for bootstrap calculation |

The arguments of the core function ciregic are described in Table 2.1. The data must contain the last observation time prior to the event, the first observation time after the event, and the event indicator. The function ciregic computes the class of generalized transformation models of semiparametric regression on interval-censored competing risk data with B-spline sieve maximum likelihood estimation. The value of $\alpha = (1, 1)^T$ for the link functions of the two competing risks is used in this simulation, which corresponds to the proportional odds model (Dabrowska & Doksum, 1988; Eriksson et al., 2015) for both event types as described in Chapter 2.2. This is because the data were simulated from proportional odds models for both

event types. Sample R code and the corresponding output of the function ciregic are listed below:

```
R> set.seed(12345)
R> fit.newdata <- ciregic(formula = Surv2(v = v, u = u, event = c) ~ z1 + z2,
                          data = newdata, alpha = c(1, 1),
                          nboot = 0, do.par = FALSE)
R> fit.newdata
Call:
ciregic.default(formula = Surv2(v = v, u = u, event = c) ~ z1 + z2,
                data = newdata, alpha = c(1, 1),
                do.par = FALSE, nboot = 0)
Event type 1
Coefficients :
        z1         z2
 0.5230574 -0.2426299
Event type 2
Coefficients :
        z1         z2
-0.3963446  0.3442936
```

There are 6 arguments in the function ciregic (see Table 2.1). The argument formula has the form of response $\sim$ predictor. The response part of the formula must be a Surv2 object in the function ciregic, and the predictor is a vector of covariates. The first argument in Surv2 is the last examination time before the event, the second

is the first examination time after the event, and the last the event type or censoring status (c $\in \{0, 1, 2\}$), with 0 indicating right censoring. The argument alpha is a vector of two parameters that represent the link functions of generalized odds rate transformation models for competing events. The support of $\alpha$ is $[0, \infty) \times [0, \infty)$. For example, $\alpha_1 = 0$ fits the Fine–Gray proportional subdistribution hazards model for event type 1 and $\alpha_2 = 1$ fits the proportional odds model for event type 2 (Fine & Gray, 1999; Eriksson et al., 2015). The argument k is a tuning parameter that controls the number of knots defining a B-spline. k = 1 is the default, but the user can choose any values satisfying $0.5 \leq$ k $\leq 1$. Using the half number of internal knots compared to the default can be achieved by choosing k = 0.5 than as the default. This choice can have a substantial effect on computation time with larger data sets. The function ciregic uses cubic B-splines. The argument nboot = 0 forces the function ciregic to returning only the estimated regression coefficients without calculating the bootstrap variance-covariance matrix for the estimated regression coefficients. The function ciregic provides the bootstrap variance-covariance matrix for the estimated regression coefficients when a value of the argument nboot is greater than or equal to 2. By setting nboot = 0 and do.par = FALSE, the function ciregic returns only the estimated regression coefficients. This is useful when it is desirable to fit the model and just get point estimates. Below is the sample R code to obtain a bootstrap variance-covariance matrix in parallel computing:

```
R> set.seed(12345)
R> fit.newdata.b <− ciregic(formula = Surv2(v = v, u = u, event = c) ∼ z1 + z2,
                            data = newdata, alpha = c(1, 1),
```

```
                    nboot = 50, do.par = TRUE)
R> summary(fit.newdata.b)
Call:
ciregic.default(formula = Surv2(v = v, u = u, event = c) ~ z1 + z2,
                data = newdata, alpha = c(1, 1),
                nboot = 50, do.par = TRUE)
Event type 1
     Estimate Std. Error z value Pr(>|z|)
z1     0.5231     0.2708   1.931   0.0534 .
z2    -0.2426     0.1067  -2.274   0.0230 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Event type 2
     Estimate Std. Error z value Pr(>|z|)
z1    -0.3963     0.2509  -1.580  0.11419
z2     0.3443     0.1296   2.657  0.00788 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The argument nboot requires a non-negative integer and denotes the number of bootstrap samples used to estimate a variance-covariance matrix of the estimated regression coefficients . In the above application, the following setting was used: nboot = 50 and do.par = TRUE. This means that 50 bootstrap samples were used to compute the variance-covariance matrix in parallel computing. The packages doPar-

Table 2.2: Monte Carlo simulation results based on 1,000 replications

| n | Event type | Parameters | %bias | MCSD | ASE | ECP |
|---|---|---|---|---|---|---|
| 100 | 1 | $\beta_{11}$ | -3.055 | 0.403 | 0.425 | 0.957 |
| | | $\beta_{12}$ | 4.645 | 0.189 | 0.207 | 0.967 |
| | 2 | $\beta_{21}$ | -1.033 | 0.394 | 0.417 | 0.960 |
| | | $\beta_{22}$ | 4.867 | 0.191 | 0.202 | 0.961 |
| 200 | 1 | $\beta_{11}$ | -0.578 | 0.282 | 0.285 | 0.954 |
| | | $\beta_{12}$ | 2.983 | 0.144 | 0.140 | 0.939 |
| | 2 | $\beta_{21}$ | 1.418 | 0.273 | 0.282 | 0.948 |
| | | $\beta_{22}$ | 2.737 | 0.139 | 0.136 | 0.936 |
| 400 | 1 | $\beta_{11}$ | 0.683 | 0.198 | 0.197 | 0.947 |
| | | $\beta_{12}$ | -0.851 | 0.097 | 0.097 | 0.952 |
| | 2 | $\beta_{21}$ | 1.812 | 0.196 | 0.195 | 0.953 |
| | | $\beta_{22}$ | -0.127 | 0.095 | 0.095 | 0.946 |
| 800 | 1 | $\beta_{11}$ | 2.160 | 0.138 | 0.138 | 0.951 |
| | | $\beta_{12}$ | 0.261 | 0.070 | 0.069 | 0.941 |
| | 2 | $\beta_{21}$ | 1.884 | 0.136 | 0.136 | 0.946 |
| | | $\beta_{22}$ | -0.011 | 0.066 | 0.067 | 0.944 |

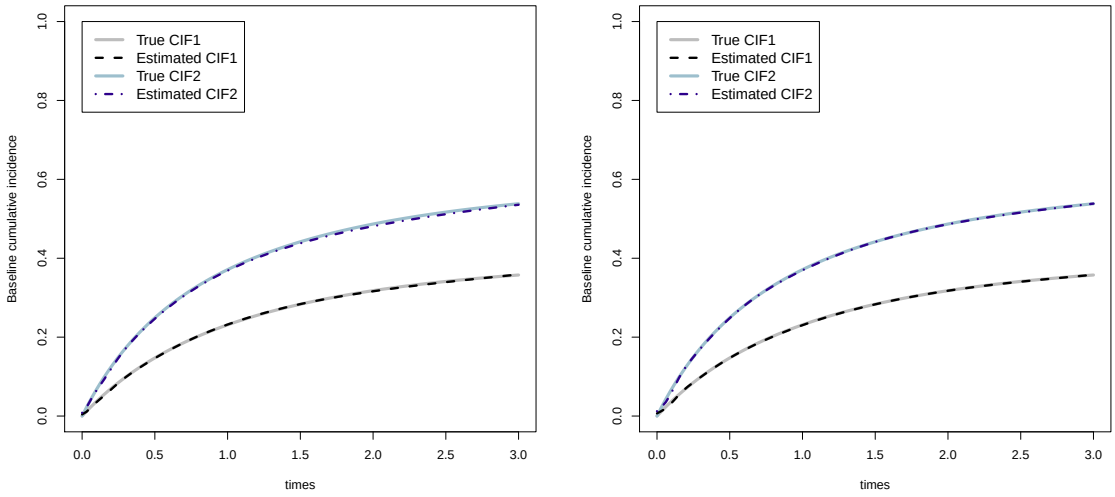
The standard error is estimated by bootstrap sampling. Monte Carlo standard deviation (MCSD), average standard error (ASE), empirical coverage probability (ECP).

Table 2.3: Computation time (seconds) based on Monte Carlo simulation using 1,000 replications

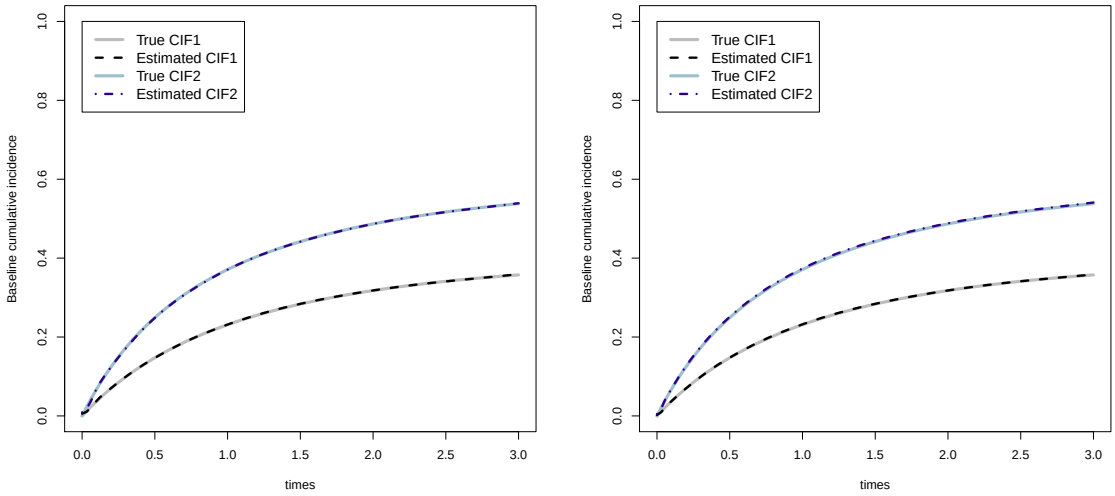
| Parallel | $n$ | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|---|
| Yes | 100 | 29.78 | 37.45 | 39.96 | 42.80 | 57.70 |
| | 200 | 29.85 | 38.59 | 41.23 | 43.96 | 65.06 |
| | 400 | 31.91 | 42.59 | 45.95 | 49.80 | 70.07 |
| | 800 | 39.33 | 49.53 | 52.95 | 56.83 | 75.22 |
| No | 100 | 81.92 | 102.22 | 108.74 | 116.00 | 145.39 |
| | 200 | 85.59 | 105.00 | 112.94 | 120.80 | 166.98 |
| | 400 | 88.70 | 116.04 | 126.58 | 136.70 | 194.84 |
| | 800 | 105.36 | 135.38 | 145.95 | 156.90 | 209.63 |

allel and parallel are implemented to set the environment for parallel computing, and the package foreach is used to perform bootstrap calculations simultaneously. The argument do.par = TRUE detects the number of cores automatically and assigns jobs to the maximum number of available cores. The total number of assigned cores is usually the same as the total number of detected cores minus one.

Extensive Monte Carlo simulations based on 1,000 replications were performed with sample sizes 100, 200, 400, and 800. The results of the simulations are shown in Table 2.2. The vector of the estimated regression coefficients is $\hat{\beta} = (\hat{\beta}_{11}, \hat{\beta}_{12}, \hat{\beta}_{21}, \hat{\beta}_{22})^T$ which are associated with the estimated regression coefficients of z1 and z2 for the two event types, respectively. Among 1,000 replications for the Monte Carlo simulations, one data set with 100 observations did not converge in at least one bootstrap sample generated in order to calculate the bootstrap standard error. Similarly, two data sets




(a) $n = 100$ (b) $n = 200$




(c) $n = 400$ (d) $n = 800$

Figure 2.2: The predicted baseline cumulative incidence functions resulted from a simulation study with sample sizes of 200, 400, and 800
Solid gray and light blue lines indicate true baseline cumulative incidence functions of the first event type and the second event type respectively. Dotted black and blue lines indicate the estimated baseline cumulative incidence functions of the first event type and the second event type respectively.

with 200 observations did not converge. Despite these very rare non-convergence issues, the simulation results show small percent biases, similar values of Monte Carlo standard deviation (MCSD) and average standard error (ASE), and values of empirical coverage probability (ECP) close to the nominal level of 0.95. Moreover, the MCSD for the different sample sizes shows a $\sqrt{n}$ convergence rate of the estimator. Figure 2.2 depicts the true baseline CIFs along with the estimated baseline CIF for both event types. This Figure illustrates that the function ciregic provides virtually unbiased estimates even with small sample sizes. Table 2.3 shows summaries of computation time for Monte Carlo simulation. In each scenario, the median computation time using parallel computing option (do.par = TRUE) to calculate bootstrap variance-covariance matrix is roughly three times more timely efficient than those without parallel computing option (do.par = FALSE).

## 2.4 Example: Analysis of HIV data using the intccr package

A data analysis from the HIV study on death and loss to HIV care in sub-Saharan Africa is presented in this chapter. The data were collected by the IeDEA-EA (East African International epidemiology Databases to Evaluate AIDS) Cohort Consortium from retrospective cohort study including HIV care and treatment programs in Kenya, Uganda, and Tanzania. The data used here include 3,053 patients who initiated ART with a cluster of differentiation 4 (CD4) cell count of at least 100 cells/$\mu l$. The data consist of 6 variables, with v being the last clinical examination time prior to the event since ART initiation, u the first clinical examination time after the event,

c the event or right-censoring indicator, and age, male and cd4 being the age at ART initiation, male gender indicator, and CD4 cell count at ART initiation, respectively.

```
R> library(intccr)
R> head(iedea, n = 5)
          v           u c       age male cd4
1 0.27104723         Inf 0 35.67146    0 192
2 0.31759068         Inf 0 45.65366    1 191
3 0.14784394 0.1724846 2 62.52977    1 102
4 0.05475701 0.3011636 1 30.77892    0 144
5 2.44490080         Inf 0 43.16496    0 664
```

In total, there were 2,232 patients in HIV care who did not experience any of the events throughout the follow-up period (c = 0, right-censored observations). Moreover, 690 patients were lost to care (c = 1), and 131 patients died while in HIV care (c = 2).

```
R> table(iedea$c)
   0    1    2
2232  690  131
```

Summary statistics regarding age by event type or censoring c are given below:

```
R> tbl.age <- rbind(summary(iedea[iedea$c == 1,]$age, digits = 4),
                    summary(iedea[iedea$c == 2,]$age, digits = 4),
                    summary(iedea[iedea$c == 0,]$age, digits = 4))
R> rownames(tbl.age) <- c("Loss to care", "Death", "In HIV care")
R> tbl.age
```

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| Loss to care | 18.45 | 28.63 | 35.21 | 36.32 | 41.65 | 78.65 |
| Death | 20.51 | 35.16 | 40.86 | 42.43 | 50.75 | 76.96 |
| In HIV care | 18.18 | 30.42 | 37.14 | 38.33 | 44.87 | 84.22 |

The median age was 35.2 years, 40.9 years, and 37.1 years for those lost to care, deceased, and still alive and in HIV care at the end of the follow-up period, respectively. Similarly, summary statistics for cd4 by event type are given below:

```
R> tbl.cd4 <- rbind(summary(iedea[iedea$c == 1,]$cd4),
                    summary(iedea[iedea$c == 2,]$cd4),
                    summary(iedea[iedea$c == 0,]$cd4))
R> rownames(tbl.cd4) <- c("Loss to care", "Death", "In HIV care")
R> tbl.cd4
```

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| Loss to care | 101 | 140.25 | 188 | 231.8101 | 262.0 | 1576 |
| Death | 102 | 131.00 | 163 | 187.0687 | 212.5 | 1135 |
| In HIV care | 101 | 152.00 | 199 | 234.9453 | 276.0 | 1332 |

The median CD4 cell count at ART initiation was 188 cells/$\mu l$, 163 cells/$\mu l$, and 199 cells/$\mu l$ for those lost to care, deceased, and still alive and in HIV care at the end of the follow-up period, respectively. For the data, $\alpha = (1, 1)^T$ was used. This means that the proportional odds model was assumed for both event types (i.e. loss to care and death). This choice was made due to the straightforward interpretation of the regression coefficient estimates under the model. For reproducibility purposes

regarding the bootstrap variance-covariance matrix of the estimated regression coefficients, the seed number 12345 was set.

```
R> set.seed(12345)
R> fit <- ciregic(formula = Surv2(v = v, u = u, event = c) ~ male + age + cd4,
                  data = iedea, alpha = c(1, 1), k = 1, nboot = 50,
                  do.par = TRUE)
```

The function ciregic is an S3 class function, and therefore the function can be used in conjunction with the generic accessor functions coef, vcov, and summary, as it is illustrated below.

```
R> summary(fit)
Call:
ciregic.default(formula = Surv2(v = v, u = u, event = c) ~ male + age + cd4,
                data = iedea, alpha = c(1, 1), k = 1,
                do.par = TRUE, nboot = 50)
Event type 1
     Estimate Std. Error z value Pr(>|z|)
male   0.2128     0.1055   2.017   0.0437 *
age   -0.0295     0.0058  -5.087   <2e-16 ***
cd4    0.0000     0.0003   0.025   0.9797
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Event type 2
     Estimate Std. Error z value Pr(>|z|)
```

male 0.5668 0.1952 2.904 0.0037 **

age 0.0314 0.0084 3.765 0.0002 ***

cd4 −0.0035 0.0018 −1.989 0.0467 *

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The function summary returns summaries of the fitted model results with asterisks indicating the corresponding level of statistical significance. The code below extracts the vector of the estimated regression coefficients and the bootstrap variance-covariance matrix, respectively.

R> coef(fit)

R> vcov(fit)

The results from the analysis presented above indicate that the odds of loss to care for males is about 24% higher compared to the corresponding odds for females (odds ratio = exp(0.2128) = 1.24). Also, older age at ART initiation by 10 years is associated with a 26% lower odds of loss to care (exp(10 × −0.0295) = 0.74). Additionally, an increased CD4 cell count at ART initiation is not associated with a higher odds of loss to care because exp(100×0) = 1. Moreover, older age by 10 years is associated with 9% higher odds of death (odds ratio = exp(10 × 0.0084) = 1.09), and, an increased CD4 cell count by 100 cells/$\mu l$ is associated with 30% lower odds of death (odds ratio = exp(100 × −0.0035) = 0.70). The predicted CIFs of loss to care and death for females with a CD4 count of 120 cells/$\mu l$ at ART initiation, according to age at ART initiation, are depicted in Figure 2.3. Fitting the proportional subdistribution

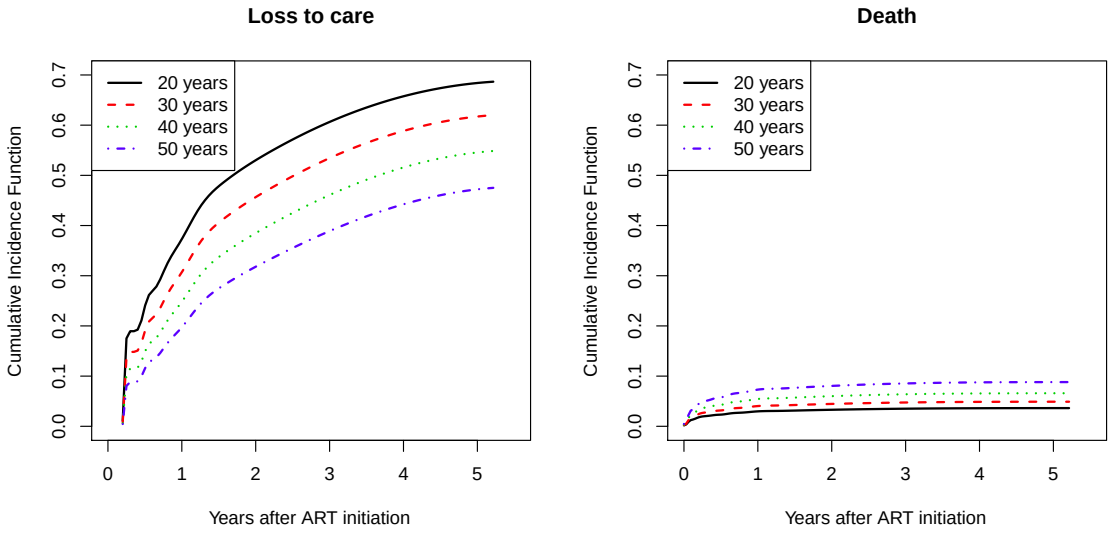



Figure 2.3: The predicted cumulative incidence functions for females aged 20 to 50 years, with CD4 count 120 cells/$\mu l$ at ART initiation

hazards model (i.e. Fine–Gray model) for loss to care and the proportional odds model for death can be performed by setting $\alpha_1 = 0$ and $\alpha_2 = 1$, as follows:

```
R> set.seed(12345)
R> fit <- ciregic(formula = Surv2(v = v, u = u, event = c) ~ male + age + cd4,
                  data = iedea, alpha = c(0, 1), nboot = 50,
                  do.par = TRUE)
```

The generic accessor function predict can be directly used with an object of class ciregic. Table 2.4 describes the arguments of the function predict. In this example, the argument object is the previously fitted model fit. In the argument covp, the user defines the desired covariate pattern for (male, age, cd4), to predicting the corresponding covariate-specific CIFs of loss to HIV care and death. There are 4 lines of output representing 4 different combinations of age by the two event types, "loss to care" (c = 1) and "death" (c = 2) respectively. The argument times produces 100

equally distributed time points between the minimum and the maximum observation time point in the data, for each event type.

```
R> tms <- fit$tms
R> par(mfrow = c(1, 2))
R> t <- seq(from = tms[1], to = tms[2], by = diff(tms) / 99)
R> pred <- lapply(c(20, 30, 40, 50),
                  function(x) {predict(object = fit,
                                       covp = c(0, x, 120),
                                       times = t)})
R> plot(pred[[1]]$t, pred[[1]]$cif1, type = "l",
     ylim = c(0, 0.4), xlim = c(0, 3),
     xlab = "Years after ART initiation",
     ylab = "Cumulative Incidence Function",
     main = "Loss to care", lwd = 2)
R> for(i in 2:4) {
    lines(pred[[i]]$t, pred[[i]]$cif1, lty = i, col = i)
  }
R> legend("topleft",
       legend = c("20 years", "30 years",
                  "40 years", "50 years"),
       lty = c(1, 2, 3, 4), col = c(1, 2, 3, 4),
       lwd = c(2, 2, 2, 2), cex = 0.9)
R> t <- seq(from = tms[1], to = tms[2], by = 3 / 99)
```

```
R> pred <- lapply(c(20, 30, 40, 50),
                  function(x) {predict(object = fit,
                                       covp = c(0, x, 120),
                                       times = t)})
R> plot(pred[[1]]$t, pred[[1]]$cif2, type = "l",
     ylim = c(0, 0.1), xlim = c(0, 3),
     xlab = "Years after ART initiation",
     ylab = "Cumulative Incidence Function",
     main = "Death", lwd = 2)
R> for(i in 2:4) {
    lines(pred[[i]]$t, pred[[i]]$cif2, lty = i, col = i)
  }
R> legend("topleft",
       legend = c("20 years", "30 years",
                  "40 years", "50 years"),
       lty = c(1, 2, 3, 4), col = c(1, 2, 3, 4),
       lwd = c(2, 2, 2, 2), cex = 0.9)
R> par(mfrow = c(1, 1))
```

Moreover, the Wald test can be performed with the object in the function ciregic in the package intccr. Below are two examples to perform the Wald test. The first example is to compare the saturated model (male, age, cd4) and the null model (model without covariate).

```
R> set.seed(12345)
```

Table 2.4: The arguments of the function predict

| Arguments | Description |
|---|---|
| object | An object of class ciregic, generated from the fitted model |
| covp | The vector of covariates |
| times | User-defined time points used to predict the cumulative incidence functions |

```
R> fit.f <- ciregic(formula = Surv2(v = v, u = u, event = c) ~
                                male + age + cd4,
                    alpha = c(1, 1), nboot = 50, do.par = TRUE,
                    data = iedea)
R> waldtest(full = fit.f)
Full model: male age cd4
Nested model:
Wald test
    Chisq df P(> Chisq)
 75.5101  6       3e-14
Wald test (cause-specific)
Event type 1
    Chisq df P(> Chisq)
 26.1007  3       9e-06
Event type 2
    Chisq df P(> Chisq)
```

```
39.5401  3        1e−08
```

The function waldtest returns output for two parts: one is the test for all covariates and another is the cause-specific test. In the above example, the $\chi^2$ statistic of overall test is 75.5 and its $p$-value is close to 0. Also, the $\chi^2$ statistic of each test for event type 1 and 2 is 26.1 and 39.5 respectively and those p-values are close to 0. These results mean that the variables male, age, and cd4 should be in the model because parameters associated with those variables are not zero. The next example is the Wald test comparing the saturated model (male, age, cd4) and nested model (male, age).

```
R> set.seed(12345)
R> fit.n <− ciregic(formula = Surv2(v = v, u = u, event = c) ∼ male + age,
                    alpha = c(1, 1), nboot = 50, do.par = TRUE,
                    data = iedea)
R> waldtest(full = fit.s, nested = fit.n)
Saturated model: male age cd4
Nested model: male age
Wald test
  Chisq df P(> Chisq)
 5.3531  2      0.0688
Wald test (cause−specific)
Event type 1
 Chisq df P(> Chisq)
 7e−04 1      0.9796
```

Event type 2

| Chisq | df | P(> Chisq) |
|---|---|---|
| 4.4474 | 1 | 0.0350 |

The $\chi^2$ statistic is 5.4 and its $p$-value is greater than 0.05. This means that parameter associated with the variable cd4 can be considered as 0 statistically in overall. However, the $\chi^2$ statistic for each event type is 5.4 with $p = 0.07$ and 4.4 with $p = 0.04$ respectively. The fitted model including the variable cd4 is not statistically significant in favor of the model not including the variable cd4 for the event type 1; however, the variable cd4 is statistically significant in the model for the event type 2.

## 2.5 Discussion

The package intccr provides a convenient and versatile tool for robust semiparametric regression analysis of the CIF based on interval-censored competing risk data. The package supports a large class of models for the CIF, including the proportional odds and the Fine–Gray proportional subdistribution hazards model as special cases. It also provides semiparametrically efficient regression coefficient estimates. To the best of my knowledge, the only other available software for the analysis of interval-censored competing risks data is the R package MIICD. That package utilizes Rubin's multiple imputation approach to deal with the unobserved event times. However, it is well known that Rubin's variance estimator is biased in cases where the imputation and the analysis models are uncongenial, a scenario that occurs frequently in practice (Meng, 1994). In addition, the MIICD package does not provide

semiparametrically efficient regression coefficient estimates and it only supports the Fine–Gray proportional subdistribution hazards model, whose interpretation is more difficult compared to the proportional odds model.

The package intccr follows the guideline for the selection of the number of knots in Section 2.1 in Bakoyannis, Yu, & Yiannoutsos (2017). Briefly, the number of internal knots for the B-spline is $N = \lfloor k \times n^{1/3} \rfloor$ where $k \in [0.5, 1]$ is a parameter that is specified by the user and $n$ is a sample size. For more details about the justification of the selection of knots, please see Section 2.1 in Bakoyannis, Yu, & Yiannoutsos (2017). Regarding the maximum number of regression coefficients to be estimated (or equivalently the number of covariates) for each event type, the following rule of thumb is suggested:

$$\text{Maximum number of covariates} = \left\lfloor \frac{\min(n_1, n_2)}{10} \right\rfloor$$

where $n_j$ for $j = 1, 2$ is the number of observations with event type $j$.

It has to be noted that, in many cases, there is no obvious interval censoring. However, the event time is typically measured in days, and the exact time of the event is not recorded. In this case, assuming that the true event time is a continuous, the exact event time is still interval censored, with the width of the censoring interval being 1 day. Such cases, can still be analyzed using the package intccr by setting $V = X - 0.5$ days and $U = X + 0.5$ days, where $X$ is the recorded event time in days. This example occurs in Dementia studies, where the time to Dementia is interval-censored while the time to death is more precisely recorded in days. Such data can be easily analyzed using the package intccr.

The simulations were ran on Intel(R) Core(TM) i5-2400 CPU 3.10GHz with 8 GB ram. Maximum number of available cores in parallel computing was 3. It is expected that the users having higher specification of their computer may see more timely efficient results.

The intccr package introduced in this Chapter provides the estimated CIFs for a particular combination of covariate values. This quantity is very appealing for graphical illustration. Also, the package provides data management functionality to reformat data sets provided in a long format (i.e. data sets with multiple lines per subject), and turn them into the wide (single-line per subject) format required by the package. One limitation of the intccr package is that, for the time being, it only allows for two event types. The plan is to update the package intccr to allow for more than two event types in the near future. The package is freely available for download from the CRAN website https://cran.r-project.org/web/packages/intccr/index.html.

# Chapter 3

## Semiparametric regression on cumulative incidence function with interval-censored competing risks data and missing cause of failure

Although the versatile tool has been developed to provide the semiparametric regression analysis of the CIF based on interval-censored competing risk data in Chapter 2, event types (i.e. causes of failure) are often partially observed. Ignoring data that contain missing event types results in a biased estimator. In this chapter, an augmented inverse probability weighted (AIPW) sieve maximum likelihood estimator is proposed as a way of the analysis of interval-censored competing risk data in the presence of missing event types. The estimator imposes weaker than usual missing at random assumptions by allowing for the inclusion of auxiliary variables that are potentially associated with the probability of missingness. It is shown that the proposed estimator is doubly robust, in the sense that it is consistent even if either the model for the probability of missingness or the model for the probability of the event type is misspecified. Extensive Monte Carlo simulation studies show a commentable performance of the proposed method even under a large amount of missing event types. The method is illustrated by using data from an HIV cohort study in sub-Saharan Africa, where a significant portion of events types is missing. The proposed method can be readily implemented in the R package intccr.

## 3.1 Introduction

Competing risks data are frequently encountered in cohort studies and clinical trials, and they refer to the situation where study participants are at risk of multiple mutually-exclusive events (Kalbfleisch & Prentice, 2011; Putter et al., 2007; Bakoyannis & Touloumi, 2012). The competing risks framework also includes situations where the scientific focus is on the first occurring event among multiple endpoints (Putter et al., 2007; Bakoyannis & Touloumi, 2012). Concurrently, competing risks data are interval-censored when the exact event time is not observed precisely and only known to lie within periodic study visits especially in clinical trials or longitudinal studies (Sun, 2006; Klein & Moeschberger, 2013). In the competing risks framework, the CIF and the CSH function are the basic identifiable quantities from the observed data (Kalbfleisch & Prentice, 2011; Putter et al., 2007; Bakoyannis & Touloumi, 2012; Koller et al., 2012; Andersen et al., 2012). The CIF is the cumulative probability of a particular event type occurring by a certain time in the presence of remaining event types, while the CSH estimates the instantaneous occurrence rate of a specific event type in the presence of the others. It is important to note that the CIF explicitly quantifies clinical prognosis and is useful for prediction purposes (Koller et al., 2012; Bakoyannis et al., 2017). Throughout Chapter 3, it is focused on making inferences about the CIF.

In many settings, the event types for some individuals are missing due to the usual non-response or by the study design. This introduces an additional complexity in the analysis of the CIF with right-censored competing risks data, since missingness can lead to a bias and a loss of statistical efficiency. Several researchers have

addressed the issue of nonparametric analysis of the CIF with missing at random (MAR) event types (Little & Rubin, 2002). The issue of nonparametric inference about the CIF with missing event types was addressed by Lee et al. (2014). For this, Lee et al. (2014) used the multiple imputation approach proposed by Lu & Tsiatis (2001) and proposed approaches for the calculation of pointwise confidence intervals and non-parametric two-sample tests. Recently, Bakoyannis et al. (2019) proposed a more general approach for nonparametric inference about transition probabilities in nonhomogeneous Markov processes with missing absorbing states, which is applicable to the competing risks problem with missing event types. This approach is based on a nonparametric pseudolikelihood estimator and provides a way to construct simultaneous confidence bands for the CIF. The latter estimator can be utilized within the framework of the nonparametric tests by Bakoyannis (2020) to perform two-sample comparisons for the CIF in settings with missing event types. It is of note that the aforementioned tests are applicable even in cases where the CIFs under comparison cross at one or more time points. The issue of semiparametric analysis of the CIF with missing event types under a MAR assumption has also received attention in the literature. Bakoyannis et al. (2010) utilized Rubin's multiple imputation methodology to estimate the parameters of the semiparametric Fine–Gray model for the CIF (Fine & Gray, 1999). Moreno-Betancur & Latouche (2013) used inverse probability weighting to analyze competing risks data with missing event types using the Klein–Andersen pseudo-value approach (Klein & Andersen, 2005) for making an inference about the CIF in a general class of semiparametric models. L. Mao & Lin (2017) proposed an EM algorithm for semiparametric analysis of the CIF under the general class of semiparametric transformation models. This approach, which provides semi-

parametric efficient estimation for the regression coefficients, allows for missing event types.

Another frequently encountered problem in studies with competing risks time-to-event data is interval censoring (Sun, 2006). Interval censoring refers to the situation where the actual event time is not precisely observed but is only known to lie between two observation times such as clinic visits. To address this problem in the framework of competing risks with fully observed event types, Hudgens et al. (2001) proposed nonparametric maximum likelihood and pseudolikelihood estimators for the CIF under interval censoring. Li (2016) proposed a semiparametric B-spline-based sieve maximum likelihood estimation approach (Y. Zhang et al., 2010) for making an inference under the Fine–Gray proportional subdistribution hazards model. Bakoyannis et al. (2017) used the semiparametric B-spline-based sieve maximum likelihood approach (Y. Zhang et al., 2010) to the more general class of semiparametric odds rate transformation models for the CIF and explicitly accounted for the boundedness constraint of the CIFs for the different event types. L. Mao et al. (2017) proposed an EM-algorithm for semiparametric analysis under the general class of semiparametric transformation models for the CIF. All three aforementioned methods for semiparametric analysis of the CIF under interval censoring provide semiparametrically efficient estimators of the regression coefficients. However, only the approach by Bakoyannis et al. (2017) can be readily implemented in practice using the R package intccr (Park et al., 2019).

Some studies with competing risks data involve both missing event types and interval censoring. This is the case for a motivating East-Africa International Epidemiologic Databases to Evaluate AIDS (EA-IeDEA) study. One of the aims of

this study is to evaluate potential prognostic factors for disengagement from HIV care and death while in care (i.e. before disengagement) after ART initiation. The nature of this scientific question requires a model for the CIF. However, a significant complication in this study is the substantial number of deaths under-reporting which is common in resource-limited settings. To address this problem, EA-IeDEA investigators have implemented a double-sampling design where a small subset of individuals who miss their clinic visit is actively outreached in the community and their vital status is eventually ascertained. This double-sampling design leads to a missing event type problem since the event type for the non-outreached individuals who miss a clinic visit is unobserved and could be either (unreported) death or disengagement from care. Moreover, the working definition of disengagement used by the clinical investigators within EA-IeDEA is being without a clinic visit for three months. However, the actual time of disengagement is not precisely observed but is only known to lie within the three-month window without clinic visits. Therefore, the event time is interval-censored. To the best of my knowledge, only Do & Kim (2017) and L. Mao et al. (2017) have considered the problem of semiparametric analysis of the CIF with both interval-censored competing risks data and missing event types. Do & Kim (2017) utilized Rubin's multiple imputation to deal with missingness in the framework of the pseudo-value approach for the CIF (Klein & Andersen, 2005). However, multiple imputation can provide biased estimates when the imputation model is misspecified. Moreover, Rubin's variance estimator is biased when the imputation model is misspecified (N. Wang & Robins, 1998; Robins & Wang, 2000) or under uncongeniality between the imputation and analysis models (Meng, 1994; N. Wang & Robins, 1998; Robins & Wang, 2000). In the simulation results presented by Do

& Kim (2017), there are cases where Rubin's variance estimates exhibit a relative bias close to 20%. Incorporation of auxiliary variables that are potentially associated with the probability of missingness in the imputation model is a cause of such uncongeniality. Nevertheless, accounting for such auxiliary variables is crucial in many settings in order to make the key MAR assumption more plausible (Collins et al., 2001; Lu & Tsiatis, 2001; Bakoyannis et al., 2019). L. Mao et al. (2017) allowed for missing event types in their method for interval-censored data. Missingness in this case is being accounted for in the expectation step of the EM algorithm by L. Mao et al. (2017). However, this approach does not incorporate auxiliary variables which may be related to the probability of a missing event type. Moreover, the computation algorithm by L. Mao et al. (2017), which simultaneously provides estimates for the models for all event types, does not explicitly incorporate the nonlinear inequality constraint that the sum of the CIFs for all event types is bounded by one. This can lead to non-convergence problems in practice. Last but not least, neither Do & Kim (2017) nor L. Mao et al. (2017) approaches are readily available using off-the-shelf software.

In this chapter, the main limitations of the currently available methods are addressed among semiparametric analysis of the CIF with interval-censored competing risks data and missing event types. More precisely, an AIPW technique is proposed to account for missing event types within the B-spline-based sieve maximum likelihood framework for interval-censored competing risks data by Bakoyannis et al. (2017). This approach, which allows for auxiliary covariates, utilizes a parametric logistic model for the probability of a missing event type and a binary logistic or multinomial model for the probability of the event type. Double robustness of the proposed esti-

mator is shown that the estimator is consistent even when either the model for the probability of missingness or the probability of the event type is misspecified. Moreover, a new function ciregic_aipw in the R package intccr is introduced. It can be used to readily implement the proposed approach in practice. Simulation studies show that the highlighted performance of the proposed method even when the model for the event type is misspecified, and that the naïve complete case analysis can provide seriously biased estimates. Also, Rubin's multiple imputation procedure for missing event types (Bakoyannis et al., 2010; Do & Kim, 2017) can provide biased estimates when the imputation model is misspecified. The proposed method is applied to the data from the motivating EA-IeDEA study.

## 3.2 Methods

### 3.2.1 Data and model

Let $(T_i, C_i)$ be the pair of event time and event type of the $i$th individual, $i = 1, \ldots, n$, where $C_i \in \{1, 2, \ldots, k\}$ and $k < \infty$. Also, let $V_i$ be the last observation time prior to the occurrence of the event and $U$ the first observation time after the event onset where $V_i, U_i \in [a, b]$, with $0 < a < b < \infty$. Next, $\Delta_i$ is defined as the event indicator that the $i$th individual experience an event during the study period. It is clear that $\Delta_i = 0$ indicates that the $i$th individual is right-censored. The censoring indicators are also defined: the interval censoring indicator $\Delta_i^{(1)}$ and the left censoring indicator $\Delta_i^{(2)}$, which satisfy $\Delta_i = \Delta_i^{(1)} + \Delta_i^{(2)}$. The indicator the $i$th individual

experience the $j$th event type, $j = 1, \ldots, k$, is defined as $\Delta_{ij}^{(1)} = \Delta_i^{(1)} I(C_i = j)$ for an interval-censored case and as $\Delta_{ij}^{(2)} = \Delta_i^{(2)} I(C_i = j)$ for a left-censored case. Note that $V_i$ is only observed for interval-censored and right-censored observations. For the left-censored observations, the last examination time prior to the event occurrence is trivially 0. Also, let $Z_i \subset \mathbb{R}^d$ be the vector of covariates of interest. Here, the observation times are independent of $(T_i, C_i)$ conditionally on $Z_i$ (independent interval censoring), and that and the distribution of $(V_i, U_i)$ does not contain the parameters of interest (non-informative interval censoring). In a situation where some event types are missing, define $R_i$ to be the response (i.e. non-missingness) indicator for the event type, with $R_i = 1$ if $\Delta_{ij}$ has been observed and $R_i = 0$ otherwise. In what follows it is assumed that $R_i = 0$ implies that $\Delta_i = 1$, which means that the right censoring status is always observed. Also, the indicators $\Delta_i^{(1)}$ and $\Delta_i^{(2)}$ are not subject to missingness. Finally, denote the vector of auxiliary variables that may be related to $R_i$ by $A_i$. Under this setup, the observable data based on an i.i.d. sample are $X_i = \left(R_i, \Delta_i^{(1)}, \Delta_i^{(2)}, R_i \underline{\Delta}_i^{(1)}, R_i \underline{\Delta}_i^{(2)}, \Delta_i^{(1)} V_i, U_i, Z_i, A_i\right)$, for $i = 1, \ldots, n$, where $\underline{\Delta}_i^{(l)} = \left(\Delta_{i1}^{(l)}, \ldots, \Delta_{ik}^{(l)}\right)^T$, $l = 1, 2$. In this work, the following MAR assumption is imposed

$$\begin{aligned} &\Pr\left(R_i = 1 | \Delta_i = 1, \Delta_i^{(1)}, \Delta_i^{(2)}, \underline{\Delta}_i^{(1)}, \underline{\Delta}_i^{(2)}, \Delta_i^{(1)} V_i, U_i, Z_i, A_i\right) \\ &\qquad = \Pr\left(R_i = 1 | \Delta_i = 1, U_i, Z_i, A_i\right), \end{aligned} \tag{3.1}$$

that is, given the observed data and the auxiliary variables, the probability of response is independent of the incomplete event type indicators $\left(\underline{\Delta}_i^{(1)}, \underline{\Delta}_i^{(2)}\right)$. Incorporating the auxiliary variables $A_i$ leads to a weaker MAR assumption (Lu & Tsiatis, 2001;

Bakoyannis et al., 2019). Note that, for simplicity, it is assumed that $R_i$ depends on the event diagnosis time $U_i$ and not the last observation time $V_i$ prior to the occurrence of the event. This is a plausible assumption in practice.

In this chapter, the CIF conditional on $Z = z$ is studies. It is defined as

$$F_j(t; z) = \Pr(T \leq t, C = j | Z = z), \quad j = 1, \ldots, k.$$

A natural choice for a class of models for the CIF is the class of semiparametric transformation models. Under this class, the CIF is expressed as

$$g_j \left\{ F_j(t; z) \right\} = \phi_j(t) + \beta_j^T z, \quad j = 1, \ldots, k,$$

where $g_j$ is a known and increasing link function, $\phi_j$ is an unspecified strictly increasing function, and $\beta_j$ is a vector of regression coefficients (Zeng et al., 2006; Bakoyannis et al., 2017; L. Mao et al., 2017). A special subset of this class is the class of the generalized odds rate transformation models defined as (Jeong & Fine, 2006; Dabrowska & Doksum, 1988; Scharfstein et al., 1998; Bakoyannis et al., 2017)

$$g_j \left( F_j; \alpha_j \right) = \begin{cases} \log \left\{ \dfrac{(1 - F_j)^{-\alpha_j} - 1}{\alpha_j} \right\} & \text{if } 0 < \alpha_j < \infty \\ \\ \log \left\{ -\log \left( 1 - F_j \right) \right\} & \text{if } \alpha_j = 0 \end{cases}$$

Special cases of this class are the Fine–Gray proportional subdistribution hazards model (Fine & Gray, 1999) when $\alpha_j = 0$ and the proportional proportional odds model when $\alpha_j = 1$.

### 3.2.2 Semiparametric estimation

When there are no missing event types, the likelihood function can be expressed as

$$
\begin{aligned}
L(\theta) \propto \prod_{i=1}^{n} \Bigg[ & \left[ \prod_{j=1}^{k} \left\{ F_j(U_i; Z_i, \theta_j) - F_j(V_i; Z_i, \theta_j) \right\}^{\Delta_{ij}^{(1)}} \right] \\
& \times \left[ \prod_{j=1}^{k} \left\{ F_j(U_i; Z_i, \theta_j) \right\}^{\Delta_{ij}^{(2)}} \right] \\
& \times \left\{ 1 - \sum_{j=1}^{k} F_j(V_i; Z_i, \theta_j) \right\}^{1-\Delta_i} \Bigg]
\end{aligned}
\tag{3.2}
$$

where $\theta = \left(\theta_1^T, \theta_2^T, \cdots, \theta_k^T\right)^T$ are the unknown parameters(Bakoyannis et al., 2017). Note that $\theta_j = \left(\phi_j, \beta_j^T\right)^T$. The maximization of likelihood (3.2) can be performed over the sieve space $\Theta_n = \mathcal{B}^k \times \Phi_n^k$, where $\mathcal{B} \subset \mathbb{R}^d$ and

$$
\Phi_n(\gamma, N_n, m) = \left\{ \phi : \phi(t; \gamma) = \sum_{s=1}^{N_n+m} \gamma_s B_{s,m}(t), \gamma \in \mathbb{R}^{N_n+m}, \gamma_1 < \cdots < \gamma_{N_n+m} \right\}
$$

is the B-spline sieve space with $N_n$ and $m$ denoting the number of internal knots and the order of the B-spline, $\gamma = (\gamma_1, \ldots, \gamma_{N_n+m})^T$ is the unknown control point vector of the B-spline coefficients, and $t \in [a, b]$. The number of internal knots $N_n$ is selected to satisfy $N_n \approx n^\nu$ such that $\max_{1 \le l \le N_n+1} |w_l - w_{l-1}| = O(n^{-\nu})$, where $w_l$ is the place of the $l$th knot (Bakoyannis et al., 2017). The optimal choice for $\nu$, in order to achieve the optimal rate of convergence of the B-spline estimator of $\phi_j$, is $\nu = 1/(1 + 2p)$, where $p$ is the degree of smoothness of the true underlying functions $\phi_j$ for $j = 1, \ldots, k$ (Bakoyannis et al., 2017). The knots are placed in

the percentiles of the distribution of the observation times $(V_i, U_i)$. Note that the restriction $\gamma_1 < \cdots < \gamma_{N_n+m}$ imposes a monotonicity constraint on the B-spline functions. Moreover, during the maximization the constraint

$$\max_z \left\{ \sum_{j=1}^{k} F_j(b; z, \theta_j) \right\} < 1$$

is also imposed. This is crucial since ignoring this constraint may lead to non-convergence issues. The estimation process can be readily implemented using the function ciregic in the R package intccr (Park et al., 2019).

In the presence of missing event types, the likelihood function (3.2) cannot be evaluated for the missing cases. To deal with this issue, the AIPW method which is similar to the one by Gao & Tsiatis (2005) is used. For this let

$$\rho\left(O_i, \xi^*\right) \equiv \Pr\left(R_i = 1 | \Delta_i = 1, U_i, Z_i, A_i; \xi^*\right),$$

where $O_i = (U_i, Z_i, A_i)^T$, be the parametric response (or equivalently the missingness) model, where $\xi^*$ is a finite dimensional parameter. Since the model $\rho$ may be misspecified, $\xi^*$ denotes the minimizer of the Kullback–Leibler divergence between the assumed and the true model. Similarly, define

$$\pi_j(O_i, \psi^*) = \Pr(C_i = j | \Delta_i = 1, U_i, Z_i, A_i; \psi^*), \quad j = 1, \ldots, k, \tag{3.3}$$

to be the parametric model for the $j$th event type, where $\psi^*$ is a finite-dimensional parameter as before. Again, $\psi^*$ is the minimizer of the Kullback–Leibler divergence between the assumed and the true model. The implicit assumption in model (3.3) is

that the probability of the $j$th event type does not depend on whether the observation is interval-censored or left-censored, and, also, that $C_i$ is conditionally independent of the last examination time prior to the event diagnosis time for the interval-censored cases. A natural choice of the model $\pi_j$ is the binary logistic (if $k = 2$) or the multinomial logistic (if $k > 2$) model.

The first stage of the analysis involves the estimation of $\xi^*$ using the observations with $\Delta_i = 1$, and $\psi^*$ based on the observations with $R_i = 1$ and $\Delta_i = 1$ (i.e. cases with an observed event type). Estimation in both cases is conducted via (parametric) maximum likelihood. Under the MAR assumption (3.1), the second stage of the analysis consists of maximizing the objective function under the AIPW framework

$$\begin{aligned}\tilde{l}(\theta; \hat{\xi}_n, \hat{\psi}_n) = \frac{1}{n} \sum_{i=1}^{n} \Bigg[ & \sum_{j=1}^{k} \tilde{\Delta}_{ij}^{(1)}(\hat{\xi}_n, \hat{\psi}_n) \log \{F_j(U_i; Z_i, \theta_j) - F_j(V_i; Z_i, \theta_j)\} \\ & + \sum_{j=1}^{k} \tilde{\Delta}_{ij}^{(2)}(\hat{\xi}_n, \hat{\psi}_n) \log \{F_j(U_i; Z_i, \theta_j)\} \\ & + (1 - \Delta_i) \log \left\{ 1 - \sum_{j=1}^{k} F_j(V_i; Z_i, \theta_j) \right\} \Bigg]\end{aligned}$$

where

$$\tilde{\Delta}_{ij}^{(l)}(\hat{\xi}_n, \hat{\psi}_n) = \frac{R_i}{\rho(O_i; \hat{\xi}_n)} \Delta_{ij}^{(l)} - \frac{R_i - \rho(O_i; \hat{\xi}_n)}{\rho(O_i; \hat{\xi}_n)} \pi_j(O_i; \hat{\psi}_n), \qquad l = 1, 2.$$

This objective function corresponds to the AIPW version of the logarithm of likelihood (3.2), multiplied by $1/n$ which does not affect the maximizer but is convenient for the consistency proof. Maximization is performed over the sieve space $\Theta_n$ under

the constraints described above for the case without missing event types. The resulting estimator is denoted as $\hat{\theta}_n$. This approach can be readily implemented using the function ciregic_aipw in the R package intccr. In Chapter 3.5, an illustrative example is provided. The example describes how to use the ciregic_aipw to perform the proposed AIPW methodology.

### 3.2.3 Properties of the proposed estimator

The proposed estimator possesses the double robustness property, that is it is consistent if either $\rho(O_i;\xi^*)$ or $\pi_j(O_i;\psi^*)$, $j=1,\ldots,k$, is correctly specified. Letting $\Theta$ denote the true (infinite-dimensional) parameter space, consistency is proved in the $L^2$-metric $d$ which is defined as follows

$$d(\theta^{(1)},\theta^{(2)}) = \left( \sum_{j=1}^{k} \left\| \beta_j^{(1)} - \beta_j^{(2)} \right\|^2 + \sum_{j=1}^{k} \left\| \phi_j^{(1)} - \phi_j^{(2)} \right\|_\Phi^2 \right)^{\frac{1}{2}},$$

for $\theta^{(1)},\theta^{(2)} \in \Theta$, where $\|\cdot\|$ is the Euclidean $L^2$-norm and

$$\left\| \phi_j^{(1)} - \phi_j^{(2)} \right\|_\Phi^2 = E\left\{ \phi_j^{(1)}(V) - \phi_j^{(2)}(V) \right\}^2 + E\left\{ \phi_j^{(1)}(U) - \phi_j^{(2)}(U) \right\}^2.$$

Now, let $\theta_0$ denote the true parameter values. Theorem 1 ensures the double robustness of the proposed estimator.

**Theorem 1.** *(Double robustness) Suppose that the interval censoring is independent and non-informative conditionally on the covariates $Z$, the MAR assumption* (3.1) *is satisfied, the regularity conditions are satisfied, and $N_j = O(n^\nu)$, $j=1,\ldots,k$,*

*where* $\nu$ *satisfies* $1/\{2(1+p)\} < \nu < 1/(2p)$. *Then, if either* $\rho(O_i;\xi^*)$ *or* $\pi_j(O_i;\psi^*)$, $j = 1,\ldots,k$, *is correctly specified,*

$$d(\hat{\theta}_n, \theta_0) \xrightarrow{p} 0.$$

*Proof.* To show the double robustness property of the proposed estimator empirical process theory will be used (Kosorok, 2008; Van der Vaar & Wellner, 1996). The standard empirical process notations are used: $Pf = \int_{\mathcal{X}} f(x)dP(x)$ and $\mathbb{P}_n = n^{-1}\sum_{i=1}^{n} f(X_i)$ for a measurable function $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is the sample space. Also, let $K$ be a generic constant, that could differ from place to place. Now, define the functions

$$\begin{aligned}\tilde{l}_{\theta,\xi,\psi}(X) = &\sum_{j=1}^{k} \tilde{\Delta}_j^{(1)}(\xi,\psi)\log\{F_j(U;Z,\theta_j) - F_j(V;Z,\theta_j)\} \\ &+ \sum_{j=1}^{k} \tilde{\Delta}_j^{(2)}(\xi,\psi)\log\{F_j(U;Z,\theta_j)\} + (1-\Delta)\log\left\{1 - \sum_{j=1}^{k} F_j(V;Z,\theta_j)\right\}\end{aligned}$$

for a generic observation $X \in \mathcal{X}$, and

$$\begin{aligned}l_{\theta}(X) = &\sum_{j=1}^{k} \Delta_j^{(1)}\log\{F_j(U;Z,\theta_j) - F_j(V;Z,\theta_j)\} \\ &+ \sum_{j=1}^{k} \Delta_j^{(2)}\log\{F_j(U;Z,\theta_j)\} + (1-\Delta)\log\left\{1 - \sum_{j=1}^{k} F_j(V;Z,\theta_j)\right\}\end{aligned}$$

Note that based on this notation, obtaining the proposed AIPW sieve estimator of $\theta$ requires to maximize $\mathbb{P}_n\tilde{l}_{\theta,\hat{\xi}_n,\hat{\psi}_n} \equiv \tilde{\mathbb{M}}_n\left(\theta;\hat{\xi}_n,\hat{\psi}_n\right)$. If there were no missing event types one would need to maximize $\mathbb{P}_n l_\theta \equiv \mathbb{M}_n(\theta)$ (Bakoyannis et al., 2017). The latter

objective function can be seen as an estimator of $Pl_\theta \equiv \mathbb{M}(\theta)$. In this work, similarly to Bakoyannis et al. (2017), assume the following regularity conditions:

C1. $Z$ and $A$ are bounded in the sense that there exists a $K \in (0, \infty)$ such that $\Pr(\|Z\| \vee \|A\| \leq K) = 1$. Moreover, $E(ZZ^T)$ is a non-singular.

C2. For $j = 1, 2, \ldots, k$, $\beta_{0,j} \in \mathcal{B}$, where $\mathcal{B}$ is a compact subset of $\mathbb{R}^d$.

C3. There exists $\eta > 0$ such that $P(U - V \geq \eta) = 1$ and the unions of the supports of $U$ and $V$ are contained in $[a, b]$ for $0 < a < b < \infty$. Also, and $0 < \min_{j \in \{1,2,\ldots,k\}} F_j(a; Z = 0) < \sum_{j=1}^{k} F_j(b; Z = 0) < 1$.

C4. $\phi_{0,j} \in \Phi$, where $\Phi$ is a set of functions whose $p$th derivative is bounded in $[a, b]$ for $p \geq 1$, and the first derivative of $\phi_{0,j}$ is strictly positive and continuous on $[a, b]$, for $j = 1, \ldots, k$.

C5. The joint density of $(V, U)$ conditional on $Z$ has bounded partial derivatives with respect to $(v, u)$, whose bounds do not depend on $(v, u, z)$.

C6. There exists $\kappa$ for $0 < \kappa < 1$ such that $a^T \mathrm{Var}(Z|V)a \geq \kappa a^T E\left(ZZ^T|V\right)a$ and $a^T \mathrm{Var}(Z|U)a \geq \kappa a^T E\left(ZZ^T|U\right)a$ a.s. for all $a \in \mathbb{R}^d$.

C7. The parametric model $\pi_j(O_i; \psi)$, $j = 1, \ldots, k$, is continuously differentiable in $\psi$. Moreover, $\hat{\psi}_n \xrightarrow{p} \psi^*$ and $\sqrt{n}\left(\hat{\psi}_n - \psi^*\right) = n^{-1/2} \sum_{i=1}^{n} \omega_i + o_p(1)$, where $E\omega_1 = 0$ and $E\|\omega_1\|^2 < \infty$.

C8. The parametric model $\rho(O_i; \xi)$ is continuously differentiable in $\xi$ and satisfies $\rho(O_i; \xi) > 0$ a.s.. Moreover, $\hat{\xi}_n \xrightarrow{p} \xi^*$ and $\sqrt{n}\left(\hat{\xi}_n - \xi^*\right) = n^{-1/2} \sum_{i=1}^{n} \phi_i + o_p(1)$, where $E\phi_1 = 0$ and $E\|\phi_1\|^2 < \infty$.

Conditions C1–C6 guarantee the consistency of $\hat{\theta}_n$ and the $\sqrt{n}$-consistency and asymptotic normality of the regression coefficient estimator for the B-spline sieve maximum likelihood estimator for interval-censored competing risks data without missing

event types (Bakoyannis et al., 2019). Conditions C7 and C8 are required for the proposed AIPW sieve maximum likelihood estimator for dealing with missing event types. These additional conditions are satisfied if the parametric models $\rho$ and $\pi_j$, $j = 1, \ldots, k$, are specified as regular generalized linear models and estimated through maximum likelihood. The positivity condition $\rho(O_i; \xi) > 0$ a.s. is expected to be satisfied in general in practice.

To show the consistency of the proposed estimator the following conditions should be proved:

(i) $\sup_{\theta \in \Theta_n} \left| \tilde{\mathbb{M}}_n(\theta; \hat{\xi}_n, \hat{\psi}_n) - \mathbb{M}(\theta) \right| \equiv \left\| \tilde{\mathbb{M}}_n(\theta; \hat{\xi}_n, \hat{\psi}_n) - \mathbb{M}(\theta) \right\|_{\Theta_n} \xrightarrow{p} 0$

(ii) $\sup_{\theta : d(\theta, \theta_0) \geq \epsilon} \mathbb{M}(\theta) < \mathbb{M}(\theta_0)$

(iii) The sequence of the estimators $\hat{\theta}_n$ satisfies

$$\tilde{\mathbb{M}}_n(\hat{\theta}_n; \hat{\xi}_n, \hat{\psi}_n) \geq \tilde{\mathbb{M}}_n(\theta_0; \hat{\xi}_n, \hat{\psi}_n) - o_p(1)$$

For condition (i),

$$\begin{aligned} \|\tilde{\mathbb{M}}_n(\theta; \hat{\xi}_n, \hat{\psi}_n) - \mathbb{M}(\theta)\|_{\Theta_n} &\leq \|\tilde{\mathbb{M}}_n(\theta; \hat{\xi}_n, \hat{\psi}_n) - \tilde{\mathbb{M}}_n(\theta; \xi^*, \psi^*)\|_{\Theta_n} \\ &\quad + \|\tilde{\mathbb{M}}_n(\theta; \xi^*, \psi^*) - \mathbb{M}_n(\theta)\|_{\Theta_n} \\ &\quad + \|\mathbb{M}_n(\theta) - \mathbb{M}(\theta)\|_{\Theta_n} \\ &\equiv A_n + B_n + C_n. \end{aligned} \tag{3.4}$$

For the first term,

$$\begin{aligned} A_n &\leq \sum_{j=1}^{k} \left\| \frac{1}{n} \sum_{i=1}^{n} \left\{ \tilde{\Delta}_{ij}^{(1)}(\hat{\xi}_n, \hat{\psi}_n) - \tilde{\Delta}_{ij}^{(1)}(\xi^*, \psi^*) \right\} \log \left\{ F_j(U_i; Z_i, \theta_j) - F_j(V_i; Z_i, \theta_j) \right\} \right\|_{\Theta_n} \\ &\quad + \sum_{j=1}^{k} \left\| \frac{1}{n} \sum_{i=1}^{n} \left\{ \tilde{\Delta}_{ij}^{(2)}(\hat{\xi}_n, \hat{\psi}_n) - \tilde{\Delta}_{ij}^{(2)}(\xi^*, \psi^*) \right\} \log \left\{ F_j(U_i; Z_i, \theta_j) \right\} \right\|_{\Theta_n} \\ &\leq \sum_{j=1}^{k} \left[ \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \tilde{\Delta}_{ij}^{(1)}(\hat{\xi}_n, \hat{\psi}_n) - \tilde{\Delta}_{ij}^{(1)}(\xi^*, \psi^*) \right\} \right| \right. \\ &\qquad \left. + \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \tilde{\Delta}_{ij}^{(2)}(\hat{\xi}_n, \hat{\psi}_n) - \tilde{\Delta}_{ij}^{(2)}(\xi^*, \psi^*) \right\} \right| \right] \end{aligned}$$

Under this inequality, conditions C1, C7, C8, and Taylor expansion it follows that $A_n \xrightarrow{p} 0$. For the second term,

$$\begin{aligned} B_n &\leq \sum_{j=1}^{k} \left\| \frac{1}{n} \sum_{i=1}^{n} \left\{ \tilde{\Delta}_{ij}^{(1)}(\xi^*, \psi^*) - \Delta_{ij}^{(1)} \right\} \log \left\{ F_j(U_i; Z_i, \theta_j) - F_j(V_i; Z_i, \theta_j) \right\} \right\|_{\Theta_n} \\ &\quad + \sum_{j=1}^{k} \left\| \frac{1}{n} \sum_{i=1}^{n} \left\{ \tilde{\Delta}_{ij}^{(2)}(\xi^*, \psi^*) - \Delta_{ij}^{(2)} \right\} \log \left\{ F_j(U_i; Z_i, \theta_j) \right\} \right\|_{\Theta_n} \\ &\leq \sum_{j=1}^{k} \left[ \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \tilde{\Delta}_{ij}^{(1)}(\xi^*, \psi^*) - \Delta_{ij}^{(1)} \right\} \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \tilde{\Delta}_{ij}^{(2)}(\xi^*, \psi^*) - \Delta_{ij}^{(2)} \right\} \right| \right] \\ &= \sum_{j=1}^{k} \left[ \left| E \left\{ \tilde{\Delta}_{ij}^{(1)}(\xi^*, \psi^*) - \Delta_{ij}^{(1)} \right\} \right| + \left| E \left\{ \tilde{\Delta}_{ij}^{(2)}(\xi^*, \psi^*) - \Delta_{ij}^{(2)} \right\} \right| \right] + o_p(1). \quad (3.5) \end{aligned}$$

After trivial algebra, the first expectation in the right side of (3.5) is

$$\begin{aligned} E\left[\tilde{\Delta}_{ij}^{(1)}(\xi^*,\psi^*)-\Delta_{ij}^{(1)}\right] &= E\left[\frac{R_i-\rho(O_i;\xi^*)}{\rho(O_i;\xi^*)}\left\{\Delta_{ij}^{(1)}-\pi_j(O_i;\psi^*)\right\}\right] \\ &= E\Bigg[\frac{E(R_i|O_i)-\rho(O_i;\xi^*)}{\rho(O_i;\xi^*)} \\ &\qquad \times\left\{E(\Delta_{ij}^{(1)}|O_i)-\pi_j(O_i;\psi^*)\right\}\Bigg]. \end{aligned}$$

If either $\rho(O_i;\xi^*)$ or $\pi_j(O_i;\psi^*)$ is correctly specified, that is if either $E(R_i|O_i) = \rho(O_i;\xi^*)$ a.s. or $E(\Delta_{ij}^{(1)}|O_i) = \pi_j(O_i;\psi^*)$ a.s., then, in light of condition C8, it follows that $E\left[\tilde{\Delta}_{ij}^{(1)}(\xi^*,\psi^*)-\Delta_{ij}^{(1)}\right]=0$. Similarly, if either $\rho(O_i;\xi^*)$ or $\pi_j(O_i;\psi^*)$ is correctly specified, then $E\left[\tilde{\Delta}_{ij}^{(2)}(\xi^*,\psi^*)-\Delta_{ij}^{(2)}\right]=0$. Therefore, in light of (3.5), $B_n \xrightarrow{p} 0$. Finally, Bakoyannis et al. (2017) showed that $C_n \xrightarrow{p} 0$ and, thus, based on (3.4), condition (i) is satisfied.

Condition (ii) has been shown by Bakoyannis et al. (2017). Finally, by Taylor expansion and conditions C1, C7, and C8 it follows that $\tilde{\Delta}_{ij}^{(1)}\left(\hat{\xi}_n,\hat{\psi}_n\right)=\tilde{\Delta}_{ij}^{(1)}\left(\xi^*,\psi^*\right)+o_p(1)$ and $\tilde{\Delta}_{ij}^{(2)}\left(\hat{\xi}_n,\hat{\psi}_n\right)=\tilde{\Delta}_{ij}^{(2)}\left(\xi^*,\psi^*\right)+o_p(1)$. Thus, $\tilde{\mathbb{M}}_n\left(\hat{\theta}_n;\hat{\xi}_n,\hat{\psi}_n\right)-\tilde{\mathbb{M}}_n\left(\theta_0;\hat{\xi}_n,\hat{\psi}_n\right) = \tilde{\mathbb{M}}_n\left(\hat{\theta}_n;\xi^*,\psi^*\right)-\tilde{\mathbb{M}}_n\left(\theta_0;\xi^*,\psi^*\right)+o_p(1)$. Now, using the same arguments to those used in the consistency proof in Bakoyannis et al. (2017) leads to the conclusion that condition (iii) is satisfied. Therefore, $d(\hat{\theta}_n,\theta_0)\xrightarrow{p}0$. □

As in the case with interval-censored competing risks data without missing event types (Bakoyannis et al., 2019), $\nu = 1/(1+2p)$ is set. Using the conditions along with arguments similar to those used in Bakoyannis et al. (2019) it can be shown that the estimator $\hat{\beta}_n$ is $\sqrt{n}$-consistent and asymptotically normal. However, it is not feasible to claim semiparametric efficiency for $\hat{\beta}_n$. Variance estimation can

be based on nonparametric bootstrap method (G. Cheng et al., 2010). The function ciregic_aipw of the R package intccr has an argument to select the desired number of bootstrap replications for variance estimation (for more details on the use of ciregic_aipw Chapter 3.5).

## 3.3 Simulation studies

In order to evaluate the performance of the proposed estimator a series of Monte Carlo simulation experiments was conducted. It is considered that there are two event types, $C = 1$ and $C = 2$, and two covariates of interest, $Z_1$ simulated from the Bernoulli distribution with probability 0.4, and $Z_2$ simulated from the standard normal distribution. There is also an auxiliary variable depending on the true event type as $A = I(C = 1) + \epsilon$, where $\epsilon \sim N(0, 1)$. The competing risks data were generated under the proportional odds models:

$$F_j(t) = \frac{\exp\left\{\phi_j(t) + \beta_j^T Z\right\}}{1 + \exp\left\{\phi_j(t) + \beta_j^T Z\right\}}, \quad j = 1, 2,$$

where

$$\exp\left\{\phi_j(t)\right\} = -\frac{\tau_j}{\rho_j}\left\{1 - \exp(\rho_j t)\right\}$$

under the improper Gompertz distribution (Jeong & Fine, 2006). Similarly to Bakoyannis et al. (2019), each set of parameters was assumed as $(\tau_1, \rho_1) = (0.4, -0.6)$ and $(\tau_2, \rho_2) = (0.75, -0.5)$. The values for the regression coefficients were $\beta_0 =$

$(\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}) = (0.5, -0.3, -0.5, 0.3)$. To generate interval censoring a series of observation time points based on the exponential distribution with a hazard parameter being equal to 3 was simulated. This led on average in one clinic visit every four months. It was assumed that the maximum study period is at 3 years. The probability of non-missingness was assumed to be

$$\text{logit}\{\Pr(R = 1|O)\} = \xi_0 + 0.5U - 0.5Z_1 + 0.6Z_2 + \xi_4 A$$

where $(\xi_0, \xi_4)$ is equal to (.9, -.5), (.6, -.1), (.6, 0), (.55, .1), or (.4, .5). These choices led to approximately 30% missing event types. Also, the other scenarios were considered, where $(\xi_0, \xi_4)$ was set to (-.1, -.5), (-.35, -.1), (-.35, 0), (-.4, .1), or (-.6, .5), which led to approximately 50% missing event types. For each simulation scenario 1,000 data sets were simulated, and the sample sizes $n = 200$ and $n = 400$ were considered. In this simulation study, the proposed AIPW method was considered by assuming the following models

$$\text{logit}\{\rho(O, \xi)\} = \xi_0 + \xi_1 U + \xi_2 Z_1 + \xi_3 Z_2 + \xi_4 A,$$

and

$$\text{logit}\{\pi_1(O, \psi)\} = \psi_0 + \psi_1 U + \psi_2 Z_1 + \psi_3 Z_2 + \psi_4 A. \tag{3.6}$$

Note that the true model $\pi_1(O, \psi)$ has a very complicated form under the proportional odds model assumed in this simulation study and, thus, model (3.6) is misspecified. However, the model $\rho(O, \xi)$ is correctly specified. Standard error estimation was based on 100 bootstrap replications. Two alternative methods were considered to compare

the performance of the proposed estimator: one is the naïve complete case analysis (CC) and the other is the multiple imputation procedure (MI) by Bakoyannis et al. (2010) based on 5 imputations and under the imputation model (3.6), as alternative approaches to deal with missingness. The B-spline-based sieve maximum likelihood approach by Bakoyannis et al. (2017) as the complete data method was performed in both approaches. Standard error for the complete case analysis was based on 50 bootstrap replications, while for the MI procedure Rubin's rules and nonparametric bootstrap for the within imputation variance estimation were used.

Simulation results are presented in Tables 3.1 through 3.5. Based on the simulation results, the naïve complete case analysis provided regression parameter estimates with substantial bias as a result of selection bias. The degree of bias was more pronounced with a larger missingness percent and in scenarios where the effect of the auxiliary variable on the probability of missingness was non-zero. The MI approach also provided regression coefficient estimates exhibiting non-negligible bias, although this bias was lower, on absolute, compared to that from the complete case analysis. The bias in the MI approach is attributed to the misspecification of the imputation model. The proposed AIPW approach provided virtually unbiased regression parameter estimates in all cases, even though model was misspecified (3.6). This provides numerical evidence for the double robustness of the proposed AIPW approach. For this approach, the average of the standard error estimates is close to the corresponding Monte Carlo standard deviation of the estimations, and the empirical coverage probabilities are close to the nominal 0.95 level.

The average of the baseline CIF based on the proposed approach, along with the corresponding true baseline CIF, is presented in Figure 3.1 through Figure 3.5.

Table 3.1: Monte Carlo simulation with 1,000 replications when ($\xi_4 = 0$)

| 30% | $n = 200$ | | | | $n = 400$ | | | |
|---|---|---|---|---|---|---|---|---|
| of missing | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ |
| i. CC† | | | | | | | | |
| % bias | -13.554 | -19.139 | 4.131 | 11.211 | -16.505 | -20.347 | 0.247 | 10.416 |
| MCSD[1] | 0.336 | 0.172 | 0.325 | 0.164 | 0.244 | 0.119 | 0.238 | 0.113 |
| ASE[2] | 0.352 | 0.171 | 0.342 | 0.165 | 0.242 | 0.117 | 0.235 | 0.113 |
| ECP[3] | 0.960 | 0.924 | 0.965 | 0.938 | 0.930 | 0.912 | 0.946 | 0.934 |
| ii. MI‡ | | | | | | | | |
| % bias | -6.203 | -6.394 | -5.206 | -5.152 | -8.215 | -6.933 | -8.771 | -6.657 |
| MCSD[1] | 0.327 | 0.166 | 0.320 | 0.159 | 0.235 | 0.116 | 0.230 | 0.114 |
| ASE[2] | 0.340 | 0.166 | 0.338 | 0.165 | 0.235 | 0.116 | 0.233 | 0.114 |
| ECP[3] | 0.957 | 0.953 | 0.963 | 0.956 | 0.945 | 0.953 | 0.953 | 0.948 |
| iii. AIPW§ | | | | | | | | |
| % bias | 0.539 | 0.056 | 1.324 | 1.283 | -1.755 | 0.272 | -2.261 | 0.690 |
| MCSD[1] | 0.341 | 0.180 | 0.338 | 0.173 | 0.246 | 0.122 | 0.243 | 0.121 |
| ASE[2] | 0.352 | 0.178 | 0.353 | 0.176 | 0.241 | 0.121 | 0.239 | 0.119 |
| ECP[3] | 0.956 | 0.944 | 0.956 | 0.950 | 0.934 | 0.952 | 0.947 | 0.946 |
| 50% | $n = 200$ | | | | $n = 400$ | | | |
| of missing | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ |
| i. CC† | | | | | | | | |
| % bias | -27.864 | -37.548 | 8.786 | 22.740 | -29.846 | -38.894 | 3.982 | 22.015 |
| MCSD[1] | 0.400 | 0.201 | 0.394 | 0.193 | 0.288 | 0.140 | 0.283 | 0.130 |
| ASE[2] | 0.434 | 0.204 | 0.420 | 0.195 | 0.292 | 0.137 | 0.280 | 0.132 |
| ECP[3] | 0.961 | 0.916 | 0.964 | 0.937 | 0.930 | 0.851 | 0.944 | 0.924 |
| ii. MI‡ | | | | | | | | |
| % bias | -11.534 | -10.886 | -9.436 | -8.413 | -12.808 | -11.154 | -12.888 | -10.303 |
| MCSD[1] | 0.371 | 0.190 | 0.365 | 0.184 | 0.270 | 0.135 | 0.266 | 0.131 |
| ASE[2] | 0.393 | 0.190 | 0.391 | 0.189 | 0.271 | 0.134 | 0.268 | 0.132 |
| ECP[3] | 0.955 | 0.945 | 0.965 | 0.953 | 0.948 | 0.947 | 0.949 | 0.937 |
| iii. AIPW§ | | | | | | | | |
| % bias | 2.293 | 0.074 | 4.757 | 3.211 | -2.274 | 0.852 | -2.242 | 2.058 |
| MCSD[1] | 0.431 | 0.234 | 0.427 | 0.226 | 0.298 | 0.152 | 0.300 | 0.153 |
| ASE[2] | 0.454 | 0.225 | 0.466 | 0.224 | 0.296 | 0.149 | 0.296 | 0.148 |
| ECP[3] | 0.960 | 0.941 | 0.966 | 0.943 | 0.952 | 0.938 | 0.937 | 0.941 |

† Complete care analysis, ‡ multiple imputation, § augmented inverse probability weighted method, [1] Monte Carlo standard deviation, [2] average standard error, [3] empirical coverage probability.

Table 3.2: Monte Carlo simulation with 1,000 replications when ($\xi_4 = -0.5$)

| 30% | $n = 200$ | | | | $n = 400$ | | | |
|---|---|---|---|---|---|---|---|---|
| of missing | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ |
| i. CC[†] | | | | | | | | |
| % bias | -18.165 | -30.041 | -2.499 | -1.649 | -21.094 | -30.982 | -5.913 | -1.879 |
| MCSD[1] | 0.336 | 0.174 | 0.320 | 0.164 | 0.247 | 0.123 | 0.234 | 0.113 |
| ASE[2] | 0.357 | 0.173 | 0.337 | 0.163 | 0.246 | 0.119 | 0.232 | 0.112 |
| ECP[3] | 0.950 | 0.899 | 0.958 | 0.949 | 0.934 | 0.864 | 0.940 | 0.943 |
| ii. MI[‡] | | | | | | | | |
| % bias | -6.846 | -7.710 | -6.099 | -6.623 | -7.978 | -7.565 | -8.581 | -7.284 |
| MCSD[1] | 0.318 | 0.163 | 0.313 | 0.157 | 0.233 | 0.116 | 0.228 | 0.114 |
| ASE[2] | 0.339 | 0.166 | 0.335 | 0.164 | 0.235 | 0.115 | 0.233 | 0.114 |
| ECP[3] | 0.960 | 0.950 | 0.963 | 0.960 | 0.947 | 0.953 | 0.960 | 0.947 |
| iii. AIPW[§] | | | | | | | | |
| % bias | 0.883 | -1.016 | 1.705 | 0.216 | -1.520 | -0.114 | -2.070 | 0.302 |
| MCSD[1] | 0.341 | 0.181 | 0.340 | 0.174 | 0.245 | 0.124 | 0.241 | 0.124 |
| ASE[2] | 0.362 | 0.180 | 0.354 | 0.177 | 0.242 | 0.121 | 0.240 | 0.119 |
| ECP[3] | 0.964 | 0.945 | 0.956 | 0.945 | 0.939 | 0.935 | 0.946 | 0.948 |
| 50% | $n = 200$ | | | | $n = 400$ | | | |
| of missing | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ |
| i. CC[†] | | | | | | | | |
| % bias | -32.616 | -46.683 | 0.387 | 10.913 | -35.139 | -48.949 | -3.645 | 9.304 |
| MCSD[1] | 0.416 | 0.208 | 0.376 | 0.191 | 0.298 | 0.144 | 0.268 | 0.130 |
| ASE[2] | 0.457 | 0.210 | 0.402 | 0.193 | 0.302 | 0.141 | 0.272 | 0.130 |
| ECP[3] | 0.956 | 0.888 | 0.964 | 0.944 | 0.914 | 0.805 | 0.956 | 0.939 |
| ii. MI[‡] | | | | | | | | |
| % bias | -12.990 | -11.383 | -11.257 | -9.198 | -13.237 | -11.231 | -13.187 | -10.474 |
| MCSD[1] | 0.373 | 0.191 | 0.361 | 0.185 | 0.271 | 0.137 | 0.263 | 0.134 |
| ASE[2] | 0.396 | 0.192 | 0.391 | 0.190 | 0.271 | 0.133 | 0.268 | 0.131 |
| ECP[3] | 0.955 | 0.949 | 0.960 | 0.950 | 0.947 | 0.937 | 0.946 | 0.940 |
| iii. AIPW[§] | | | | | | | | |
| % bias | 0.766 | 1.002 | 2.289 | 4.154 | -1.677 | 1.563 | -1.460 | 2.930 |
| MCSD[1] | 0.433 | 0.239 | 0.429 | 0.234 | 0.307 | 0.158 | 0.306 | 0.158 |
| ASE[2] | 0.519 | 0.233 | 0.473 | 0.230 | 0.306 | 0.152 | 0.303 | 0.151 |
| ECP[3] | 0.967 | 0.949 | 0.961 | 0.950 | 0.947 | 0.935 | 0.936 | 0.935 |

[†] Complete care analysis, [‡] multiple imputation, [§] augmented inverse probability weighted method, [1] Monte Carlo standard deviation, [2] average standard error, [3] empirical coverage probability.

Table 3.3: Monte Carlo simulation with 1,000 replications when ($\xi_4 = -0.1$)

| 30% | $n = 200$ | | | | $n = 400$ | | | |
|---|---|---|---|---|---|---|---|---|
| of missing | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ |
| i. CC[†] | | | | | | | | |
| % bias | -15.210 | -22.154 | 2.878 | 9.021 | -18.530 | -23.714 | -1.259 | 8.238 |
| MCSD[1] | 0.334 | 0.173 | 0.323 | 0.164 | 0.247 | 0.120 | 0.237 | 0.113 |
| ASE[2] | 0.357 | 0.173 | 0.345 | 0.166 | 0.245 | 0.118 | 0.237 | 0.114 |
| ECP[3] | 0.955 | 0.925 | 0.966 | 0.943 | 0.932 | 0.898 | 0.947 | 0.943 |
| ii. MI[‡] | | | | | | | | |
| % bias | -6.658 | -6.866 | -5.754 | -5.642 | -8.839 | -7.688 | -9.456 | -7.372 |
| MCSD[1] | 0.323 | 0.166 | 0.317 | 0.160 | 0.235 | 0.115 | 0.231 | 0.113 |
| ASE[2] | 0.343 | 0.167 | 0.339 | 0.166 | 0.237 | 0.117 | 0.234 | 0.116 |
| ECP[3] | 0.962 | 0.955 | 0.967 | 0.959 | 0.948 | 0.953 | 0.950 | 0.947 |
| iii. AIPW[§] | | | | | | | | |
| % bias | 0.883 | -1.016 | 1.705 | 0.216 | -2.053 | 0.059 | -2.574 | 0.463 |
| MCSD[1] | 0.341 | 0.181 | 0.340 | 0.174 | 0.248 | 0.122 | 0.245 | 0.122 |
| ASE[2] | 0.362 | 0.180 | 0.354 | 0.177 | 0.243 | 0.122 | 0.242 | 0.121 |
| ECP[3] | 0.964 | 0.945 | 0.956 | 0.945 | 0.937 | 0.943 | 0.941 | 0.949 |
| 50% | $n = 200$ | | | | $n = 400$ | | | |
| of missing | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ |
| i. CC[†] | | | | | | | | |
| % bias | -29.659 | -40.710 | 7.552 | 21.368 | -32.481 | -43.319 | 2.410 | 20.366 |
| MCSD[1] | 0.404 | 0.203 | 0.389 | 0.194 | 0.293 | 0.143 | 0.284 | 0.132 |
| ASE[2] | 0.442 | 0.207 | 0.422 | 0.196 | 0.297 | 0.139 | 0.281 | 0.132 |
| ECP[3] | 0.958 | 0.904 | 0.970 | 0.937 | 0.923 | 0.833 | 0.945 | 0.920 |
| ii. MI[‡] | | | | | | | | |
| % bias | -11.915 | -11.380 | -9.832 | -8.757 | -13.455 | -11.622 | -13.255 | -10.797 |
| MCSD[1] | 0.375 | 0.192 | 0.368 | 0.186 | 0.273 | 0.137 | 0.269 | 0.134 |
| ASE[2] | 0.397 | 0.192 | 0.394 | 0.191 | 0.273 | 0.135 | 0.270 | 0.133 |
| ECP[3] | 0.952 | 0.947 | 0.962 | 0.952 | 0.949 | 0.945 | 0.949 | 0.935 |
| iii. AIPW[§] | | | | | | | | |
| % bias | 2.345 | 0.521 | 4.691 | 3.885 | -2.800 | 1.058 | -2.602 | 2.344 |
| MCSD[1] | 0.434 | 0.236 | 0.428 | 0.229 | 0.301 | 0.157 | 0.303 | 0.156 |
| ASE[2] | 0.472 | 0.227 | 0.470 | 0.226 | 0.300 | 0.151 | 0.302 | 0.151 |
| ECP[3] | 0.966 | 0.943 | 0.953 | 0.943 | 0.954 | 0.941 | 0.946 | 0.943 |

[†] Complete care analysis, [‡] multiple imputation, [§] augmented inverse probability weighted method, [1] Monte Carlo standard deviation, [2] average standard error, [3] empirical coverage probability.

Table 3.4: Monte Carlo simulation with 1,000 replications when ($\xi_4 = 0.1$)

| 30% | $n = 200$ | | | | $n = 400$ | | | |
|---|---|---|---|---|---|---|---|---|
| of missing | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ |
| i. CC† | | | | | | | | |
| % bias | -12.628 | -16.674 | 5.670 | 13.793 | -15.684 | -18.440 | 1.203 | 12.647 |
| MCSD[1] | 0.335 | 0.173 | 0.327 | 0.165 | 0.244 | 0.118 | 0.238 | 0.113 |
| ASE[2] | 0.350 | 0.171 | 0.343 | 0.165 | 0.241 | 0.117 | 0.235 | 0.113 |
| ECP[3] | 0.953 | 0.924 | 0.961 | 0.939 | 0.932 | 0.921 | 0.948 | 0.935 |
| ii. MI‡ | | | | | | | | |
| % bias | -6.153 | -5.730 | -5.096 | -4.411 | -8.572 | -7.238 | -8.971 | -6.994 |
| MCSD[1] | 0.326 | 0.166 | 0.319 | 0.160 | 0.235 | 0.116 | 0.229 | 0.114 |
| ASE[2] | 0.339 | 0.166 | 0.337 | 0.165 | 0.236 | 0.116 | 0.233 | 0.115 |
| ECP[3] | 0.959 | 0.950 | 0.965 | 0.952 | 0.945 | 0.946 | 0.958 | 0.949 |
| iii. AIPW§ | | | | | | | | |
| % bias | 0.539 | 0.056 | 1.324 | 1.283 | -1.862 | 0.143 | -2.373 | 0.531 |
| MCSD[1] | 0.341 | 0.180 | 0.338 | 0.173 | 0.247 | 0.123 | 0.242 | 0.122 |
| ASE[2] | 0.353 | 0.178 | 0.353 | 0.176 | 0.241 | 0.121 | 0.240 | 0.119 |
| ECP[3] | 0.956 | 0.944 | 0.956 | 0.950 | 0.942 | 0.943 | 0.947 | 0.941 |
| 50% | $n = 200$ | | | | $n = 400$ | | | |
| of missing | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ |
| i. CC† | | | | | | | | |
| % bias | -26.485 | -35.389 | 10.377 | 25.324 | -29.191 | -36.689 | 4.878 | 24.550 |
| MCSD[1] | 0.396 | 0.202 | 0.389 | 0.196 | 0.289 | 0.140 | 0.286 | 0.132 |
| ASE[2] | 0.428 | 0.203 | 0.423 | 0.196 | 0.290 | 0.137 | 0.280 | 0.132 |
| ECP[3] | 0.963 | 0.912 | 0.971 | 0.933 | 0.931 | 0.853 | 0.941 | 0.914 |
| ii. MI‡ | | | | | | | | |
| % bias | -11.110 | -11.276 | -9.079 | -8.630 | -12.792 | -11.016 | -12.923 | -10.242 |
| MCSD[1] | 0.373 | 0.190 | 0.365 | 0.185 | 0.270 | 0.134 | 0.267 | 0.132 |
| ASE[2] | 0.393 | 0.190 | 0.391 | 0.189 | 0.270 | 0.134 | 0.267 | 0.132 |
| ECP[3] | 0.955 | 0.945 | 0.962 | 0.958 | 0.940 | 0.943 | 0.944 | 0.938 |
| iii. AIPW§ | | | | | | | | |
| % bias | 2.526 | -0.374 | 4.668 | 2.802 | -2.595 | 1.113 | -2.628 | 2.342 |
| MCSD[1] | 0.427 | 0.234 | 0.426 | 0.227 | 0.301 | 0.153 | 0.302 | 0.154 |
| ASE[2] | 0.457 | 0.224 | 0.466 | 0.223 | 0.297 | 0.150 | 0.298 | 0.150 |
| ECP[3] | 0.956 | 0.943 | 0.958 | 0.937 | 0.951 | 0.934 | 0.935 | 0.933 |

† Complete care analysis, ‡ multiple imputation, § augmented inverse probability weighted method, [1] Monte Carlo standard deviation, [2] average standard error, [3] empirical coverage probability.

Table 3.5: Monte Carlo simulation with 1,000 replications when ($\xi_4 = 0.5$)

| 30% | $n = 200$ | | | | $n = 400$ | | | |
|---|---|---|---|---|---|---|---|---|
| of missing | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ |
| i. CC[†] | | | | | | | | |
| % bias | -8.424 | -7.855 | 10.944 | 22.273 | -10.788 | -9.978 | 6.606 | 20.756 |
| MCSD[1] | 0.334 | 0.170 | 0.333 | 0.165 | 0.243 | 0.118 | 0.243 | 0.115 |
| ASE[2] | 0.348 | 0.168 | 0.349 | 0.166 | 0.237 | 0.116 | 0.237 | 0.114 |
| ECP[3] | 0.954 | 0.938 | 0.967 | 0.931 | 0.933 | 0.937 | 0.946 | 0.912 |
| ii. MI[‡] | | | | | | | | |
| % bias | -5.830 | -5.538 | -4.734 | -4.006 | -7.316 | -6.868 | -7.745 | -6.489 |
| MCSD[1] | 0.324 | 0.165 | 0.318 | 0.160 | 0.235 | 0.115 | 0.230 | 0.114 |
| ASE[2] | 0.337 | 0.166 | 0.336 | 0.164 | 0.235 | 0.115 | 0.234 | 0.114 |
| ECP[3] | 0.959 | 0.944 | 0.969 | 0.952 | 0.939 | 0.949 | 0.953 | 0.951 |
| iii. AIPW[§] | | | | | | | | |
| % bias | 1.562 | 1.050 | 2.554 | 2.395 | -1.620 | -0.142 | -2.090 | 0.300 |
| MCSD[1] | 0.345 | 0.181 | 0.345 | 0.176 | 0.252 | 0.126 | 0.247 | 0.125 |
| ASE[2] | 0.359 | 0.180 | 0.371 | 0.179 | 0.244 | 0.121 | 0.243 | 0.121 |
| ECP[3] | 0.953 | 0.940 | 0.962 | 0.947 | 0.938 | 0.947 | 0.941 | 0.948 |
| 50% | $n = 200$ | | | | $n = 400$ | | | |
| of missing | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ | $\beta_{11}$ | $\beta_{12}$ | $\beta_{21}$ | $\beta_{22}$ |
| i. CC[†] | | | | | | | | |
| % bias | -21.830 | -26.442 | 15.977 | 33.185 | -24.989 | -27.229 | 10.162 | 33.219 |
| MCSD[1] | 0.387 | 0.197 | 0.391 | 0.197 | 0.285 | 0.137 | 0.288 | 0.134 |
| ASE[2] | 0.417 | 0.201 | 0.442 | 0.199 | 0.282 | 0.136 | 0.287 | 0.134 |
| ECP[3] | 0.959 | 0.915 | 0.972 | 0.923 | 0.930 | 0.899 | 0.945 | 0.886 |
| ii. MI[‡] | | | | | | | | |
| % bias | -10.184 | -10.645 | -8.141 | -7.964 | -11.575 | -10.141 | -11.457 | -9.226 |
| MCSD[1] | 0.373 | 0.190 | 0.362 | 0.186 | 0.273 | 0.133 | 0.267 | 0.131 |
| ASE[2] | 0.391 | 0.190 | 0.391 | 0.189 | 0.270 | 0.133 | 0.268 | 0.132 |
| ECP[3] | 0.953 | 0.946 | 0.965 | 0.946 | 0.940 | 0.937 | 0.944 | 0.935 |
| iii. AIPW[§] | | | | | | | | |
| % bias | 3.026 | -0.118 | 5.326 | 3.423 | -2.882 | 0.722 | -2.672 | 2.091 |
| MCSD[1] | 0.438 | 0.236 | 0.441 | 0.233 | 0.311 | 0.160 | 0.311 | 0.162 |
| ASE[2] | 0.468 | 0.232 | 0.577 | 0.232 | 0.306 | 0.153 | 0.309 | 0.154 |
| ECP[3] | 0.954 | 0.956 | 0.966 | 0.940 | 0.946 | 0.930 | 0.951 | 0.943 |

[†] Complete care analysis, [‡] multiple imputation, [§] augmented inverse probability weighted method, [1] Monte Carlo standard deviation, [2] average standard error, [3] empirical coverage probability.

It is evident that the proposed AIPW estimator is virtually unbiased in all cases. To sum up, this simulation study provided numerical evidence for the double robustness of the proposed AIPW estimator and, also, its asymptotic normality.

The proposed AIPW estimator also outperformed the CC and MI approaches which provided biased estimates. Moreover, the AIPW estimator was more computationally efficient compared to the MI approach which requires performing the analysis multiple times. Under 100 replications with sample size of 400 and 50% missing event type, the computation time of the AIPW estimator is 5.96 minutes in average with 1.09 minutes of standard deviation. However, those of the MI approach is 20.92 minutes in average with 2.28 minutes of standard deviation.

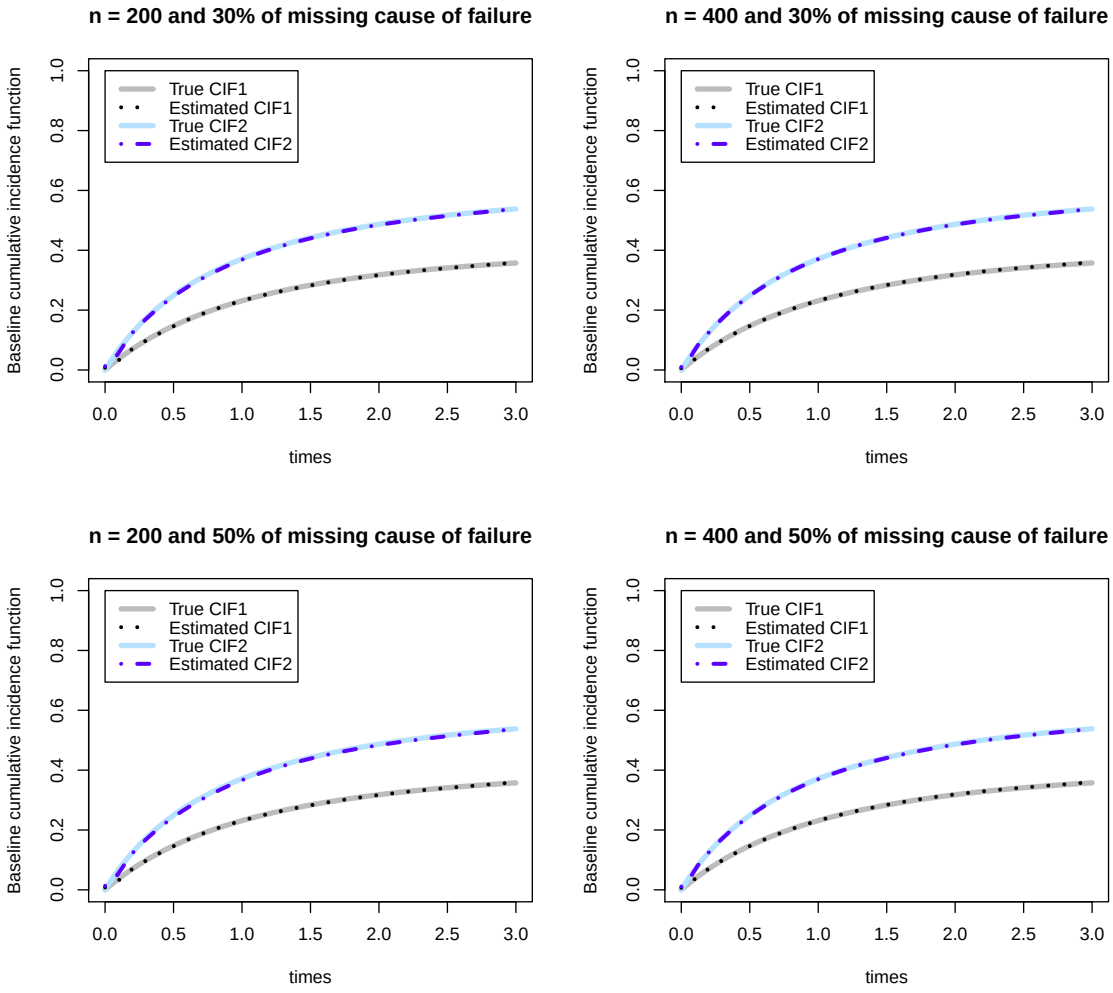



Figure 3.1: The predicted baseline cumulative incidence functions resulted from a simulation study with sample sizes of 200 and 400 when $\xi_4 = 0$

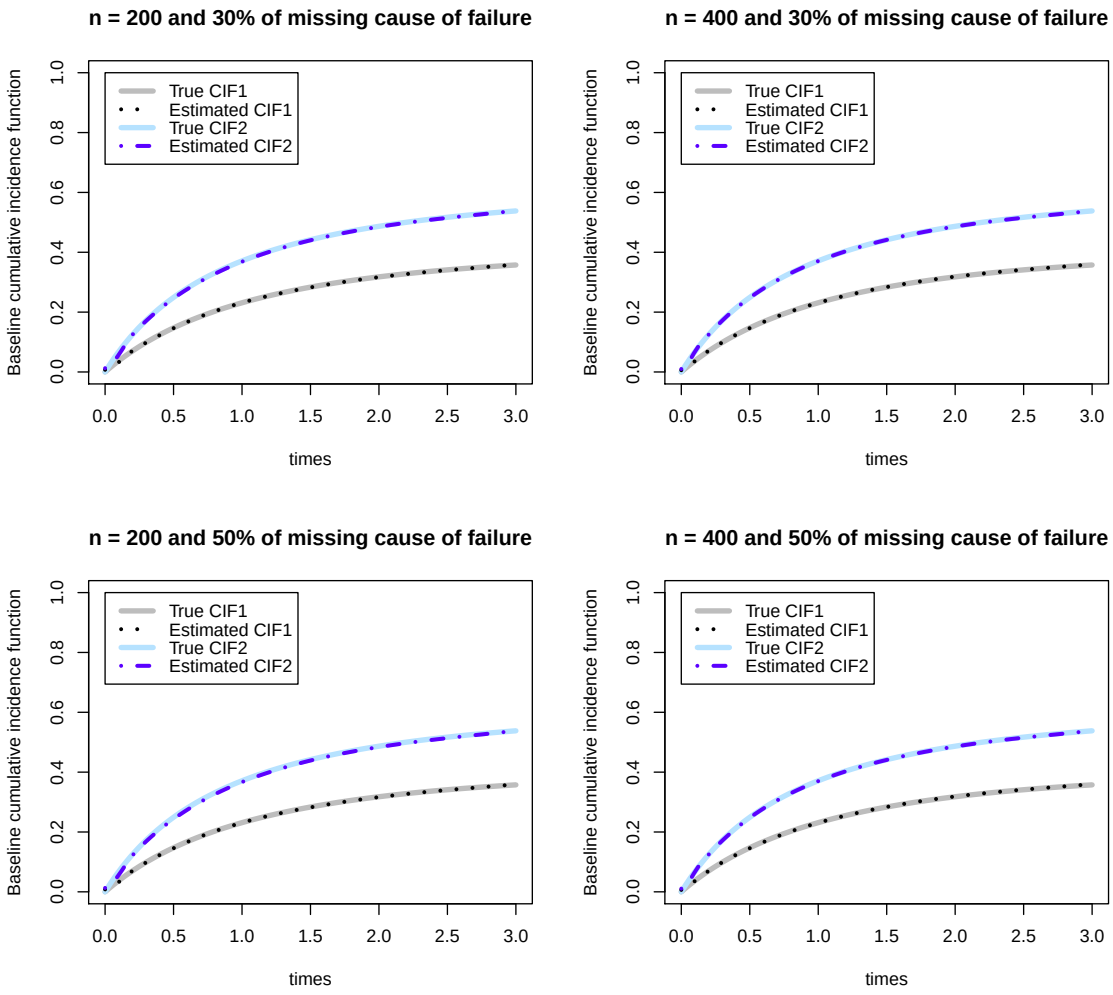



Figure 3.2: The predicted baseline cumulative incidence functions resulted from a simulation study with sample sizes of 200 and 400 when $\xi_4 = -0.5$

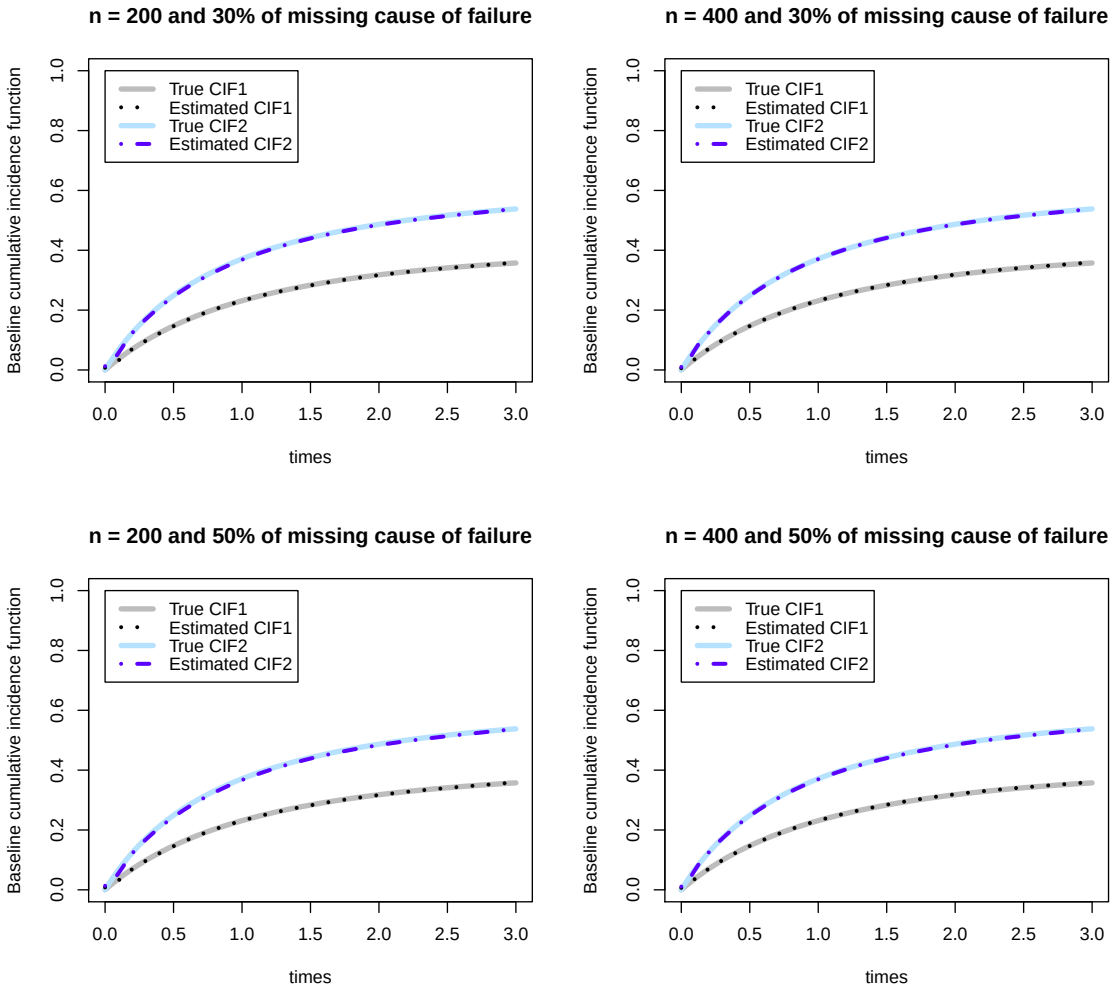



Figure 3.3: The predicted baseline cumulative incidence functions resulted from a simulation study with sample sizes of 200 and 400 when $\xi_4 = -0.1$

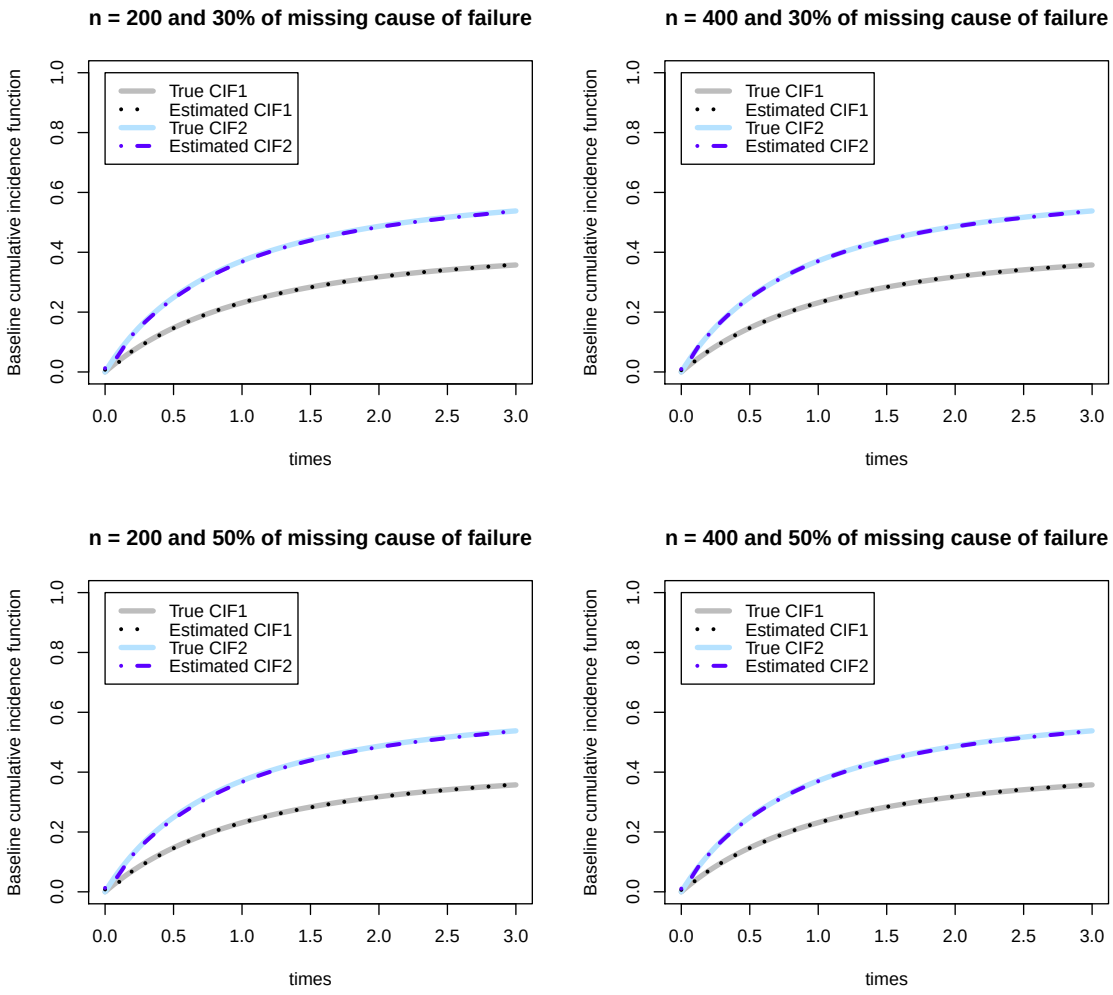



Figure 3.4: The predicted baseline cumulative incidence functions resulted from a simulation study with sample sizes of 200 and 400 when $\xi_4 = 0.1$

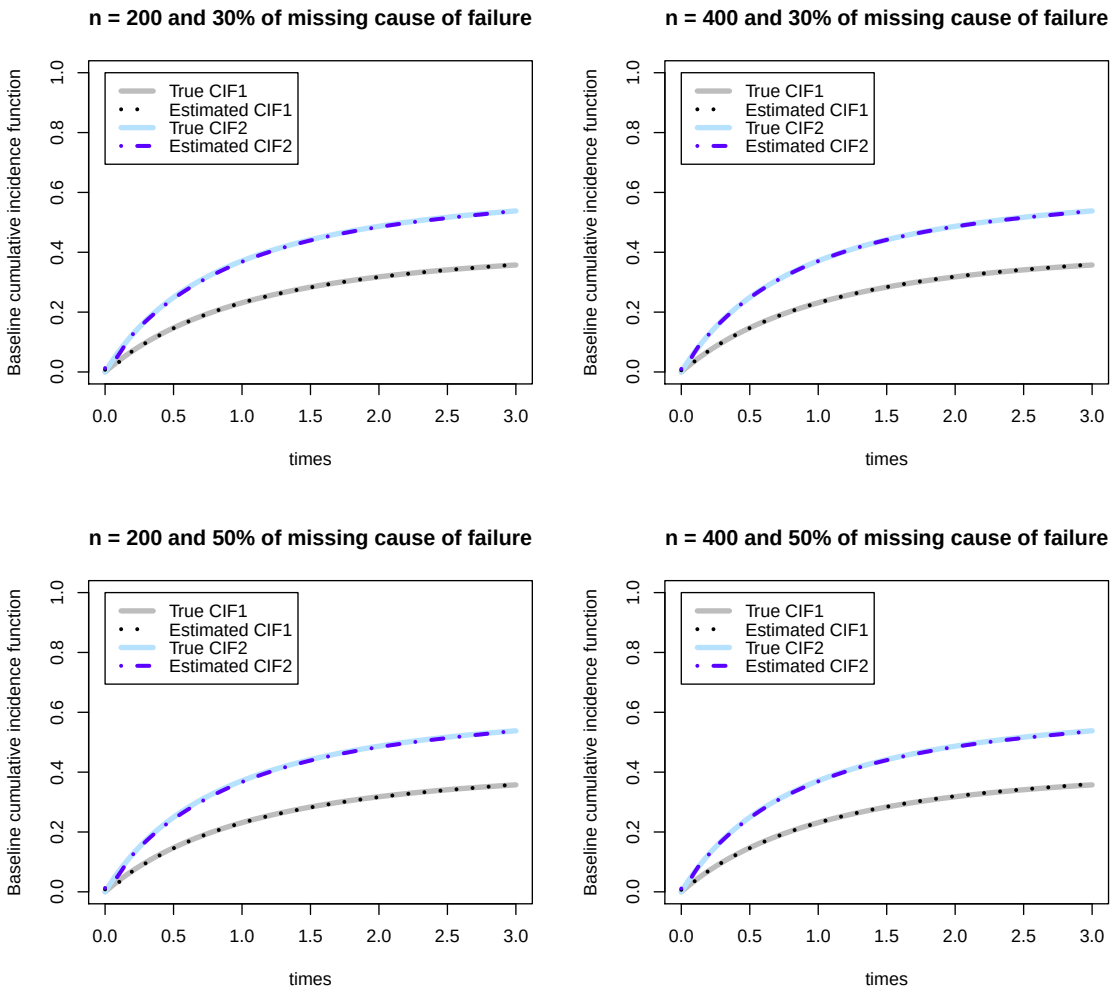



Figure 3.5: The predicted baseline cumulative incidence functions resulted from a simulation study with sample sizes of 200 and 400 when $\xi_4 = 0.5$

## 3.4 Analysis of HIV data

The proposed AIPW approach was used to analyze the data from the motivating EA-IeDEA study described in Chapter 3.1, using the ciregic_aipw function in the R package intccr. The goal of this analysis was to evaluate gender, age and CD4 cell count as potential prognostic factors for death and disengagement from care, as well as to estimate the covariate-specific CIFs of these event types. Descriptive characteristics of the study sample are presented in Table 3.6. The total sample size was 48,691 patients. In total, 2,094 (4.3%) of them were observed to die (reported deaths), while 20,477 (42.1%) patients were identified as lost to clinic. Of them, 4,890 (23.9%) were successfully traced by outreach workers and had their true vital status actively ascertained. This indicates that there is a large portion of missing event types among those who where classified as losses to HIV care. 516 (10.6%) of successfully traced individuals were found to be deceased and this indicates a significant death under-reporting problem.

The data were analyzed using the naïve complete case analysis and the proposed AIPW approach. For convenience of interpretation, the proportional odds models were considered in both event types in these analyses. Standard error estimation in both cases was based on nonparametric bootstrap using 100 bootstrap replications. For the AIPW approach linear binary logistic models were assumed for the probability of non-missingness (i.e. of successful double sampling) and the probability of death. In both models, there exist covariates the time $U$, age, gender, CD4 and the number of outreach workers. Note that the latter covariate is an auxiliary covariate which is not of scientific interest but is expected to be associated

Table 3.6: Descriptive characteristics of the study sample

| | In HIV care ($N = 26,120$) n(%) | Loss to care ($N = 20,477$) n(%) | Death ($N = 2,094$) n(%) |
|---|---|---|---|
| Gender | | | |
| *Female* | 17,511(67.0) | 13,655(66.7) | 1,125(53.7) |
| *Male* | 8,609(33.0) | 6,822(33.3) | 969(46.3) |
| Double sampling | | | |
| *Yes* | 0(-) | 4,890(23.9) | 0(-) |
| *No* | 0(-) | 15,587(76.1) | 0(-) |
| Vital status | | | |
| *Dead* | 0(-) | 516(10.6) | 0(-) |
| *Alive* | 0(-) | 4,374(89.4) | 0(-) |
| | Median(IQR) | Median(IQR) | Median(IQR) |
| Age (years) | 37.8(31.8, 45.5) | 36.3(30.6, 43.3) | 38.0(31.8, 45.2) |
| CD4 (cells/$\mu l$) | 206(95, 338) | 147 (62.4, 264) | 80 (24, 165) |

with the probability of successful outreach/double sampling (i.e. non-missingness). Since the missingness occurs only on the subgroup of patients who were identified as lost (20,477 patients) in this application, the non-missingness and death probability logistic models were fitted using this subset of patients. Results from the complete case and the proposed AIPW analyses are listed in Table 3.7. Calculation of point estimates from the data set of 48,691 observations based on the AIPW approach using the ciregic_aipw function required only about 1.1 minutes. In addition, the computation time that provides both point estimates and the standard error based on 100 bootstrap replications required about 79.9 minutes when parallel computing option is not selected and 42.2 minutes with parallel computing (utilizing three cores) on a quad-core personal computer. Based on the proposed AIPW approach, a higher CD4 count is associated with a higher CIF of disengagement from care. Based on

Table 3.7: Covariate effects on the CIF of disengagement from care and death based on the naïve complete case analysis and the proposed AIPW approach

| Outcome | Covariates | Naïve $\hat{\beta}$ ($p$-value) | Proposed AIPW $\hat{\beta}$ ($p$-value) |
|---|---|---|---|
| Disengagement | Gender *Male versus Female* | 0.184 ($< 0.001$) | $-0.019$ (0.245) |
| | CD4 at ART initiation *per 100 cells/$\mu l$* | 0.010 (0.119) | 0.226 ($< 0.001$) |
| | Age at ART initiation *per 10 years* | $-0.205$ ($< 0.001$) | $-0.065$ (0.049) |
| Death | Gender *Male versus Female* | 0.356 ($< 0.001$) | 0.092 ($< 0.001$) |
| | CD4 at ART initiation *per 100 cells/$\mu l$* | $-0.480$ ($< 0.001$) | $-0.359$ (0.001) |
| | Age at ART initiation *per 10 years* | 0.030 (0.223) | 0.004 ($< 0.001$) |

the naïve analysis male gender and younger age significantly associated with the risk of disengagement, while the effect of CD4 count on disengagement is less pronounced compared to the AIPW approach. The analysis of the CIF of death revealed that old male gender and a lower CD4 cell count are prognostic of death based on the AIPW approach. The effect of male gender and CD4 cell count on mortality was more pronounced in the naïve analysis.

The predicted CIFs for disengagement from care and death for a 30-year-old male patient by CD4 cell count from the naïve complete case analysis and the AIPW approach are depicted in Figures 3.6 and 3.7. The complete case analysis underestimates the predicted CIFs for both disengagement from care and death, compared to the AIPW approach. This is because the complete case analysis selectively discards events only and not right-censored observations.

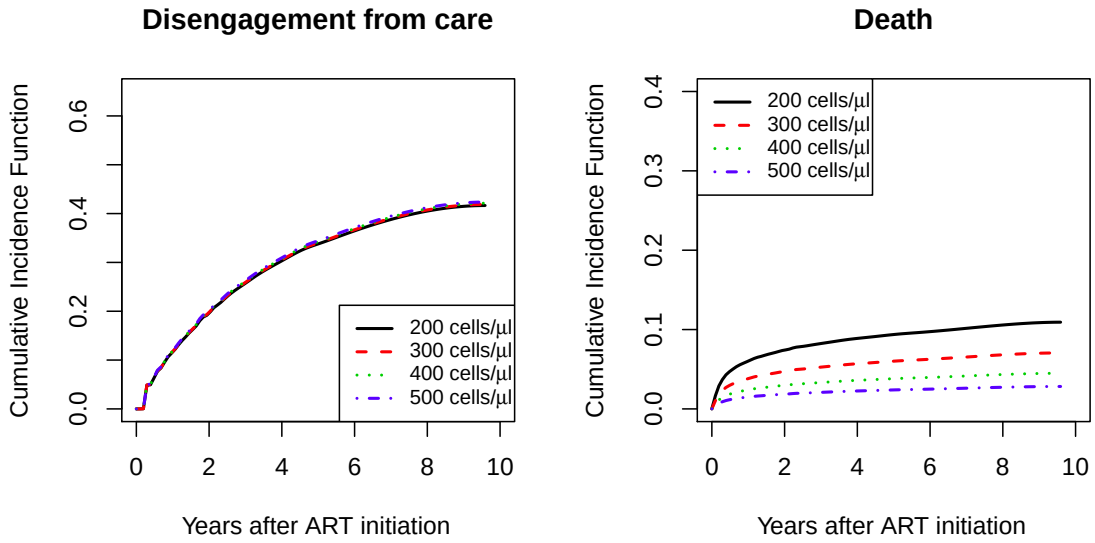



Figure 3.6: The predicted cumulative incidence functions from the naïve analysis for a 30-year-old male patient

## 3.5 Illustration of the R function ciregic_aipw

The proposed AIPW method was implemented in the existing R package intccr (Park et al., 2019). The corresponding function ciregic_aipw for the analysis of interval-censored competing risks data and missing event types is provided in R version 3.5.2 or higher (R Core Team, 2019). Currently, the function allows for only two event types. The package installation and loading can be performed as follows:

```
R> install.packages("intccr")
R> library(intccr)
```

In this illustration, the simulated data set (simdata_aipw), which is available in the intccr package, will be analyzed. This data set consists of 200 observations with 7 variables: id, v, u, c, z1, z2, and a. The description of these variables is provided in Table 3.8. The first 6 observations in the data set (simdata_aipw) are listed below.

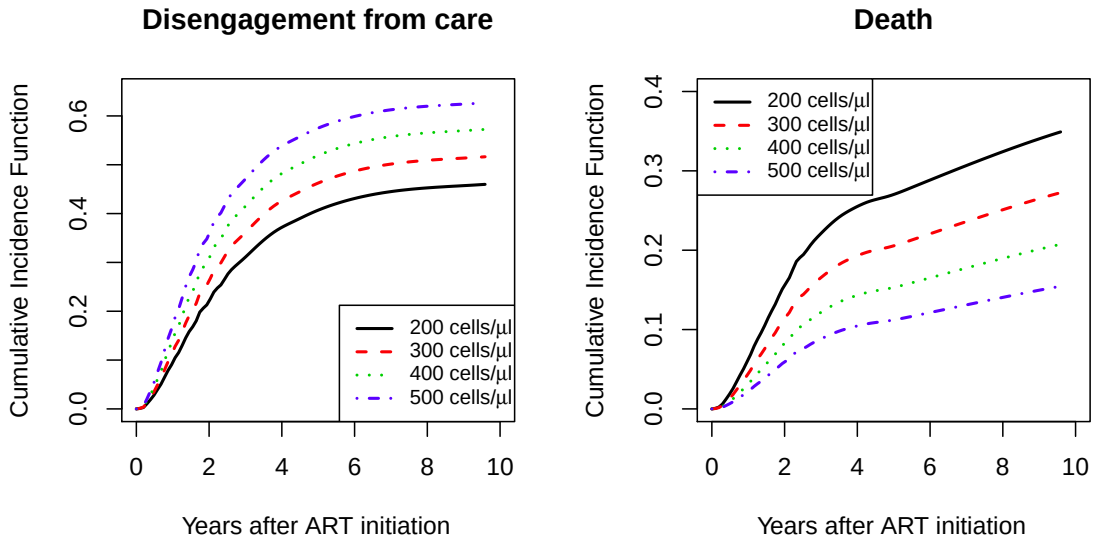



Figure 3.7: The predicted cumulative incidence functions from the proposed AIPW method for a 30-year-old male patient

R> head(simdata_aipw)

| | id | v | u | c | z1 | z2 | a |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 0.0000000 | 0.1779317 | 2 | 1 | 0.2239254 | 0.6279651 |
| 2 | 2 | 1.4760692 | 1.9341271 | NA | 1 | −1.1562233 | 1.0021440 |
| 3 | 3 | 0.5704245 | 1.5265510 | 2 | 1 | 0.4224185 | 0.2843777 |
| 4 | 4 | 1.0087580 | 1.7452873 | NA | 1 | −1.3247553 | −1.0017791 |
| 5 | 5 | 0.1232930 | 0.3463802 | 2 | 0 | 0.1410843 | −0.6172219 |
| 6 | 6 | 2.6582404 | Inf | 0 | 0 | −0.5360480 | 1.8281942 |

The first observation (id = 1) is left-censored and the corresponding event type is c = 2. The second observation (id = 2) is interval-censored. The event occurred in (v, u) = (1.476, 1.934), but the corresponding event type is missing.

R> table(simdata_aipw$c)

| 0 | 1 | 2 |
|---|---|---|

Table 3.8: Variables in the data set simdata_aipw

| Variables | Description |
|---|---|
| id | a unique individual identifier |
| v | last observation time prior to the event |
| u | first observation time after the event |
| c | an event type |
| z1 | a binary covariate |
| z2 | a continuous covariate |
| a | an auxiliary variable |

31 45 39

R> sum(is.na(simdata_aipw$c))

[1] 85

The simdata_aipw has 50.3 % missing event types among 169 observations that are not right-censored. The function ciregic_aipw fits a two parametric models; one is a logistic regression model for the probability of non-missingness using 169 observations and the other is for the probability of event type using 84 observations. Table 3.9 presents the arguments of the function. The argument formula consists of the object function Surv2(v, u, event) and a linear combination of covariates (for the simdata_aipw data set the formula is Surv2(v = v, u = u, event = c) $\sim$ z1 + z2). A set of auxiliary variables is allowed in the argument aux. Multiple auxiliary variables can be put into the argument. For example, users simply type aux = a for a single auxiliary variable or aux = a + b for two auxiliary variables. However,

the default setting of aux is NULL, which means that the models for the probability of missingness and event type do not contain an auxiliary variable. The argument sub defines a subset of the observation that contains missing event type. This is not applicable in most applications. However, in some special cases such as the motivating study, only a subset of the observations with an event type are subject to being missing. The argument alpha is a vector of nonnegative values that govern the link function under the class of odds rate transformation models (for more details see Chapter 3.2.1). Note that the function allows for different models for each event type. The argument k requires a value between 0.5 and 1, with a default value of 1. Based on k, the number of internal knots is defined as largest integer which is less than or equal to $kn^{1/3}$. Using a smaller number k reduces the computation time at the expense of a cruder B-spline approximation in finite samples. The remaining arguments, nboot and do.par, are options to define the number of bootstrap replications and the use of parallel computing. If the nboot = 0 then the ciregic_aipw function returns only point estimates without standard errors and p-values. If nboot > 0, one needs to set a seed number for reproducibility of the bootstrap standard errors is as follows:

```
R> set.seed(2019)
```

Obtaining point estimates for the regression coefficients using this data set requires the code:

```
R> set.seed(2019)
R> fit <- ciregic_aipw(formula = Surv2(v = v, u = u, event = c) ~ z1 + z2,
                       aux = a, data = simdata_aipw, alpha = c(1, 1),
                       nboot = 0, do.par = FALSE)
```

Table 3.9: Argument of the function ciregic_aipw

| Variables | Description |
|---|---|
| formula | a model formula |
| aux | a set of auxiliary variables (optional) |
| data | a data frame |
| sub | a subset of the observation that are subject to being missing (optional argument) |
| alpha | a parameter specifying the link functions |
| k | a parameter that controls the number of internal knots |
| nboot | a number of bootstrap replications for standard error estimation |
| do.par | a logical constant to utilize parallel computing |

```
R> summary(fit)

Call:

ciregic_aipw.default(formula = Surv2(v = v, u = u, event = c) ~ z1 + z2,
                     aux = a, data = simdata_aipw, alpha = c(1, 1),
                     do.par = FALSE, nboot = 0)

Event type 1

 Coefficients :

      z1        z2

0.25067 0.01175

Event type 2

 Coefficients :
```

```
      z1         z2
−0.19678  0.08918
```

Point estimates for the regression coefficients and bootstrap standard errors based on 50 replications without parallel computing are obtained as follows:

```
R> set.seed(2019)
R> fit.npar <− ciregic_aipw(formula = Surv2(v = v, u = u, event = c) ∼ z1 + z2,
                            aux = a, data = simdata_aipw,
                            alpha = c(1, 1), nboot = 50,
                            do.par = FALSE)
  |========================================| 100%
Completed bootstrapping: 50 out of 50
> summary(fit.npar)
Call:
ciregic_aipw.default(formula = Surv2(v = v, u = u, event = c) ∼ z1 + z2,
                     aux = a, data = simdata_aipw,
                     alpha = c(1, 1), do.par = FALSE, nboot = 50)
Event type 1
   Estimate Std. Error z value Pr(>|z|)
z1   0.2507     0.3450   0.727    0.468
z2   0.0118     0.1888   0.062    0.950
Event type 2
   Estimate Std. Error z value Pr(>|z|)
z1  −0.1968     0.3640  −0.541    0.589
```

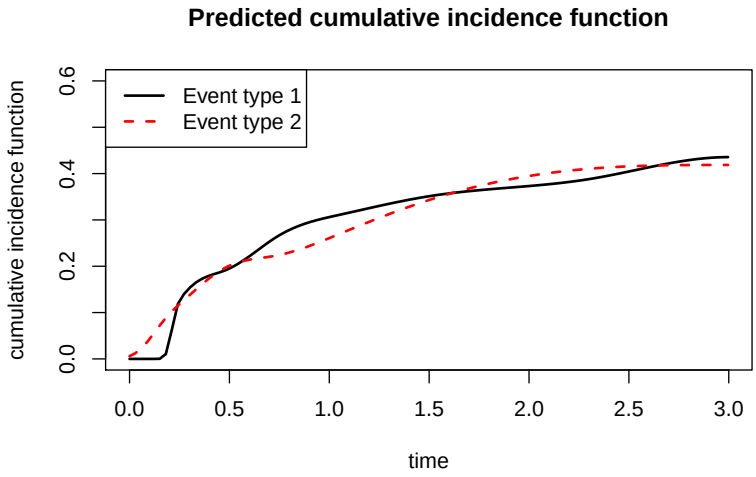



Figure 3.8: The estimated baseline cumulative incidence function

z2 0.0892 0.1902 0.469 0.639

A warning message is automatically generated if there are bootstrap replications that did not converge. The generic function summary provides the summary table for both event types. The output consists of the function call, estimates with its bootstrap standard error, z score, and p-value with significant stars. The significant stars appear when at least one covariate satisfies levels of significance.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The parallel computing option do.par = TRUE selects the maximum number of cores minus one. For example, 3 available cores are assigned in quad core system because the user needs one core to run the operating system. The parallel computing offers faster bootstrap standard error computation, and returns the same result if the same seed number was defined. Moreover, the function predict returns the covariate-specific predicted CIF. The generic function predict provides a corresponding predicted CIF to a sequence of time points and a combination of covariates. The

following R code shows how to draw a plot for the predicted baseline CIFs. The resulting plot is depicted in Figure 3.8, a different value in the argument covp provides the predicted CIFs with for the required covariate pattern (e.g. covp = c(1, .5)).

```
R> t <− seq(from = fit$tms[1], to = fit$tms[2],
            by = diff(fit$tms) / 99)
R> pred <− predict(object = fit, covp = c(0, 0), times = t)
R> plot(pred$t, pred$cif1, type = "l", ylim = c(0, .6), lwd = 2,
     main = ‘‘Predicted cumulative incidence function ’’,
     xlab = ‘‘time ’’, ylab = ‘‘cumulative incidence function ’’)
R> points(pred$t, pred$cif2, type = ‘‘l ’’, col = 2, lty = 2, lwd = 2)
R> legend(‘‘topleft ’’, legend = c(‘‘Event type 1’’, ‘‘Event type 2’’),
       lty = 1:2, col = 1:2, lwd = c(2, 2))
```

## 3.6 Discussion

In this work, complex issue was addressed in semiparametric analysis of the CIF with interval-censored competing risks data, missing event types, and potentially auxiliary covariates under a missing at random (MAR) assumption. The proposed approach utilizes an inverse probability weighting within the semiparametric B-spline-based sieve maximum likelihood estimation framework for interval-censored competing risks data by Bakoyannis et al. (2017). The general class of odds rate transformation models was considered in this approach. The standard error of the regression coefficient can be estimated via the use of nonparametric bootstrap method.

It was shown that the proposed estimator possesses the double robustness property, that is it is consistent even if either the model for the probability of non-missingness or the model for the event type probability is misspecified, but not both. The double robustness property of the proposed estimator was also justified numerically via a series of simulation experiments. On the contrary, the naïve complete case analysis and the multiple imputation approach for missing causes of failure (Bakoyannis et al., 2010) provided biased estimates. The simulation studies provided also numerical evidence for the asymptotic normality of the AIPW regression coefficient estimator. The proposed method is readily applicable using the ciregic_aipw function which has been incorporated in the R package intccr (Park et al., 2019). Importantly, this function supports parallel computing for a considerably faster bootstrap standard error estimation. Chapter 3.5 provides an illustrative example on how to use this function in practice.

The issue of semiparametric analysis of the CIF based on interval-censored competing risks data has not received much attention in the literature. To the best of my knowledge, only Do & Kim (2017) and L. Mao et al. (2017) have considered this problem. Do & Kim (2017) utilized Rubin's multiple imputation to deal with missingness in the framework of the pseudo-value approach for the CIF (Klein & Andersen, 2005). However, Rubin's multiple imputation can provide biased estimates when the imputation model is misspecified. Unlike their method, the proposed AIPW estimator possesses the double robustness property and, thus, it is consistent even if the event type probability model is misspecified, provided that the non-missingness probability model is correct. L. Mao et al. (2017) allowed for missing event types in their EM-algorithm-based approach for interval-censored competing risks data.

This approach, unlike the AIPW method, is not readily applicable using off-the-shelf software and, also, does not explicitly incorporate the nonlinear inequality constraint that the sum of the CIFs for all event types is bounded by one, which can lead to non-convergence problems. Moreover, the proposals by Do & Kim (2017) and L. Mao et al. (2017) cannot be used with auxiliary variables, as discussed in Chapter 3.1, even though such covariates can be crucial for making the key MAR assumption more plausible in practice (Collins et al., 2001; Lu & Tsiatis, 2001; Bakoyannis et al., 2019). In contrast, the proposed AIPW approach can easily incorporate auxiliary variables in the models for the probability of non-missingness and the event type probability. The use of auxiliary variables was illustrated in the human immunodeficiency virus data application.

In conclusion, the proposed AIPW approach of the B-spline-based sieve maximum likelihood method was considered as a robust and flexible analytical method for the analysis of the CIF based on interval-censored competing risks with missing event types. Interval censoring and missing event types are common problems which are typically met in studies based on electronic health records and can lead to bias, as illustrated in the simulation experiments. The availability of the ciregic_aipw function in the R package intccr has the potential to increase the impact of the proposed work in real-life medical research. Currently, the ciregic_aipw function allows for only two event types which is sufficient for many applications. However, the plan is to extend the function to allowing an arbitrary (finite) number of event types.

# Chapter 4

## Competing risks regression analysis on the cumulative incidence function for left-truncated and interval-censored data

Prognostic models for chronic conditions are crucial in modern medicine as they inform medical decision making at the patient level and health policy at the population level. However, an estimation of such models in studies with time-to-event outcomes is typically complicated due to left truncation, interval censoring, and competing risks. Ignoring these issues can lead to biased estimates and, therefore, invalid prognostic models. In this chapter, the issue of semiparametric analysis of the CIF, which explicitly quantifies clinical prognosis, is addressed in left-truncated and interval-censored competing risks data. This approach is considered in the general class of semiparametric odds rate transformation models which includes the proportional odds and the proportional subdistribution hazards model as special cases. The proposed estimation approach is based on semiparametric sieve maximum likelihood and utilizes B-splines. Simulation studies show that the performance of the method under both left truncation and interval censoring is excellent. The proposed methodology is applied to analyze dementia data that are subject to interval censoring, left truncation, and the competing risk of death. The method is readily applicable using the ciregic_lt function in the R package intccr.

## 4.1 Introduction

Prognostic models of chronic conditions are crucial in modern medicine and public health. They can be used to estimate the prognostic effects of various characteristics and to provide prognosis estimates which are important for informing medical decision making at the patient level. For example, prognostic models can be used for classifying patients into risk groups, informing patients about their likely course of disease, and guiding physicians and patients in their decision making about the optimal treatment option (Royston et al., 2009; Moons et al., 2009). Moreover, these models can be used for identifying appropriate patients for recruitment in clinical trials (Moons et al., 2009). Prognostic models are also important for informing health policy at the population level, such as for guiding resource allocation. Finally, prognosis estimates, such as the risk of a particular outcome, are also of key importance in quality of life studies (Fine & Gray, 1999). The estimation of prognostic models for chronic conditions in studies with time-to-event outcomes is typically challenging. First, many individuals die prior to the occurrence of the outcome of interest which leads to a competing risks situation. Second, the outcome onset time is frequently not precisely observed but is only known to lie between two observation times such as clinic visits. This issue is known as interval censoring (Z. Zhang & Sun, 2010). Finally, individuals in many studies are recruited some time after the onset of risk under study, which is known as left truncation or late entry.

This research was motivated by the Indianapolis-Ibadan Dementia Project. The data have been described in Hendrie et al. (1995); Hall et al. (2009); Hendrie et al. (2017). In this study, participants were evaluated for dementia every 2 to 3

years, which means that the exact dementia onset time was unknown and was only known to lie within a 2 to 3-year time interval. In addition to interval censoring, the majority of the participants died throughout the follow-up period prior to developing dementia, leading to a competing risks situation. Finally, the time origin in this study was at the 65 years of age landmark. However, the vast majority of participants were enrolled after the age of 65 (left truncation). Given that participants had to be dementia-free at enrollment, those who were enrolled after the age of 65 constituted a sample of healthier individuals compared to those who were alive and dementia-free just at 65. Ignoring this left truncation issue can lead to significantly biased estimates (Bakoyannis & Touloumi, 2017).

When the scientific interest of the study is focused on prognostic factors or prognostic models, and the study involves competing risks, such as death, the most relevant estimand is the CIF (Koller et al., 2012; Andersen et al., 2012; Bakoyannis et al., 2017). The CIF, which is identifiable based on competing risks data, represents the cumulative probability of the outcome of interest (Putter et al., 2007; Bakoyannis & Touloumi, 2012). The issue of semiparametric analysis of the CIF under the proportional subdistribution hazards model has been addressed by Fine & Gray (1999), which used inverse probability weighting techniques to account for right censoring. Recently, L. Mao & Lin (2017) proposed a semiparametrically efficient estimation approach for semiparametric transformation models for the CIF. However, neither of these proposals are applicable to left-truncated or interval-censored competing risks data. The issue of left truncation within the framework of the proportional subdistribution hazards model has received attention without, however, allowing for interval censoring (X. Zhang et al., 2011; Geskus, 2011; P.-S. Shen, 2011). There is also some

recent work on semiparametric analysis of the CIF with interval-censored competing risks data which, nevertheless, does not also address the issue of left truncation. Li (2016) proposed a B-spline-based sieve maximum likelihood approach for semiparametric analysis of the CIF under the proportional subdistribution hazards model of Fine and Gray. Bakoyannis et al. (2017) extended this work to the general class of odds rate transformation models which includes the proportional subdistribution hazards and the proportional odds models as special cases. They also explicitly incorporated the nonlinear inequality constraint that the sum of the CIF for the different causes is bounded by 1 for any covariate pattern. This is important since ignoring this constraint may lead to non-convergence problems. Importantly, the latter approach is readily implemented using the function ciregic in the R package intccr (Park et al., 2019). L. Mao et al. (2017) proposed a nonparametric maximum likelihood approach and provided an EM algorithm for computation. However, they did not consider the boundedness constraint for the CIF and, also, there is no general software to readily implement this approach in practice. All three aforementioned approaches for interval-censored competing risks data are semiparametrically efficient with respect to the regression coefficients. To the best of my knowledge, there is no methodology for semiparametric analysis of the CIF with both left-truncated and interval-censored competing risks data.

In this work, the issue of semiparametric analysis of the CIF is addressed in left-truncated and interval-censored competing risks data. This extends the B-spline-based sieve maximum likelihood methodology for the CIF under the general class of semiparametric odds rate transformation models by Bakoyannis et al. (2017), to allow for independent left truncation conditionally on the covariates. More precisely, the

sieve likelihood function is modified by dividing each likelihood contribution with the probability of being event-free at the left truncation time. The justification for this modification is that the likelihood contribution of a left-truncated individual must be conditional on the fact that this individual is event-free at the left truncation time. Simulation studies show that the method works well and that the method by Bakoyannis et al. (2017) which ignores left truncation can lead to substantial bias with left-truncated and interval-censored competing risks data. The proposed approach is readily applicable using the new ciregic_lt function of the R package intccr. A short tutorial describes how to use this function in Chapter 4.5. Finally, potential prognostic factors for dementia and death are evaluated in the motivating Indianapolis-Ibadan Dementia Project.

The structure of this Chapter is as follows. In Chapter 4.2, it begins by introducing some notations, the class of generalized odds rate transformation models for the CIF, and the likelihood function for left-truncated and interval-censored competing risks data. The B-spline based sieve maximum likelihood estimation approach is also presented. In Chapter 4.3, a series of simulation experiments is conducted to evaluate the finite-sample performance of the proposed method and the method by Bakoyannis et al. (2017) which ignores left truncation. In Chapter 4.4, the proposed methodology is applied to the motivating study, the Indianapolis-Ibadan Dementia Project, to fit a simple prognostic model for dementia and death. Chapter 4.5 includes a short tutorial about the use of the R function ciregic_lt which is contained in the package intccr. Finally, concluding remarks are discussed in Chapter 4.6.

## 4.2 Methods

### 4.2.1 Data and models

Suppose that a random sample of $n$ individuals is observed over the observation time interval $[a, b]$, with $0 < a < b < \infty$. Let $T_i$ denote the time to event and $C_i$ the event type, with $C_i \in \{1, 2, \ldots, k\}$. The total number of event types $k$ is considered to be finite. With interval-censored data $T_i$ is not precisely observed but is only known to lie between two examination times. Let $(E_{i1}, \ldots, E_{im_i})$ be the set of examination times for the $i$th individual, with $a \leq E_{i1} < E_{i2} < \cdots < E_{im_i} \leq b$ almost surely and $m_i < \infty$ for all $i = 1, \ldots, n$. Next, let $V_i$ denote the last examination time where the $i$th individual was event-free and $U_i$ denote the first examination time where this individual was diagnosed with an event, with $P(V_i < U_i) = 1$ for all $i$. Under this setup one only observes the interval $(V_i, U_i]$ for which $T_i \in (V_i, U_i]$. Let $W_i$ be the potential left truncation time for the $i$th individual which satisfies $P(W_i \leq E_{i1}) = 1$. Now define $\Delta_i^{(1)} = I(V_i < T_i \leq U_i)$, which is the indicator that the $i$th individual is interval-censored and $\Delta_i^{(2)} = I(T_i \leq U_i)$, with $U_i = E_{i1}$, which indicates that the $i$th individual is left-censored. The corresponding event-type specific indicators are defined as $\Delta_{ij}^{(1)} = \Delta_i^{(1)} I(C_i = j)$ and $\Delta_{ij}^{(2)} = \Delta_i^{(2)} I(C_i = j)$. Based on these variables, the indicator of any event is defined as $\Delta_i = \Delta_i^{(1)} + \Delta_i^{(2)} = \sum_{j=1}^{k} \left( \Delta_{ij}^{(1)} + \Delta_{ij}^{(2)} \right)$. Also, the left truncation indicator is defined as $\Xi_i = I(W_i > 0)$. Finally, let $Z_i$ denote the covariate vector of interest for the $i$th individual, with $Z_i \in \mathbb{R}^q$, $q < \infty$. Letting $\underline{\Delta}_i^{(1)} = \left( \Delta_{i1}^{(1)}, \ldots, \Delta_{ik}^{(1)} \right)$ and $\underline{\Delta}_i^{(2)} = \left( \Delta_{i1}^{(2)}, \ldots, \Delta_{ik}^{(2)} \right)$, the observed data are $n$ i.i.d.

copies of $X_i = \left(\underline{\Delta}_i^{(1)}, \underline{\Delta}_i^{(2)}, \Xi_i, W_i, V_i, U_i, Z_i\right)$. Similarly to other works for the CIF under left truncation or interval censoring, it is assume that

A1 Both $W_i$ and $(E_{i1}, \ldots, E_{im_i})$ are independent of $(T_i, C_i)$ conditionally on $Z_i$.

A2 Both $W_i$ and $(E_{i1}, \ldots, E_{im_i})$ are non-informative about the parameters of the conditional distribution $(T_i, C_i)$ given $Z_i$.

(Geskus, 2011; X. Zhang et al., 2011; Li, 2016; Bakoyannis et al., 2017; L. Mao & Lin, 2017).

The CIF for the $j$th event type conditionally on the covariate $Z$ is defined as

$$F_j(t; Z) = \Pr(T \leq t, C = j|Z), \quad j = 1, \ldots, k,$$

and represents the cumulative probability of the $j$th event type in the presence of the other event types (Putter et al., 2007; Bakoyannis & Touloumi, 2012). A general class of models for the CIF is the class of semiparametric transformation models, defined as

$$g_j\left\{F_j(t; Z)\right\} = \phi_j(t) + \beta_j^T Z, \quad j = 1, \ldots, k, \tag{4.1}$$

where $g_j(\cdot)$ is a known increasing link function, $\phi_j(\cdot)$ is an unspecified but strictly increasing function of time, and $\beta_j$ is a vector of regression coefficients that corresponds to the vector of covariates $Z$. A special subset of this class is the class of generalized

odds rate transformation models, which is defined as

$$g_j(F_j;\alpha_j) = \begin{cases} \log\left\{\dfrac{(1-F_j)^{-\alpha_j}-1}{\alpha_j}\right\} & \text{if } 0<\alpha_j<\infty \\ \\ \log\left\{\log(1-F_j)\right\} & \text{if } \alpha_j = 0 \end{cases} \tag{4.2}$$

(Jeong & Fine, 2006; Dabrowska & Doksum, 1988; Scharfstein et al., 1998; Bakoyannis et al., 2017). This class includes the proportional odds model, with $\alpha_j = 1$, and the proportional subdistribution hazards model, with $\alpha_j = 0$, as special cases. Since the link function parameter $\alpha_j$ depends on $j$, it is possible to specify different models for the CIFs of different event types.

### 4.2.2 Semiparametric sieve estimation

In the absence of left truncation, and under the assumption of independent and non-informative observation times conditionally on the covariates $Z_i$ (assumptions A1 and A2 for the examination times only), the likelihood function for interval-censored competing risks data is

$$\tilde{L}_n(\theta;X) \propto \prod_{i=1}^{n}\left[\left[\prod_{j=1}^{k}\left\{F_j(U_i;Z_i,\theta_j)-F_j(V_i;Z_i,\theta_j)\right\}^{\Delta_{ij}^{(1)}}\right]\left[\prod_{j=1}^{k}\left\{F_j(U_i;Z_i,\theta_j)\right\}^{\Delta_{ij}^{(2)}}\right] \times \left\{1-\sum_{j=1}^{k}F_j(V_i;Z_i,\theta_j)\right\}^{1-\Delta_i}\right]$$

where the CIFs are parameterized as semiparametric transformation models (4.1), with link functions specified as in (4.2) (Bakoyannis et al., 2017). Under this model specification, $\theta = \left(\phi, \beta^T\right)^T$, with $\phi = (\phi_1, \ldots, \phi_k)^T$ and $\beta = \left(\beta_1^T, \ldots, \beta_k^T\right)^T$. Under left truncation some individuals are recruited after the onset of risk under study and, therefore, the likelihood contribution for an interval-censored observation with the $j$th event type is

$$\Pr(V_i < T_i \leq U_i, C = j | T_i > W_i, Z_i) = \frac{F_j(U_i; Z_i) - F_j(V_i; Z_i)}{1 - \sum_{j=1}^{k} F_j(W_i; Z_i)}.$$

The corresponding likelihood contribution for a left-truncated and left-censored individual with the $j$th event type is

$$\Pr(T_i \leq U_i, C = j | T_i > W_i, Z_i) = \frac{F_j(U_i; Z_i)}{1 - \sum_{j=1}^{k} F_j(W_i; Z_i)}.$$

Finally, the likelihood contribution for a left-truncated and right-censored individual is

$$\Pr(T_i > V_i | T_i > W_i, Z_i) = \frac{1 - \sum_{j=1}^{k} F_j(V_i; Z_i)}{1 - \sum_{j=1}^{k} F_j(W_i; Z_i)}.$$

Note that in all these cases, the likelihood contribution of a left-truncated observation is that of a non-left-truncated observation divided by the probability of being event-free at the left truncation time $W_i$. Therefore, the likelihood function for

left-truncated and interval-censored competing risks data is

$$
\begin{aligned}
L_n(\theta; X) \propto \prod_{i=1}^{n} & \left\{ 1 - \sum_{j=1}^{k} F_j(W_i; Z_i, \theta_j) \right\}^{-\Xi_i} \\
& \times \left[ \prod_{j=1}^{k} \{ F_j(U_i; Z_i, \theta_j) - F_j(V_i; Z_i, \theta_j) \}^{\Delta_{ij}^{(1)}} \right. \\
& \quad \times \prod_{j=1}^{k} \{ F_j(U_i; Z_i, \theta_j) \}^{\Delta_{ij}^{(2)}} \\
& \quad \left. \times \left\{ 1 - \sum_{j=1}^{k} F_j(V_i; Z_i, \theta_j) \right\}^{1-\Delta_i} \right].
\end{aligned}
$$

The resulting log-likelihood function is

$$
\begin{aligned}
l_n(\theta) = \sum_{i=1}^{n} & \left[ \sum_{j=1}^{k} \Delta_{ij}^{(1)} \log \{ F_j(U_i; Z_i, \theta_j) - F_j(V_i; Z_i, \theta_j) \} \right. \\
& \quad + \sum_{j=1}^{k} \Delta_{ij}^{(2)} \log \{ F_j(U_i; Z_i, \theta_j) \} \\
& \quad + (1 - \Delta_i) \log \left\{ 1 - \sum_{j=1}^{k} F_j(V_i; Z_i, \theta_j) \right\} \\
& \quad \left. - \Xi_i \log \left\{ 1 - \sum_{j=1}^{k} F_j(W_i; Z_i, \theta_j) \right\} \right]
\end{aligned}
\tag{4.3}
$$

The corresponding parameter space is $\Theta = \Phi^k \times \mathcal{B}^k$, where $\Phi \ni \phi_j$, $j = 1, \ldots, k$, is an infinite dimensional parameter space and $\mathcal{B} \ni \beta_j$, $j = 1, \ldots, k$, is a (finite dimensional) subset of $\mathbb{R}^q$. Direct maximization of the log-likelihood (4.3) over the infinite dimensional parameter space $\Theta$ may lead to inconsistency and can be computationally burdensome (X. Shen & Wong, 1994; Y. Zhang et al., 2010). A solution to this problem is to consider a sequence of smaller parameter spaces $\{\Theta_n\}_{n \geq 1}$ that approx-

imations $\Theta$ and the approximation error tends to zero as $n \to \infty$ (X. Shen & Wong, 1994). The sequence of approximating parameter spaces $\{\Theta_n\}_{n\geq 1}$ is known as a sieve. Maximization of the log-likelihood (4.3) over $\Theta_n$ gives the sieve maximum likelihood estimate of $\theta_0$. In this work, sieve spaces of monotone B-spline functions are defined as

$$\Phi_n(\gamma, N_n, m) = \Bigg\{ \phi : \phi(t;\gamma) = \sum_{s=1}^{N_n+m} \gamma_s B_{s,m}(t), \quad \gamma \in \mathbb{R}^{N_n+m}, \gamma_1 < \cdots < \gamma_{N_n+m} \Bigg\}, \tag{4.4}$$

where $N_n$ and $m$ are the number of internal knots and the order of the B-spline, respectively, $\gamma = (\gamma_1, \ldots, \gamma_{N_n+m})^T$ is the vector of the unknown control points for the B-spline, which are the parameters to be estimated, and $t \in [a, b]$ (Y. Zhang et al., 2010; Li, 2016; Bakoyannis et al., 2017). The number of internal knots $N$ increases with the same size $n$ and satisfies $N_n \approx n^\nu$, such that $\max_{1\leq l\leq N_n+1} |s_l - s_{l-1}| = O(n^{-\nu})$, where $s_l$ is the place of the $l$th knot. The choice of $\nu$ that leads to the optimal convergence rate is $1/(1+2p)$, where $p$ is related to the (common) smoothness of the true functions $\phi_j$, $j = 1, \ldots, k$ (Bakoyannis et al., 2017). Maximization of the log-likelihood (4.3) over $\Theta_n = \Phi_n^k(\gamma, N_n, m) \times \mathcal{B}^k$ gives the proposed B-spline-based sieve maximum likelihood estimate. In order to take into account the special features of the CIF, maximization is performed under the following constraints:

(C1) Monotonicity constraints for the coefficients of the B-spline functions in (4.4)

$$\gamma_1 < \cdots < \gamma_{N_n+m}$$

which are required due to the monotonicity of the CIF.

(C2) The nonlinear inequality constraint

$$\max_{z} \left\{ \sum_{j=1}^{k} F_j(b; z, \theta_j) \right\} < 1$$

which is required due to the boundedness of the sum of the CIFs, by virtue of being a probability, by 1.

(C3) The nonlinear equality constraint

$$\max_{z} \left\{ \sum_{j=1}^{k} F_j \left( \min_{i \in \{1,\ldots,n\}} (w_i); z, \theta_j \right) \right\} = 0$$

which is required since the CIFs at 0 are equal to 0.

Standard error estimation can be performed using either the nonparametric bootstrap method or via the least squares approach by J. Huang et al. (2008); Y. Zhang et al. (2010).

### 4.2.3 Practical implementation of the method

Estimation of the parameters of model (4.1) with left-truncated and interval-censored competing risks data requires first to define the number and the placement of the internal knots for the B-splines. In practice, following Bakoyannis et al. (2017), the number of the internal knots can be set to $N_n = \lfloor n^{1/3} \rfloor$, where $\lfloor x \rfloor$ is the largest integer that is less than or equal to $x$ (Bakoyannis et al., 2017). Due to the mono-

tonicity constraints, the proposed approach is less sensitive to the placement of the internal knots compared to unconstrained B-splines. A natural choice is to place the internal knots in the quantiles of the empirical distribution of the observation times (Bakoyannis et al., 2017). Next, one needs to select the link function parameter $\alpha$. Since the true link functions are unknown in practice, one can perform a grid search over a plausible range of combinations $(\alpha_1, \alpha_2)$ to select the optimal link function parameters (Bakoyannis et al., 2017). However, accounting for the additional uncertainty due to this type of model selection is an open problem in the literature (Zeng et al., 2006; M. Mao & Wang, 2010). In many applications, practitioners do not perform model selection for $\alpha$ and adopt the proportional odds model or the proportional subdistribution hazards model since they provide more intuitive interpretations of the regression coefficients. The last step of the estimation process is to obtain parameter estimates by maximizing the sieve log-likelihood function (4.3) under the constraints C1–C3. This requires numerical optimization software that allows for equality, linear inequality and nonlinear inequality constraints.

The proposed methodology can be readily implemented using the new function ciregic_lt which is contained in the R package intccr. The package intccr can be downloaded from the CRAN website or can be directly installed via R studio. This package also includes the function ciregic which performs semiparametric regression with interval-censored competing risks data but without left truncation (Bakoyannis et al., 2017; Park et al., 2019). Currently, the package supports only two event types. The function ciregic_lt uses cubic B-splines and sets, by default, the number of internal knots to be $N = \lfloor n^{1/3} \rfloor$, where $\lfloor x \rfloor$ is the largest integer that is less than or equal to $x$. The knots are automatically placed at the quantiles of the empirical distribution

of the observation times. The user needs to specify the link function parameter value $\alpha = (\alpha_1, \alpha_2)^T$. For standard error estimation, the user can select either the least squares approach by J. Huang et al. (2008) or the nonparametric bootstrap method (G. Cheng et al., 2010). The function also provides an option for parallel computing functionality in order to speed up computations during the bootstrap standard error estimation. It has to be mentioned that the function ciregic_lt utilizes the R package alabama to perform the constrained optimization required for the computation of the proposed B-spline-based sieve maximum likelihood estimation. A short tutorial illustrating the use of the function ciregic_lt is provided in Chapter 4.5.

## 4.3 Simulation studies

In order to evaluate the performance of the proposed method, a series of simulation experiments was conducted. It was considered that there are two event types and two covariates, $Z_1$ and $Z_2$. $Z_1$ was a binary variable drawn from the Bernoulli distribution with a probability of 0.4 and $Z_2$ was a continuous variable drawn from the standard normal distribution. The CIFs of the two event types were assumed to satisfy the proportional odds models

$$F_j(t; Z) = \frac{\exp\left\{\phi_j(t) + \beta_j^T Z\right\}}{1 + \exp\left\{\phi_j(t) + \beta_j^T Z\right\}}, \quad j = 1, 2.$$

The improper Gompertz distribution was used to parameterize $\exp\{\phi_j(t)\}$ as

$$\exp\{\phi_j(t)\} = -\frac{\tau_j}{\rho_j}\{1 - \exp(\rho_j t)\}$$

(Jeong & Fine, 2006). The corresponding parameters were set to be $(\tau_1, \rho_1) = (0.4, -0.48)$ and $(\tau_2, \rho_2) = (0.6, -0.32)$ in order to satisfy

$$\lim_{t\to\infty} \sum_{j=1}^{2} F_j(t; Z = 0) = 1.$$

The true regression coefficients were set to be $\beta = (0.5, -0.3, -0.5, 0.3)^T$. Under these models, the event types were generated first and then the corresponding event times were simulated. The total follow-up period was set to be 5 years. To introduce interval censoring, a series of observation time points was simulated from the exponential distribution with a hazard parameter of 3. Left truncation times were simulated from the exponential distribution with a hazard parameter of 2. If the simulated left truncation time was greater than or equal to the simulated event time, the corresponding observation was discarded to generate a left truncation situation. The simulation of individuals continued until the total number of observations in the data set was equal to the required sample size under each scenario. There were two scenarios regarding left truncation; in the first scenario only 50% of the observations were subject to left truncation, while in the second all the observations were subject to left truncation (i.e. $W_i > 0$ for all $i = 1, \ldots, n$). Under this simulation set up the sample sizes 250, 500, and 1000 were considered, and 1,000 Monte Carlo simulations for each simulation scenario were conducted. In this simulation study, the performance was evaluated by comparing the proposed approach with interval-censored competing risks data by Bakoyannis et al. (2017) which ignores left truncation. The analysis of the simulated data sets was conducted using the functions ciregic, for the

naïve method, and ciregic_lt, for the proposed approach. Standard error estimation was based on the least squares approach by J. Huang et al. (2008).

Simulation results regarding the regression coefficients are presented in Tables 4.1 and 4.2. Under a 50% left truncation (Table 4.1), the proposed approach provided virtually unbiased estimates, the Monte Carlo standard deviation (MCSD) of the estimates was close to average of the standard error estimates (ASE), and the empirical coverage probability (ECP) was close to the nominal level in all cases. On the contrary, the naïve method which ignores left truncation provided regression coefficients exhibiting some bias. The degree of bias did not change with sample size. When all the individuals were subject to left truncation (Table 4.2), the proposed approach still provided almost unbiased regression coefficient and standard error estimates, and the corresponding ECP was close to the nominal level. The naïve approach provided regression coefficients with substantial bias in all cases. Also, the corresponding 95% confidence intervals exhibited a poor coverage probability for cases with larger sample sizes (i.e. $n = 500$ and $n = 1,000$).

Simulation results regarding the baseline CIFs are depicted in Figures 4.1 and 4.2. The left side of these figures displays the average of the baseline CIF estimates based on the proposed method and the right side shows the corresponding estimates for the naïve approach which ignores left truncation. In both left truncation scenarios, the proposed estimator provided practically unbiased estimates of the baseline CIFs. On the contrary, the naïve approach underestimated the baseline CIFs in both cases. This underestimation is a result of the fact that the left truncated data involve healthier individuals. The degree of underestimation was more pronounced for the case where 100% of the observations were subject to left truncation.

Table 4.1: Simulation results of comparison of the proposed method with naïve method for 50% left truncation

| | | Proposed method | | | | Naïve method | | | |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | coefficients | %bias | MCSD[1] | ASE[2] | ECP[3] | %bias | MCSD[1] | ASE[2] | ECP[3] |
| 250 | $\beta_{11}$ | -2.206 | 0.265 | 0.272 | 0.954 | -9.909 | 0.246 | 0.249 | 0.951 |
| | $\beta_{12}$ | 1.466 | 0.126 | 0.136 | 0.966 | -7.624 | 0.116 | 0.126 | 0.966 |
| | $\beta_{21}$ | -1.220 | 0.264 | 0.265 | 0.953 | -9.560 | 0.245 | 0.252 | 0.956 |
| | $\beta_{22}$ | 1.435 | 0.126 | 0.135 | 0.957 | -7.689 | 0.115 | 0.126 | 0.958 |
| 500 | $\beta_{11}$ | -0.537 | 0.188 | 0.187 | 0.946 | -9.088 | 0.173 | 0.171 | 0.937 |
| | $\beta_{12}$ | 0.483 | 0.087 | 0.092 | 0.961 | -8.373 | 0.080 | 0.086 | 0.951 |
| | $\beta_{21}$ | 0.362 | 0.184 | 0.183 | 0.946 | -8.358 | 0.169 | 0.173 | 0.946 |
| | $\beta_{22}$ | 0.746 | 0.086 | 0.092 | 0.965 | -8.075 | 0.079 | 0.086 | 0.958 |
| 1000 | $\beta_{11}$ | 0.038 | 0.130 | 0.130 | 0.944 | -8.674 | 0.120 | 0.119 | 0.928 |
| | $\beta_{12}$ | 0.254 | 0.062 | 0.064 | 0.962 | -8.333 | 0.057 | 0.059 | 0.937 |
| | $\beta_{21}$ | 0.560 | 0.130 | 0.127 | 0.940 | -8.064 | 0.121 | 0.121 | 0.930 |
| | $\beta_{22}$ | 0.035 | 0.062 | 0.064 | 0.960 | -8.550 | 0.058 | 0.059 | 0.929 |

Standard error estimated by the least-square method; [1]Monte Carlo standard deviation; [2]average standard error of the estimates; [3]empirical coverage probability

Table 4.2: Simulation results of comparison of the proposed method with naiïve method for 100% left truncation

| | | Proposed method | | | | Naïve method | | | |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | coefficients | %bias | MCSD[1] | ASE[2] | ECP[3] | %bias | MCSD[1] | ASE[2] | ECP[3] |
| 250 | $\beta_{11}$ | -1.577 | 0.298 | 0.309 | 0.958 | -22.228 | 0.240 | 0.250 | 0.937 |
| | $\beta_{12}$ | 0.399 | 0.140 | 0.155 | 0.970 | -22.175 | 0.112 | 0.126 | 0.942 |
| | $\beta_{21}$ | -1.466 | 0.305 | 0.300 | 0.947 | -23.169 | 0.247 | 0.252 | 0.931 |
| | $\beta_{22}$ | 1.024 | 0.139 | 0.154 | 0.973 | -21.321 | 0.110 | 0.126 | 0.947 |
| 500 | $\beta_{11}$ | -3.247 | 0.212 | 0.214 | 0.951 | -23.618 | 0.172 | 0.172 | 0.890 |
| | $\beta_{12}$ | -1.355 | 0.101 | 0.106 | 0.962 | -22.968 | 0.080 | 0.086 | 0.880 |
| | $\beta_{21}$ | -2.889 | 0.212 | 0.208 | 0.941 | -23.868 | 0.173 | 0.174 | 0.888 |
| | $\beta_{22}$ | -1.144 | 0.102 | 0.106 | 0.958 | -22.769 | 0.081 | 0.086 | 0.897 |
| 1000 | $\beta_{11}$ | -2.325 | 0.145 | 0.149 | 0.957 | -23.228 | 0.115 | 0.120 | 0.846 |
| | $\beta_{12}$ | -0.409 | 0.074 | 0.074 | 0.962 | -22.321 | 0.057 | 0.060 | 0.804 |
| | $\beta_{21}$ | -2.363 | 0.149 | 0.145 | 0.943 | -23.588 | 0.121 | 0.121 | 0.844 |
| | $\beta_{22}$ | -0.463 | 0.076 | 0.074 | 0.952 | -22.320 | 0.059 | 0.060 | 0.796 |

Standard error estimated by the least-square method; [1]Monte Carlo standard deviation; [2]average standard error of the estimates; [3]empirical coverage probability

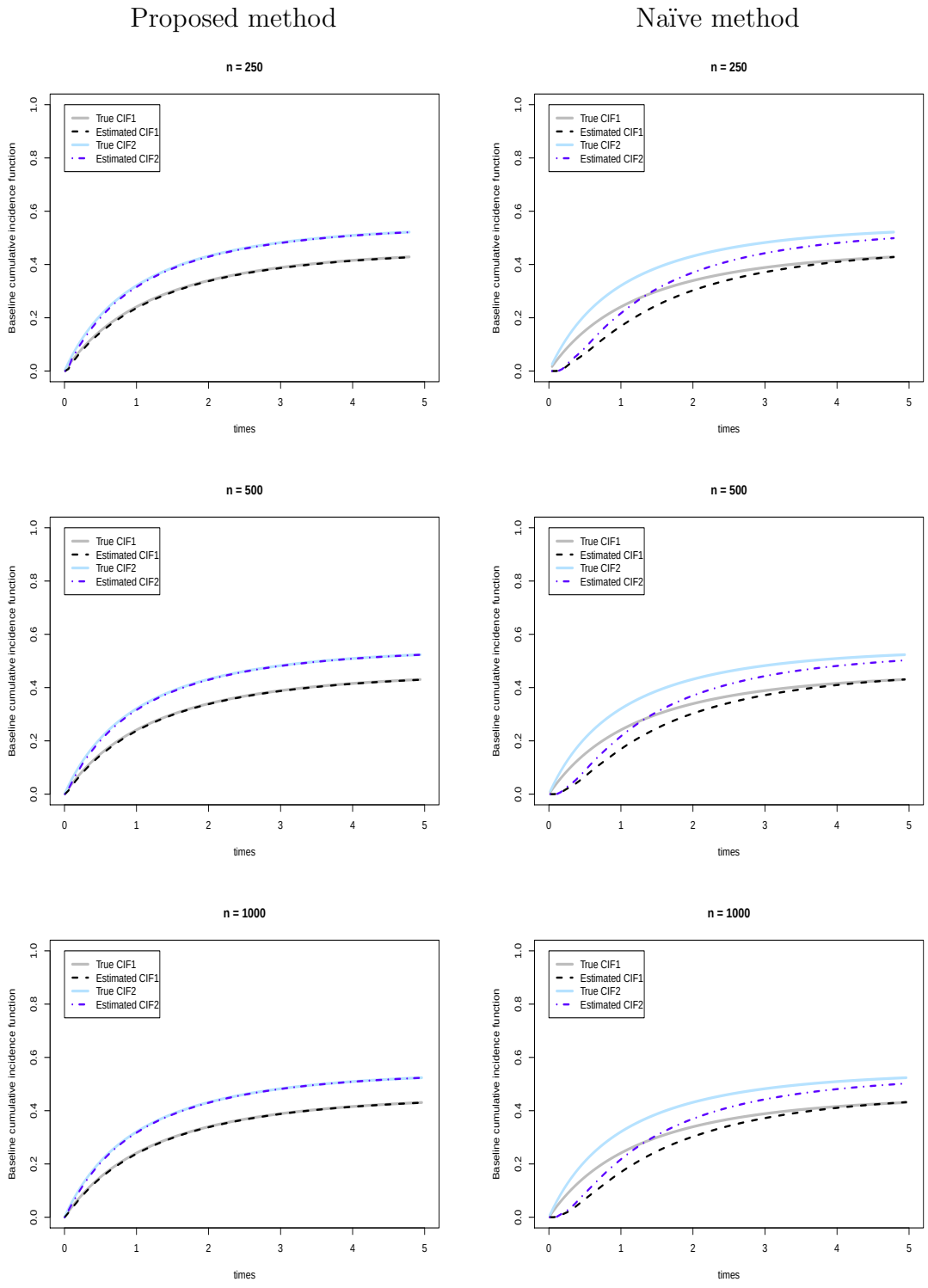



Figure 4.1: The predicted baseline cumulative incidence functions resulted from a simulation study with sample sizes of 250, 500, and 1,000 under a 50% left truncation

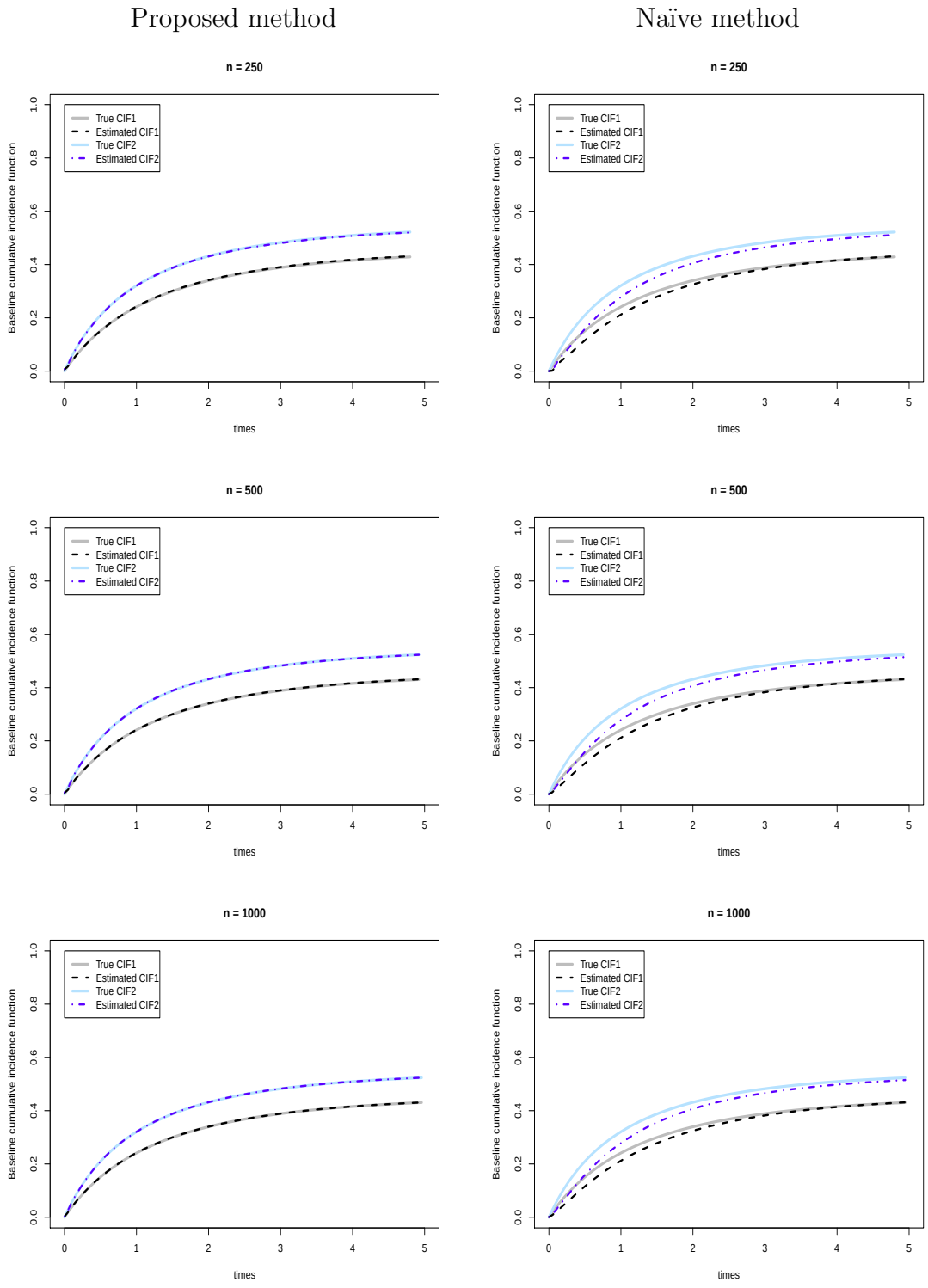



Figure 4.2: The predicted baseline cumulative incidence functions resulted from a simulation study with sample sizes of 250, 500, and 1,000 under a 100% left truncation

In summary, the simulation study provided numerical evidence for the validity of the proposed approach. It also showed that the naïve approach which ignores left truncation can provide seriously biased regression coefficient and baseline CIF estimates. Therefore, the use of the proposed approach in settings with left-truncated and interval-censored competing risks data is crucial.

## 4.4 Analysis of dementia data

The proposed method is illustrated using data from the Indianapolis-Ibadan Dementia Project, which included data of elderly African Americans from Indianapolis. Note that, even though dementia onset times were clearly interval-censored, times to death were more precisely observed. These times were recorded in days and, thus, the interval censoring window was just one day. Strictly speaking, death is also interval-censored, and the one-day interval censoring window satisfies condition C3 in the Appendix of Bakoyannis et al. (2017), which is required for the validity of the B-spline-based sieve maximum likelihood approach for interval-censored competing risks data. The characteristics of the study participants are described in Table 4.3. The study sample consisted of 4,103 individuals. Of them, 450 (11.0%) developed dementia and 2,232 (54.4%) died prior to diagnosis of dementia. Even though the time origin for this analysis was the age of 65 years, individuals were enrolled at any age after 65 years. This posed a problem because these individuals were healthy (i.e. alive and dementia-free) up until they were enrolled in the study. In this sample, 99.1% of individuals were left-truncated. Among individuals who had diagnosed with

dementia, a female share is 67.6%. 33.4% of them had alcohol use and 49.7% of them did not smoke at baseline. The data were analyzed using the R package intccr (Park et al., 2019). The proposed method, implemented using the function ciregic_lt of the intccr package, was compared to the naïve method, implemented using the function ciregic of the intccr package, that does not consider left truncation. In this illustration, gender, alcohol use, and smoking status were considered as potential prognostic factors of dementia. In this analysis, 157 individuals, who had a missing record in alcohol use or smoking status at baseline, were eliminated.

Results from the data analyses are provided in Table 4.4. Based on the proposed analysis males and those with alcohol use and smoking at baseline were less likely to develop dementia. This is attributed to the fact that such individuals have a significantly higher probability of death, as it is estimated using the proposed approach. The most striking difference between the naïve and the proposed was that there was no evidence that smoking is a prognostic factor for dementia based on the naïve analysis.

The covariate-specific CIFs of both dementia and death for individuals without alcohol use and smoking at baseline is depicted by gender in Figure 4.3. The naïve analysis appears to underestimate the CIFs not only for dementia but also death prior to dementia in both males and females. This phenomenon was also observed in this simulation studies and is attributed to the fact that the left-truncated observations reflect healthier individuals. Not accounting for left truncation results in not accounting for the fact that the older individuals have lived up and are dementia-free until the point of entry into the study, which is expected to result in biased estimates.

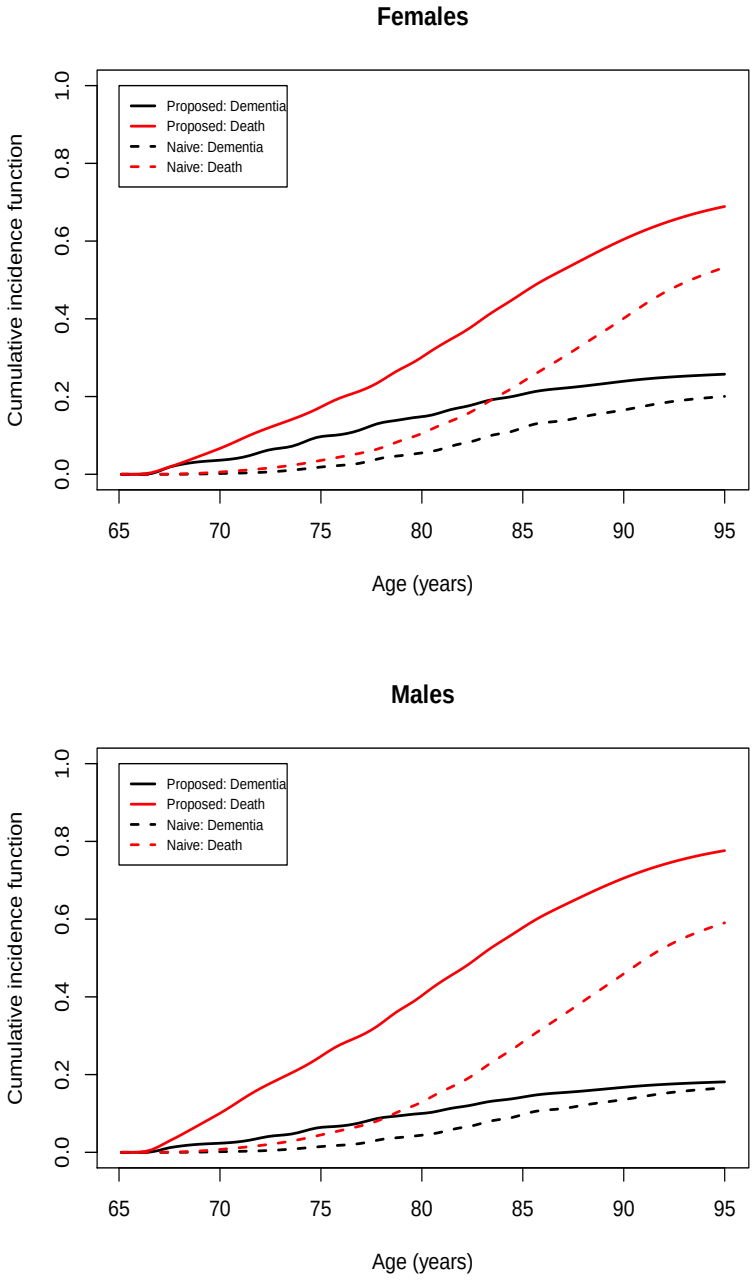



Figure 4.3: The predicted covariate-specific cumulative incidence function of dementia and death for individuals without alcohol use and smoking at baseline

Table 4.3: Descriptive characteristics of the study sample

| | Alive | Dementia | Death |
|---|---|---|---|
| | $(N = 1,347)$ | $(N = 436)$ | $(N = 2,166)$ |
| | $n(\%)$ | $n(\%)$ | $n(\%)$ |
| Left truncation | 1,335 (99.1) | 433 (99.3) | 2,145 (99.0) |
| Gender | | | |
| *Female* | 968 (71.9) | 292 (67.0) | 1,286 (59.4) |
| *Male* | 379 (28.1) | 144 (33.0) | 880 (40.6) |
| Alcohol use | | | |
| *Yes* | 456 (33.9) | 146 (33.5) | 949 (43.8) |
| *No* | 891 (66.1) | 290 (66.5) | 1,217 (56.2) |
| Smoking status | | | |
| *Yes* | 762 (56.6) | 224 (51.4) | 1,453 (67.1) |
| *No* | 585 (43.4) | 212 (48.6) | 713 (32.9) |

Alcohol use and smoking status were measured at baseline.

Table 4.4: Covariate effects on the CIF of dementia and death based on the naïve analysis and the proposed method

| | | Proposed method | | |
|---|---|---|---|---|
| Outcome | Covariate | OR* | 95% CI** | $p$-value |
| Dementia | Gender (*Male versus Female*) | 0.639 | (0.539, 0.757) | <.001 |
| | Alcohol use (*Yes versus No*) | 0.793 | (0.667, 0.941) | 0.008 |
| | Smoking (*Yes versus No*) | 0.537 | (0.459, 0.628) | <.001 |
| Death | Gender (*Male versus Female*) | 1.566 | (1.303, 1.882) | <.001 |
| | Alcohol use (*Yes versus No*) | 1.262 | (1.046, 1.522) | 0.015 |
| | Smoking (*Yes versus No*) | 1.861 | (1.570, 2.207) | <.001 |
| | | Naïve method | | |
| Outcome | Covariate | OR* | 95% CI** | $p$-value |
| Dementia | Gender (*Male versus Female*) | 0.792 | (0.640, 0.980) | 0.032 |
| | Alcohol use (*Yes versus No*) | 0.736 | (0.594, 0.912) | 0.005 |
| | Smoking (*Yes versus No*) | 0.987 | (0.793, 1.227) | 0.904 |
| Death | Gender (*Male versus Female*) | 1.267 | (1.104, 1.455) | 0.001 |
| | Alcohol use (*Yes versus No*) | 1.359 | (1.182, 1.563) | <.001 |
| | Smoking (*Yes versus No*) | 2.202 | (1.902, 2.550) | <.001 |

Standard error estimated by the least-square method; *Odds Ratio; **95% confidence interval

## 4.5 The R function ciregic_lt

The new R function ciregic_lt, which is included in the existing package intccr, can be used to readily implement the proposed methodology (Park et al., 2019). The R package intccr can be downloaded from R CRAN website or can be directly install via in R command line. A short tutorial describes how to use this function. In this Chapter, Windows 10 version of R was used to perform the analysis. Installation of the R package intccr can be performed as follows:

```
R> install.packages(" intccr ")
R> library(intccr)
```

Here the data set longdata_lt, which is in a long format and is included in the intccr package, will be analyzed. The R function dataperp_lt in the R package intccr is required to perform the necessary data management for reshaping data format from long to ready-to-use format because the function ciregic_lt requires a data set in a single data point per subject. The arguments of this function are presented in Table 4.5. The first 5 observations in longdata_lt are

```
R> head(longdata_lt, n = 5)
```

| | id | t | c | w | z1 | z2 |
|---|---|---|---|---|---|---|
| 1 | 1 | 0.3884379 | 0 | 0.06788707 | 1 | −0.6292596 |
| 2 | 1 | 0.5892272 | 2 | 0.06788707 | 1 | −0.6292596 |
| 3 | 2 | 0.3620252 | 0 | 0.01001262 | 1 | 1.0583622 |
| 4 | 2 | 0.4355463 | 2 | 0.01001262 | 1 | 1.0583622 |
| 5 | 3 | 0.8264931 | 0 | 0.79281511 | 0 | 0.8313488 |

Table 4.5: Variables in the data set longdata_lt

| Variables | Description |
|---|---|
| id | a unique identification number |
| t | an observation time |
| c | an event type |
| w | a left truncation time |
| z1 | a binary covariate |
| z2 | a continuous covariate |

Reshaping longdata_lt in ready-to-use wide format by using dataprep_lt can be performed as follows:

```
R> dat <− dataprep_lt(data = longdata_lt, ID = id,
                      W = w, time = t, event = c, Z = c(z1, z2))
```

The first 5 observations in dat are

```
R> head(dat, n = 5)
```

| | id | w | v | u | c | z1 | z2 |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 0.06788707 | 0.3884379 | 0.5892272 | 2 | 1 | −0.6292596 |
| 2 | 2 | 0.01001262 | 0.3620252 | 0.4355463 | 2 | 1 | 1.0583622 |
| 3 | 3 | 0.79281511 | 0.8264931 | 1.4923544 | 2 | 0 | 0.8313488 |
| 4 | 4 | 0.59540435 | 1.5835315 | 1.9462826 | 1 | 0 | −1.7417131 |
| 5 | 5 | 0.24879273 | 2.5284975 | 2.6537938 | 2 | 0 | 0.6452470 |

Table 4.6: Argument of the function ciregic_lt

| Variables | Description |
|---|---|
| formula | a formula describing regression model |
| data | a data frame |
| alpha | a pair of parameters for the link functions |
| k | a parameter that controls the number of internal knots |
| nboot | a number of bootstrap replicates to estimate standard error |
| do.par | a logical constant that define parallel computing |

The dataprep_lt function captures the left truncation time and the time to last visit prior to the event, the time to the first visit after the event, and the corresponding event type. Event type c should be coded as 0, for right-censored observations, 1 or 2.

```
R> table(dat$c)

  0   1   2
 20 124 131
```

The data frame dat contains 20 (7.3%) right-censored observations. The function ciregic_lt has 6 arguments described in Table 4.6. The argument formula formulates the desired statistical model and consists of two elements; the first is the function Surv2(v, u, w, event), which includes the observation times v and u, the left truncation time w and the event type c, and the second is a linear combination of covariates (e.g. Surv2(v = v, u = u, w = w, event = c) $\sim$ z1 + z2). The argument alpha defines the pair of link function parameters $\alpha = (\alpha_1, \alpha_2)$ of the class of generalized odds rate

transformation models for the two event types. The default value is alpha = c(1, 1) which means that the default choice is the proportional odds model for both event types. The argument k $\in (0.5, 1)$ controls the number of internal knots. The actual number of internal knots is set to be $\lfloor kn^{1/3} \rfloor$, with the default value of k being 1. One can use a smaller value to speed up the computation, for example by setting k = 0.5, at the price of a cruder B-spline approximation of the baseline CIFs. The argument nboot defines the number of bootstrap samples for the calculation of the standard errors of the estimated regression coefficients. If nboot = 0 then the least squares approach by J. Huang et al. (2008) is used for standard error estimation. Finally, requesting parallel computing for bootstrap standard error estimation (if nboot > 0), in order to speed up computation, requires setting do.par = TRUE. The analysis of the dat data set, which contains left-truncated and interval-censored competing risks data, can be performed using the ciregic_lt function is as follows:

```
R> fit <- ciregic_lt(formula =
                       Surv2(v = v, u = u, w = w, event = c) ~ z1 + z2,
                       data = dat, alpha = c(1, 1), nboot = 0, do.par = FALSE)
R> summary(fit)

Call:
ciregic_lt .default(formula = Surv2(v = v, u = u, w = w, event = c)
                       ~ z1 + z2, data = dat, alpha = c(1, 1), k = 1,
                       do.par = FALSE, nboot = 0)
Event type 1

   Estimate Std. Error z value Pr(>|z|)
```

```
z1    0.5215      0.2727    1.913    0.0558 .
z2   −0.1835      0.1390   −1.320    0.1868
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Event type 2
   Estimate Std. Error z value Pr(>|z|)
z1   −0.4206      0.2622   −1.604     0.109
z2    0.1581      0.1373    1.151     0.249
```

In order to reproduce the standard error estimates in terms of the nonparametric bootstrap method, a seed number is required. In the example below, the model performs 50 bootstrap replicates to estimate standard error estimation.

```
R> set.seed(2019)
R> fit.npar <− ciregic_lt(formula =
                          Surv2(v = v, u = u, w = w, event = c) ~ z1 + z2,
                          data = dat, alpha = c(1, 1),
                          nboot = 50, do.par = FALSE)
  |========================================| 100%
Completed bootstrapping: 50 out of 50
R> summary(fit.npar)

Call:
 ciregic_lt .default(formula = Surv2(v = v, u = u, w = w, event = c)
                     ~ z1 + z2, data = dat, alpha = c(1, 1), k = 1,
                     do.par = FALSE, nboot = 50)
```

Event type 1

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| z1 | 0.5215 | 0.2898 | 1.800 | 0.0719 | . |
| z2 | −0.1835 | 0.1453 | −1.263 | 0.2066 | |

− − −

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Event type 2

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| z1 | −0.4206 | 0.2931 | −1.435 | 0.151 |
| z2 | 0.1581 | 0.1435 | 1.102 | 0.270 |

A warning message appears if there was at least one bootstrap replicate that did not converge. The warning message contains information on the number of bootstrap sample(s) that did not converge and how many successful bootstrap samples there were out of the nboot replicates. The function summary is a generic function that provides a summary table with estimates, standard errors, z values and $p$-values for both event types. If one chooses to utilize parallel computing by setting do.par = TRUE, the function ciregic_lt selects the maximum number of available cores minus one. Parallel computing can speed up considerably the bootstrap standard error estimation process. Finally, the function predict provides the estimated CIFs for both event types. It requires a sequence of time points and a combination of covariate values to predict the CIF. These estimates can then be plotted using the function plot or other graphical function in R. Below is an example of how to estimate and plot the baseline CIFs (e.g. the CIFs for the covariate pattern covp = c(0, 0)) for the

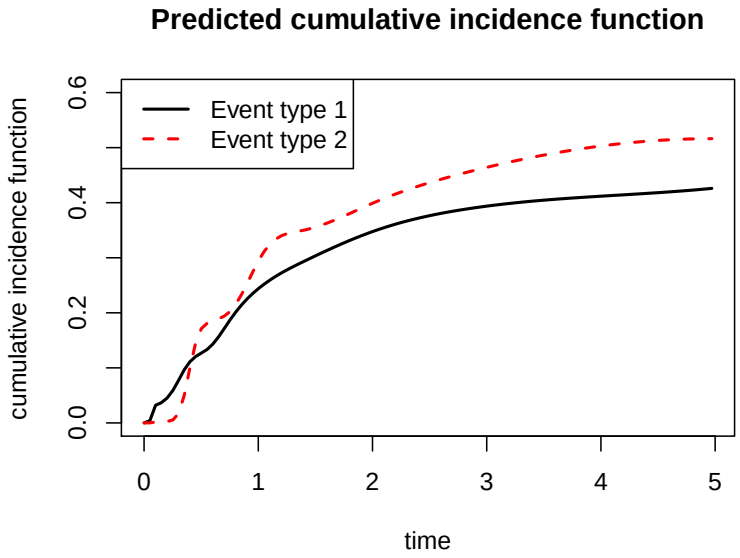



Figure 4.4: The estimated baseline cumulative incidence functions

two event types. Figure 4.4 is the resulting plot for the estimated baseline CIFs for two event types.

R> t <− seq(from = fit$tms[1], to = fit$tms[2],

by = diff(fit$tms) / 99)

R> pred <− predict(object = fit, covp = c(0, 0), times = t)

R> plot(pred$t, pred$cif1, type = "l", ylim = c(0, .6), lwd = 2,

main = ‘‘Predicted cumulative incidence function ’’,

xlab = ‘‘time ’’, ylab = ‘‘cumulative incidence function ’’)

R> points(pred$t, pred$cif2, type = ‘‘l ’’, col = 2, lty = 2, lwd = 2)

R> legend(‘‘topleft ’’, legend = c(‘‘event type 1’’, ‘‘event type 2’’),

lty = 1:2, col = 1:2, lwd = c(2, 2))

## 4.6 Conclusions

In this chapter, the issue of semiparametric analysis of the CIF is addressed under left-truncated and interval-censored competing risks data. More precisely, this approach is extended the semiparametric B-spline-based sieve maximum likelihood estimator for interval-censored competing risks data by Bakoyannis et al. (2017) to also account for left truncation. The general class of generalized odds rate transformation models, which includes the proportional odds and the proportional subdistribution hazards models as special cases, is considered. Standard error estimation can be performed either via nonparametric bootstrap method by G. Cheng et al. (2010) or using the least squares approach by J. Huang et al. (2008). The simulation experiments provide numerical evidence for the validity of the proposed methodology. Additionally, the simulation experiments show that ignoring left truncation can lead to substantial bias. The proposed approach can be readily applied using the new function ciregic_lt which is included in the R package intccr (Park et al., 2019). A short tutorial on the use of the ciregic_lt function was provided in Chapter 4.5.

The CIF explicitly quantifies clinical prognosis in time-to-event studies with competing risks and, thus, it is a crucial quantity in medical decision making. Specifically, it can be used for classifying patients into risk groups, informing patients about their likely course of disease, and guiding physicians and patients in their decision making about the optimal treatment option (Royston et al., 2009; Moons et al., 2009). Moreover, the CIF can be used for identifying appropriate patients for recruitment in clinical trials (Moons et al., 2009). This estimand is also important for informing health policy, such as for guiding resource allocation, and is also of key importance

in quality of life studies (Fine & Gray, 1999). Fitting prognostic models in practice is frequently complicated due to interval censoring, left truncation, and competing risks, situations that are common in cohort studies of chronic diseases. Several semiparametric approaches have been proposed for the CIF that account for either left truncation (X. Zhang et al., 2011; Geskus, 2011; P.-S. Shen, 2011) or interval censoring (Li, 2016; Bakoyannis et al., 2017; L. Mao et al., 2017). To the best of my knowledge, the issue of dealing with both left truncation and interval censoring in the framework of semiparametric analysis of the CIF with competing risks data has not been addressed so far. Chapter 4 addressed this significant gap in the literature.

In motivating study, dementia onset times were interval-censored while death times were more precisely observed. These times were recorded in days and, thus, the interval censoring window was just one day. Therefore, strictly speaking, death is also interval-censored, and the one-day interval censoring window satisfies condition C3 in the Appendix of Bakoyannis et al. (2017), which is required for the validity of the B-spline-based sieve maximum likelihood approach for interval-censored competing risks data (Bakoyannis et al., 2017). This phenomenon is encountered in many applications and in particular studies of chronic conditions because they typically involve death as a competing risk. The analysis of the dementia data revealed that ignoring left truncation leads to underestimated CIFs of dementia and death. This underestimation if attributed to the fact that individuals who enter the study after the age of 65 are dementia-free and, thus, consist a sample of healthier individuals. Moreover, it was found using the proposed methodology that males, individuals who smoke, and those who use alcohol have a lower CIF of dementia. This counter-intuitive, from an aetiology perspective, result is attributed to the fact that such individuals

are more likely to die prior to developing dementia. This result illustrates that the analysis of the CIF is not appropriate for the evaluation of risk factors from an aetiology perspective (Koller et al., 2012; Andersen et al., 2012). On the contrary, the CIF quantifies clinical prognosis and is useful in clinical decision making (Koller et al., 2012; Andersen et al., 2012; Bakoyannis et al., 2017). For example, the estimated model based on the dementia data could be used in clinical practice for the calculation of the estimated cumulative probability of dementia for an older adult.

An interesting area for future research is the quantification of the predictive accuracy of models for the CIF with left-truncated and interval-censored competing risks data. Such an approach can be either based on time-dependent receiver operating characteristic curves (Saha & Heagerty, 2010) or Brier scores (Schoop et al., 2011). The standard methods for quantifying predictive accuracy with competing risks data are not applicable in the presence of interval censoring and left truncation. Furthermore, incorporating internal time-dependent covariates (Cortese & Andersen, 2010) for dynamic predictions with left-truncated and interval-censored competing risks data is another important, from a practical standpoint, topic for future research. For this, one can consider either joint models (Elashoff et al., 2008; X. Huang et al., 2011; Andrinopoulou et al., 2014) or more flexible, and computationally efficient, landmark approaches (Nicolaie et al., 2013). Finally, the plan is to extend the R function ciregic_lt to allow for more than two event types.

# Chapter 5

# Discussion

The analysis of the CIF for interval-censored competing risks data benefits from many advanced methodologies. However, a universal tool that helps the application of method is even more elusive. The goal in this dissertation was to fill out the gap between the development and application of methods. My dissertation addresses not only development of software in Chapters 2 but also new methodologies that deal with missing event types in Chapter 3 and left truncation in Chapter 4.

In Chapter 2, a comprehensive R package intccr was developed to offer semiparametric regression analysis for the CIF on interval-censored competing risks data. The R package intccr covers a class of semiparametric generalized odds rate transformation models including the Fine–Gray model and the proportional odds model as special cases. The selection of the number of internal knots in the B-splines approximation follows the guideline in Section 2.1 in Bakoyannis, Yu, & Yiannoutsos (2017). Concisely, the number of internal knots for the B-splines can be chosen by $\lfloor k \times n^{1/3} \rfloor$ where $k \in [0.5, 1]$ is a parameter that determines smoothness and $n$ is a sample size. Please see Section 2.1 in Bakoyannis, Yu, & Yiannoutsos (2017) for more details. Data sometimes consist of a relatively large number of variables with a small number of data points or observations. This type of data may result in non-convergence issue during the estimating process. To avoid the problem, the following rule of thumb is suggested for the maximum number of regression coefficients to be estimated (or

equivalently the number of covariates) for each event type:

$$\text{Maximum number of covariates} = \left\lfloor \frac{\min(n_1, n_2)}{10} \right\rfloor$$

where $n_j$ for $j = 1, 2$ is the number of observations with event type $j$. The R package intccr is an important development in estimating the CIF for interval-censored competing risks data. One limitation of the R package intccr is that, for the time being, it only allows for two event types. It is indispensable to update the package intccr to allow for more than two event types in the near future.

In Chapter 3, the AIPW technique was proposed to account for missing event types in interval-censored competing risks data. This extended the B-spline-based sieve maximum likelihood method by incorporating parametric models for the probability of missingness and event type into the estimator. Also, weaker missing at random assumption was imposed to the estimator by involving auxiliary variables that may be associated with the probability of missingness. The methodology offers doubly robust estimator under any portions of missing event types. This means that the estimator is consistent even if either the model for the probability of missingness or the model for the probability of event type is incorrectly specified. In practice, interval censoring and missing event types are common problems which are typically met in studies based on electronic health records and can lead to bias, as illustrated in the simulation experiments. The R function ciregic_aipw, that applies the methodology and was built in the R package intccr, has the potential needs in real-world applications. The R function ciregic_aipw allows for only two event types which is

sufficient for many applications. However, it is required to extend the function to allow for an arbitrary (finite) number of event types in the future.

In Chapter 4, the issue of semiparametric analysis of the CIF was addressed under left-truncated and interval-censored competing risks data. To the best of my knowledge, the issue of dealing with both left truncation and interval censoring in the framework of semiparametric analysis of the CIF with competing risks data has not been addressed. The proposed methodology is the extension of the B-spline-based sieve maximum likelihood method for interval-censored competing risks data proposed by Bakoyannis et al. (2017) by introducing a contribution of the left truncation. A class of generalized odds rate transformation models, including the proportional odds and the proportional subdistribution hazards models as special cases, is considered. Moreover, the least-squares method in a class of semiparametric odds rate transformation models to estimate the standard error of the coefficient was built in the R function ciregic_lt in addition to the nonparametric bootstrap method in the procedure of estimating the standard error of the coefficient. Future research should consider the quantification of the predictive accuracy of models for the CIF with left-truncated and interval-censored competing risks data. Furthermore, another interesting research topic for future research is dynamic predictions by incorporating internal time-dependent covariates in left-truncated and interval-censored competing risks data.

In summary, throughout this dissertation novel statistical methodologies and comprehensive tool were developed to resolve the issue–interval censoring, left truncation, and missing event types–of semiparametric regression analysis for the CIF on competing risks data. Further studies are necessary to develop a closed form of stan-

dard error estimate of the regression coefficient and certainly required to investigate the asymptotic property of the standard error estimates followed from the AIPW method.

References

Aalen, O. O., & Johansen, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, *5*(3), 141–150.

Allignol, A. (2017). kmi: Kaplan-Meier multiple imputation for the analysis of cumulative incidence functions in the competing risks setting [Computer software manual]. (R package version 0.5.3)

Andersen, P. K., Geskus, R. B., de Witte, T., & Putter, H. (2012). Competing risks in epidemiology: possibilities and pitfalls. *International journal of epidemiology*, *41*(3), 861–870.

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

Bakoyannis, G. (2020). Nonparametric tests for transition probabilities in nonhomogeneous Markov processes. *Journal of Nonparametric Statistics*, *In press*.

Bakoyannis, G., Siannis, F., & Touloumi, G. (2010). Modelling competing risks data with missing cause of failure. *Statistics in Medicine*, *29*(30), 3172-3185.

Bakoyannis, G., & Touloumi, G. (2012). Practical methods for competing risks data: A review. *Statistical Methods in Medical Research*, *21*(3), 257–272.

Bakoyannis, G., & Touloumi, G. (2017). Impact of dependent left truncation in semiparametric competing risks methods: a simulation study. *Communications in Statistics-Simulation and Computation*, *46*(3), 2025–2042.

Bakoyannis, G., Yu, M., & Yiannoutsos, C. T. (2017). Semiparametric regression on cumulative incidence function with interval-censored competing risks data. *Statistics in Medicine*, *36*(23), 3683–3707.

Bakoyannis, G., Zhang, Y., & Yiannoutsos, C. T. (2019). Nonparametric inference for Markov processes with missing absorbing state. *Statistica Sinica*, *29*(4), 2083–2104.

Chen, D.-G., Sun, J., & Peace, K. E. (2012). *Interval-censored time-to-event data: methods and applications*. Chapman & Hall/CRC Biostatistics Series.

Cheng, G., Huang, J. Z., et al. (2010). Bootstrap consistency for general semiparametric M-estimation. *The Annals of Statistics*, *38*(5), 2884–2915.

Cheng, Y.-J., Wang, M.-C., & Tsai, C.-Y. (2019). Estimations of the joint distribution of failure time and failure type with dependent truncation. *Biometrics*, *75*(2), 428–438.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330-351.

Cortese, G., & Andersen, P. K. (2010). Competing risks and time-dependent covariates. *Biometrical Journal*, *52*(1), 138–158.

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

Delord, M. (2017). MIICD: Multiple imputation for interval censored data [Computer software manual]. (R package version 2.4)

Dlugosz, S., Peng, L., & Li, R. (2016). cmprskQR: Analysis of competing risks using quantile regressions [Computer software manual]. (R package version 0.9.1)

Do, G., & Kim, Y.-J. (2017). Analysis of interval censored competing risk data with missing causes of failure using pseudo values approach. *Journal of Statistical Computation and Simulation*, *87*(4), 631-639.

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

Eriksson, F., Li, J., Scheike Harder, T., & Zhang, M.-J. (2015). The proportional odds cumulative incidence model for competing risks. *Biometrics*, *71*(3), 687-695.

Fine, J. P. (2001). Regression modeling of competing crude failure probabilities. *Biostatistics*, *2*(1), 85–97.

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

Gao, G., & Tsiatis, A. A. (2005). Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. *Biometrika*, *92*(4), 875-891.

Gerds, T. A. (2017). prodlim: Product-limit estimation for censored event history analysis [Computer software manual]. (R package version 1.6.1)

Gerds, T. A., Scheike, T. H., Blanche, P., & Ozenne, B. (2017). riskregression: Risk regression models and prediction scores for survival analysis with competing risks [Computer software manual]. (R package version 1.4.3)

Geskus, R. (2011). Cause-specific cumulative incidence estimation and the Fine and Gray model under both left truncation and right censoring. *Biometrics*, *67*, 39–49.

Grambauer, N., & Neudecker, A. (2011). compei: Event-specific incidence rates for competing risks data [Computer software manual]. (R package version 1.0)

Gray, B. (2017). cmprsk: Subdistribution analysis of competing risks [Computer software manual]. (R package version 2.2-7)

Groeneboom, P., Jongbloed, G., & Wellner, J. A. (2008). The support reduction algorithm for computing non-parametric function estimates in mixture models. *Scandinavian Journal of Statistics*, *35*(3), 385–399.

Hall, K. S., Gao, S., Baiyewu, O., Lane, K. A., Gureje, O., Shen, J., ... others (2009). Prevalence rates for dementia and Alzheimer's disease in African Americans: 1992 versus 2001. *Alzheimer's & Dementia*, *5*(3), 227–233.

Hendrie, H. C., Osuntokun, B. O., Hall, K. S., Ogunniyi, A. O., Hui, S. L., Unverzagt, F. W., ... others (1995). Prevalence of Alzheimer's disease and dementia in two communities: Nigerian Africans and African Americans. *The American Journal of Psychiatry*.

Hendrie, H. C., Zheng, M., Li, W., Lane, K., Ambuehl, R., Purnell, C., ... others (2017). Glucose level decline precedes dementia in elderly African Americans with diabetes. *Alzheimer's & Dementia*, *13*(2), 111–118.

Huang, J., Zhang, Y., & Hua, L. (2008). *A least-squares approach to consistent information estimation in semiparametric models* (Tech. Rep. No. Technical Report 2008-3). Department of Biostatistics: University of Iowa.

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

Hudgens, M. G., Satten, G. A., & Longini Jr., I. M. (2001). Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation. *Biometrics*, *57*(1), 74-80.

Jeong, J.-H., & Fine, J. P. (2006). Parametric regression on cumulative incidence function. *Biostatistics*, *8*(2), 184–196.

Kalbfleisch, J., & Prentice, R. (2011). *The statistical analysis of failure time data*. Wiley & Sons.

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

Klein, J. P., & Moeschberger, M. (2013). *Survival analysis: Techniques for censored and truncated data*. Springer, New York.

Koller, M. T., Raatz, H., Steyerberg, E. W., & Wolbers, M. (2012). Competing risks and the clinical community: Irrelevance or ignorance? *Statistics in Medicine*, *31*(11-12), 1089–1097.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer, New York.

Lee, M., Dignam, J. J., & Han, J. (2014). Multiple imputation methods for nonparametric inference on cumulative incidence with missing cause of failure. *Statistics in Medicine*, *33*(26), 4605–4626.

Li, C. (2016). The Fine–Gray model under interval censored competing risks data. *Journal of Multivariate Analysis*, *143*, 327–344.

Little, R., & Rubin, D. (2002). *Statistical analysis with missing data*. Wiley & Sons.

Lu, K., & Tsiatis, A. A. (2001). Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics*, *57*(4), 1191–1197.

Maathuis, M. H. (2005). Reduction algorithm for the NPMLE for the distribution function of bivariate interval-censored data. *Journal of Computational and Graphical Statistics*, *14*(2), 352–362.

Maathuis, M. H. (2013). MLEcens: Computation of the MLE for bivariate (interval) censored data [Computer software manual]. (R package version 0.1-4)

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

Mao, L., Lin, D.-Y., & Zeng, D. (2017). Semiparametric regression analysis of interval-censored competing risks data. *Biometrics*, *73*(3), 857-865.

Mao, M., & Wang, J.-L. (2010). Semiparametric efficient estimation for a class of generalized proportional odds cure models. *Journal of the American Statistical Association*, *105*(489), 302–311.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*(4), 538–558.

Moons, K. G., Royston, P., Vergouwe, Y., Grobbee, D. E., & Altman, D. G. (2009). Prognosis and prognostic research: what, why, and how? *British Medical Journal*, *338*, b375.

Moreno-Betancur, M., & Latouche, A. (2013). Regression modeling of the cumulative incidence function with missing causes of failure using pseudo-values. *Statistics in Medicine*, *32*(18), 3206–3223.

Nicolaie, M., Van Houwelingen, J., De Witte, T., & Putter, H. (2013). Dynamic prediction by landmarking in competing risks. *Statistics in Medicine*, *32*(12), 2031–2047.

Pan, W. (2000). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics*, *56*(1), 199–203.

Park, J., Bakoyannis, G., & Yiannoutsos, C. T. (2019). Semiparametric competing risks regression under interval censoring using the R package intccr. *Computer Methods and Programs in Biomedicine*, *173*, 167–176.

Peng, L., & Fine, J. P. (2009). Competing risks quantile regression. *Journal of the American Statistical Association*, *104*(488), 1440–1453.

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

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Robins, J. M., & Wang, N. (2000). Inference for imputation estimators. *Biometrika*, *87*(1), 113–124.

Royston, P., Moons, K. G., Altman, D. G., & Vergouwe, Y. (2009). Prognosis and prognostic research: developing a prognostic model. *British Medical Journal*, *338*, b604.

Saha, P., & Heagerty, P. (2010). Time-dependent predictive accuracy in the presence of competing risks. *Biometrics*, *66*(4), 999–1011.

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

Scheike, T. H., Martinussen, T., Martinussen, J., & Hols, K. (2017). timereg: Flexible regression models for survival data [Computer software manual]. (R package version 1.9.1)

Scheike, T. H., & Zhang, M.-J. (2011). Analyzing competing risk data using the R timereg package. *Journal of Statistical Software*, *38*(2).

Scheike, T. H., Zhang, M.-J., & Gerds, T. A. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, *95*(1), 206–220.

Schoop, R., Beyersmann, J., Schumacher, M., & Binder, H. (2011). Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal*, *53*(1), 88–112.

Shen, P.-S. (2011). Proportional subdistribution hazards regression for left-truncated competing risks data. *Journal of Nonparametric Statistics*, *23*(4), 885–895.

Shen, X., & Wong, W. H. (1994). Convergence rate of sieve estimates. *The Annals of Statistics*, *22*(2), 580–615.

Sun, J. (2006). *The statistical analysis of interval-censored failure time data*. Springer-Verlag New York.

Therneau, T. M., & Lumley, T. (2016). survival: Survival analysis [Computer software manual]. (R package version 2.41-3)

Van der Vaar, A., & Wellner, J. (1996). *Weak convergence and empirical processes with applications to statistics*. Springer, New York.

Wang, M.-C., Brookmeyer, R., & Jewell, N. P. (1993). Statistical models for prevalent cohort data. *Biometrics*, *49*(1), 1–11.

Wang, N., & Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, *85*(4), 935–948.

Wolbers, M., Koller, M. T., Witteman, J. C. M., & Steyerberg, E. W. (2009). Prognostic models with competing risks: Methods and application to coronary risk prediction. *Epidemiology*, *20*(4), 555-561.

Wolfson, D., Best, A., Addona, V., Wolfson, J., & Gadalla, S. (2019). Benefits of combining prevalent and incident cohorts: An application to myotonic dystrophy. *Statistical Methods in Medical Research*, *28*(10–11), 3333–3345.

Zeng, D., Yin, G., & Ibrahim, J. G. (2006). Semiparametric transformation models for survival data with a cure fraction. *Journal of the American Statistical Association*, *101*(474), 670-684.

Zhang, X., Zhang, M.-J., & Fine, J. (2011). A proportional hazards regression model for the subdistribution with right-censored and left-truncated competing risks data. *Statistics in Medicine*, *30*, 1933–1951.

Zhang, Y., Hua, L., & Huang, J. (2010). A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. *Scandinavian Journal of Statistics*, *37*(2), 338–354.

Zhang, Z., & Sun, J. (2010). Interval censoring. *Statistical Methods in Medical Research*, *19*(1), 53–70.

Curriculum Vitae

# Jun Park

## Education

- Ph.D. in Biostatistics, Minor in Epidemiology 2020

  Indiana University, Indianapolis, IN

- M.S. in Statistics 2014

  Rutgers University, New Brunswick, NJ

- B.S. in Statistics 2011

  Soongsil University, Seoul, Republic of Korea

## Working Experience

- Intern May 2018 - August 2018

  Merck Biostatistics and Research Decision Sciences, Upper Gwynedd, PA

- Research Assistant January 2017 - December 2019

  Department of Biostatistics, Indiana University, Indianapolis, IN,

- Teaching Assistant August 2015 - December 2016

  Department of Mathematical Sciences, Indiana University Purdue University Indianapolis, Indianapolis, IN

## Awards

- Graduate and Professional Education Grant, IUPUI 2019

- Travel award, Richard M. Fairbanks School of Public Health, IUPUI 2018
- Outstanding Beginning Graduate Student Award, IUPUI 2016
- First Place for the 7th SAS Data Mining Championship, SAS Korea 2009
- Award of Excellent for Achievement, Soongsil University 2009
- Scholarship, Soongsil University 2008 - 2010