

MULTINOMIAL LOGISTIC REGRESSION AND  
PREDICTION ACCURACY FOR  
INTERVAL-CENSORED COMPETING RISKS  
DATA

by

**Yongli Shuai**

B. A., Xi'an Jiaotong University, China, 1997

M. S., Wright State University, 2004

Submitted to the Graduate Faculty of  
the Graduate School of Public Health in partial fulfillment  
of the requirements for the degree of

**Doctor of Philosophy**

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH  
GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Yongli Shuai

It was defended on

December 8<sup>th</sup>, 2017

and approved by

Jong H. Jeong, Ph. D.

Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Yu Cheng, Ph. D.

Associate Professor

Department of Statistics

The Dietrich School of Arts & Sciences

University of Pittsburgh

Chung-Chou H. Chang, Ph. D.

Professor

Departments of Medicine and Biostatistics

School of Medicine and Graduate School of Public Health

University of Pittsburgh

Ying Ding, Ph. D.

Assistant Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Dissertation Advisors: Jong H. Jeong, Ph. D.

Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh,

Yu Cheng, Ph. D.

Associate Professor

Department of Statistics

The Dietrich School of Arts & Sciences

University of Pittsburgh

Copyright © by Yongli Shuai  
2018

# MULTINOMIAL LOGISTIC REGRESSION AND PREDICTION ACCURACY FOR INTERVAL-CENSORED COMPETING RISKS DATA

Yongli Shuai, PhD

University of Pittsburgh, 2018

## ABSTRACT

Interval-censored competing risks data are ubiquitous in biomedical research fields. The direct parametric modeling of the cumulative incidence functional (CIF) is appealing due to its intuitive probability interpretation and easy implementation. This dissertation is to study and extend the multinomial logistic regression (MLR) model to interval-censored competing risks data. The MLR model naturally guarantees the additivity property of the event-specific probabilities under competing risks. A cubic B-Spline-based sieve method is then adopted to add flexibility into the proposed MLR model. The second study objective is to develop the prediction error (PE) as a model-free metric to evaluate and validate the prediction accuracy for interval-censored competing risks data. Adopting the method of the pseudo-value estimator, this dissertation work proposes a novel approach to estimate the PE under the interval-censored competing risks setting. Simulation studies are presented to assess performance of the MLR model and the PE in different scenarios. The proposed methods were then applied to a community-based study of cognitive impairment in aging population.

**Public Health Significance:** Interval-censored competing risks data could be often encountered in biomedical research that is essential for public health, such as rehabilitation and pain medicine. The proposed methods provide precise yet flexible modeling of such data with straightforward interpretation on how predictors affect the CIF, as well as useful tools to evaluate and validate the prediction accuracy of the developed models.

**Keywords:** Cumulative incidence function; Parametric model; Odds ratios; B-Spline; Prediction error.

## TABLE OF CONTENTS

<b>1.0 INTRODUCTION</b> . . . . .	1
<b>2.0 MULTINOMIAL LOGISTIC REGRESSION FOR INTERVAL-CENSORED COMPETING RISKS DATA</b> . . . . .	7
2.1 Introduction . . . . .	7
2.2 Method . . . . .	8
2.2.1 Multinomial Logistic Regression (MLR) for the CIF . . . . .	8
2.2.2 Independence of Irrelevant Alternative (IIA) in the MLR model . . . . .	10
2.2.3 Maximum Likelihood Estimation and Inference in the MLR Model . . . . .	11
2.2.4 Cubic B-Spline-based Sieve MLE . . . . .	13
2.3 Simulation Studies . . . . .	16
2.3.1 MLR Model Correctly Specified . . . . .	16
2.3.2 MLR Model Misspecified for Fine and Gray Model . . . . .	22
2.3.3 MLR Model Misspecified for Weibull Distributed Failure Time . . . . .	26
2.4 Application to the Study of Dementia Epidemiology . . . . .	32
2.5 Discussion . . . . .	35
<b>3.0 PREDICTION ACCURACY FOR MODELLING INTERVAL-CENSORED COMPETING RISKS DATA</b> . . . . .	37
3.1 Introduction . . . . .	37
3.2 Method . . . . .	39
3.2.1 Definition . . . . .	39
3.2.2 Estimating the Prediction Error . . . . .	41

3.3 Simulation Studies . . . . .	44
3.3.1 PE from Correctly Specified MLR Model . . . . .	45
3.3.2 PE from MLR Misspecified for Fine and Gray Model . . . . .	48
3.3.3 PE from MLR Misspecified for Weibull Distributed Failure Time . . . . .	53
3.4 Application to the Study of Dementia Epidemiology . . . . .	58
3.5 Discussion . . . . .	60
<b>BIBLIOGRAPHY . . . . .</b>	<b>62</b>

## LIST OF TABLES

2.1 MLR Parameter estimation accuracy from simulation. . . . .	19
--	----



## LIST OF FIGURES

2.1	Estimated CIFs for cause-1 under correctly specified MLR. . . . .	20
2.2	Estimated CIFs for cause-2 under correctly specified MLR. . . . .	20
2.3	CIFs for cause-1 under correctly specified MLR with a cubic B-Spline (knots=3). . . . .	21
2.4	Estimated CIFs for cause-2 under correctly specified MLR with a cubic B-Spline (knots=3). . . . .	21
2.5	Estimated cause-1 CIFs from misspecified MLR with linear base; true CIFs from Fine and Gray. . . . .	23
2.6	Estimated cause-2 CIFs from misspecified MLR with linear base; true CIFs from Fine and Gray. . . . .	24
2.7	Estimated cause-1 CIFs from misspecified MLR with polynomial base; true CIFs from Fine and Gray . . . . .	24
2.8	Estimated cause-2 CIFs from misspecified MLR with linear base; true CIFs from Fine and Gray . . . . .	25
2.9	Estimated cause-1 CIFs from misspecified MLR with a cubic B-Spline (knots=3); true CIFs from Fine and Gray. . . . .	25
2.10	Estimated cause-1 CIFs from misspecified MLR with a cubic B-Spline (knots=3); true CIFs from Fine and Gray. . . . .	26
2.11	Estimated cause-1 CIFs from misspecified MLR with a linear base; true CIFs from Weibull ( <i>shape</i> = 0.5). . . . .	28
2.12	Estimated cause-2 CIFs from misspecified MLR with a linear base; true CIFs from Weibull ( <i>shape</i> = 0.5). . . . .	29

2.13	Estimated cause-1 CIFs from misspecified MLR with a linear base; true CIFs from Weibull ( <i>shape</i> = 2).	29
2.14	Estimated cause-2 CIFs from misspecified MLR with a linear base; true CIFs from Weibull ( <i>shape</i> = 2).	30
2.15	Estimated cause-1 CIFs from misspecified MLR with a cubic B-Spline (knots=3); true CIFs from Weibull ( <i>shape</i> = 0.5).	30
2.16	Estimated cause-2 CIFs from misspecified MLR with a cubic B-Spline (knots=3); true CIFs from Weibull ( <i>shape</i> = 0.5).	31
2.17	Estimated cause-1 CIFs from misspecified MLR with a cubic B-Spline (knots=4); true CIFs from Weibull ( <i>shape</i> = 2).	31
2.18	Estimated cause-2 CIFs from misspecified MLR with a cubic B-Spline (knots=4); true CIFs from Weibull ( <i>shape</i> = 2).	32
2.19	Estimated CIFs for the MCI from the MLR model with a cubic B-Spline (knots=6).	34
3.1	PE Estimates For MLR CIF-1	46
3.2	PE Estimates For MLR CIF-2.	46
3.3	PE Estimates For MLR CIF-1 with a cubic B-Spline (knots=3).	47
3.4	PE Estimates For MLR CIF-2 with a cubic B-Spline (knots=3).	47
3.5	PE for the cause-1 CIF based on MLR model with a linear function; true data from the Fine and Gray model.	49
3.6	PE for the cause-2 CIF based on MLR model with a linear function; true data from the Fine and Gray model.	50
3.7	PE for the cause-1 CIF based on MLR model with a polynomial function; true data from the Fine and Gray model.	50
3.8	PE for the cause-2 CIF based on MLR model with a polynomial function; true data from the Fine and Gray model.	51
3.9	PE for the cause-1 CIF based on MLR model with or without a polynomial function; true data from the Fine and Gray model.	51

3.10 PE for the cause-2 CIF based on MLR model with or without a polynomial function; true data from the Fine and Gray model. . . . .	52
3.11 PE for the cause-1 CIF based on MLR model with cubic spline; true data from the Fine and Gray model. . . . .	52
3.12 PE for the cause-2 CIF based on MLR model with cubic spline; true data from the Fine and Gray model. . . . .	53
3.13 PE for the cause-1 CIF based on MLR model with a linear function; true data from Weibull distributed ( $shape = 0.5$ ). . . . .	54
3.14 PE for the cause-2 CIF based on MLR model with a linear function; true data from Weibull distributed ( $shape = 0.5$ ). . . . .	55
3.15 PE for the cause-1 CIF based on MLR model with a cubic B-spline; true data from Weibull distributed ( $shape = 0.5$ ). . . . .	55
3.16 PE for the cause-2 CIF based on MLR model with a cubic B-spline; true data from Weibull distributed ( $shape = 0.5$ ). . . . .	56
3.17 PE for the cause-1 CIF based on MLR model with a cubic B-spline; true data from Weibull distributed ( $shape = 2$ ). . . . .	56
3.18 PE for the cause-2 CIF based on MLR model with a cubic B-spline; true data from Weibull distributed ( $shape = 2$ ). . . . .	57
3.19 PE for the cause-1 CIF based on MLR model with a cubic B-spline; true data from Weibull distributed ( $shape = 2$ ). . . . .	57
3.20 PE for the cause-2 CIF based on MLR model with a cubic B-spline; true data from Weibull distributed ( $shape = 2$ ). . . . .	58
3.21 PE based on the different models with a B-Spline. . . . .	59
3.22 PE estimates comparison among all different models. . . . .	60

## 1.0 INTRODUCTION

In medical studies, subjects are often followed with periodical examination and their events of interest are only observed to fall between two consecutive examination times instead of the exact time. The resulting data are called interval-censored data, where the time of an event of interest is only known to have occurred in some interval  $(L, R)$ . Interval-censored data are emerged extensively in biomedical research fields including clinical, laboratory and epidemiological data. [Chen et al. \(2012\)](#) summarized interval-censored data as current status data (also called case 1 data, while each subject has only one follow-up examination to infer the occurrence of an event of interest), case 2 data (including interval-censored, right-censored and left-censored data) and mixed data (including case 1, case 2 and non-censored data). Right or left censored data could be thought as special cases of interval-censored data which have an infinity as the right or left boundary. However, as incomplete data in the survival analysis, they have different data structures and as well censoring mechanisms. Thus the theory and method developed under right or left censored data can not be directly applied to the interval-censored data without extra care.

Interval-censored data have been an active research area for decades. The theory and method to analyze interval-censored data have been developed on non-parametric estimation and comparison of the survival function(s) ([Peto, 1973](#); [Turnbull, 1976](#)); parametric or semi-parametric regression models for the hazard or failure-time functions ([Sun et al., 2015](#)), and re-sampling approach including imputation to take advantage of existing methods ([Tanner and Wong, 1987](#); [Wei and Tanner, 1991](#)).

Competing risks analysis arises while subjects face two or more mutually exclusive causes of failure during research period. In the framework of competing risks setting, only one of

possible events will occur and can be observed if the follow-up sustains. All the other competing events are precluded from being observed and thus censored. Competing risks data are also a form of incomplete data and this incomplete characteristic issues the primary challenge in research.

Research on competing risks data is very active in the literature mainly focused on right-censored competing risks data. The cause-specific hazard, subdistribution hazard and the cumulative incidence function (CIF) are three fundamental quantities for competing risks analysis. Under right-censored competing risks data setting, the commonly developed methods include non-parametric test established by [Gray \(1988\)](#), the cause-specific hazard regression ([Prentice et al., 1978](#); [Lunn and McNeil, 1995](#); [Kalbfleisch and Prentice, 2002](#); [Scheike and Zhang, 2008](#); [Belot et al., 2010](#)), [Fine and Gray \(1999\)](#) subdistribution hazards model and its extension on likelihood based semi-parametric transformation models ([Eriksson et al., 2015](#); [Mao and Lin, 2016](#)). [Jeong and Fine \(2006; 2007\)](#) proposed direct parametric modelling CIF and parametric regression on CIF under right censored competing risks data. [Klein and Andersen \(2005\)](#) developed CIF regression approach which is based on the pseudo-observation method and generalized estimating equation approach (GEE). Their method focused on direct effects of covariates on CIFs. These models share the form of generalized transformation regression and have similar parameter interpretation corresponding to the link function in practice.

In biomedical research, subjects often are exposed to two or more mutually exclusive failure causes and their follow-ups are periodical. This arises the necessity of interval-censored competing risks data analysis. However, the literature on interval-censored competing risks data is relatively limited although this data setting is ubiquitous in biomedical research.

For current status data under competing risks setting, [Jewell et al. \(2003\)](#) compared Non-Parametric Maximum Likelihood Estimator (NPMLE) of the subdistribution functions with simple parametric models and also ad-hoc non-parametric estimators. [Jewell and Kalbfleisch \(2004\)](#) developed an iterative version of the Pool Adjacent Violators Algorithm for maximum likelihood estimator (MLE) and demonstrate its convergence which were also applied to NPMLE of the sub-distribution functions. The asymptotic theory for the NPMLE with the

current status data under competing risks was established by [Maathuis \(2006\)](#), [Groeneboom et al. \(2008a; 2008b\)](#). [Maathuis and Hudgens \(2011\)](#) further developed the limiting distribution of NPMLE to current status competing risks data with discrete or grouped observation times. They also constructed the point-wise confidence intervals for the sub-distribution functions in the continuous, discrete and grouped models. [Sun and Shen \(2009\)](#) discussed proportional hazards regression model for current status data under competing risks and developed the parametric maximum likelihood estimator (MLE) via a two-step iterative algorithm with separated maximization. They established the consistency and convergence rate for the MLE and also showed that the estimates of regression coefficients are efficient and have asymptotically normal distributions.

For general interval-censored lifetime data, simple imputation or multiple imputation, which transforms interval-censored data into right-censored data and then use methods of right-censored data, is still common in practice. In addition to the typical imputation approach, [Sun et al. \(2015\)](#) proposed two modified partial likelihood estimation procedures for proportional hazards model, which use only the left end point of the observed intervals as imputation and all distinct intervals defined by [Peto \(1973\)](#) to construct the partial likelihood function respectively. A key advantage of their methods is that the estimation of the underlying baseline hazard function is avoided.

[Hudgens et al. \(2001\)](#) derived NPMLE of CIF for interval-censored competing risks data and developed an EM algorithm. They also proposed an alternative pseudo-likelihood estimator to avoid having additional undefined regions. They suggested further research to investigate the theoretical properties (consistency, rates of convergence, and asymptotic distributions) of the two proposed estimators. [Werren \(2011\)](#) developed a naive pseudo-likelihood estimator of CIF for arbitrary interval-censored competing risks data and established the asymptotic theory and point-wise confidence intervals for the CIF.

[Frydman and Liu \(2013\)](#) developed the NPMLE of the cumulative intensities for the interval-censored competing risks data and claimed the estimators are asymptotically unbiased. They modified and extended the method of [Hudgens et al. \(2001\)](#) via the construction of a common supports for all sub-distributions to choose a different version of NPMLE. They obtained the NPMLE of cumulative intensities via an estimating equation method.

Semi-parametric regression models generally gain estimation efficiency and flexibility in contrast to non-parametric methods. Semi-parametric regression on CIF has been studied to evaluate the effects of covariates directly on CIFs. [Ren \(2015\)](#) provided a proportional sub-distribution hazards model and inference procedures for interval-censored data under competing risks. This semi-parametric method is extended from the work of [Heller \(2011\)](#) on estimating equation method for interval-censored data. The proposed estimating equations utilize the rank information of event time pairs and the inverse probability weighting is used to account for the missing mechanism.

[Mao et al. \(2017\)](#) studied a class of semi-parametric regression (including proportional and non-proportional models) for the sub-distribution of a competing risk. For arbitrary interval-censored data with cause of failure information partially missing, they discussed NPMLE and naive estimators while each risk is estimated separately by treating other competing risk(s) as right censored. A EM-type algorithm which extended the self-consistency formula of [Turnbull \(1976\)](#) is provided for estimation. They also established the consistency, asymptotic normality, and semi-parametric efficiency of the NPMLE.

Spline-based sieve maximum likelihood method has been adapted into popular semi-parametric models such as Cox and Fine-Gray models for flexibility while modelling CIF. [Zhang et al. \(2010\)](#) approximated the baseline cumulative hazard function of the Cox model by using a monotone B-spline function for interval-censored data. They established the spline-based sieve semi-parametric maximum likelihood estimation method and studied the asymptotic properties. [Li \(2016\)](#) applied this spline-based semi-parametric maximum likelihood estimation method to the Fine-Gray model for interval-censored competing risks data. The approach is to approximate the log baseline cumulative subdistribution hazard by using a monotone B-spline function. They recommend the spline-based estimator over the NPMLE for reduced dimension, smoothed estimation and fast convergence rate if the degree of spline is properly specified.

To improve efficiency of nonparametric and semiparametric methods, [Jeong and Fine \(2006\)](#) suggested direct parametric modelling of the CIF. [Hudgens et al. \(2014\)](#) extended

direct parametric inference for the CIF to the interval-censored competing risks data. They proposed both maximum likelihood estimator (MLE) and a naive estimator. The full maximum likelihood fits all causes simultaneously but the naive estimator allows separate estimation for each cause. Consequently the naive likelihood is appropriate only under arbitrary interval-censored setting, while the full maximum likelihood is also suitable under an independent inspection setting which allows future examination times to be possibly determined by the historical observation.

However these methods of direct CIF modelling have difficulty to constrain the intrinsic additivity among the predicted probabilities, i.e.  $F_1(t|Z) + \dots + F_k(t|Z) + S(t|Z) = 1$ , where  $Z$  is a vector of predictor variables (Gerds et al., 2012; Shi et al., 2013; Hudgens et al., 2014). Shi et al. (2013) showed through simulations that ignoring the additivity constraint would lead to larger variability and lower coverage rates in prediction of CIFs. Ge and Chen (2012) developed a fully specified parametric model via a subdistribution model for the primary cause and conditional distributions for competing causes under a Bayesian framework. The current methods for direct CIF modelling also have limitations on parameter interpretation or prediction of CIFs. Under right censored competing risks data, Gerds et al. (2012) reviewed those CIF parametric models based on the transformation models with different link functions. They suggested a Multinomial Logistic Regression (MLR) as an alternative logit model, which assures those desirable properties. This direct CIF modelling approach has both the inherent additivity constraint and apparent parameter interpretation. Given limited availability of regression models for competing risks data, especially under interval-censored setting, it has urgent necessity to develop method on direct CIF modelling for interval-censored competing risks data. In the first part of this dissertation we investigate and incorporate the MLR model (Gerds et al., 2012) into the interval-censored competing risks setting.

Parametric modelling of the CIF requires the standard metrics to evaluate and validate the prediction accuracy. In the second part of this dissertation our goal is to develop a model-free prediction accuracy metric in the framework of interval-censored competing risks data. It is important that a prediction accuracy metric is robust from model assumptions



and thus is able to detect the model misspecification. However, prediction accuracy research for the interval-censored data is still limited. Brier score (Brier, 1950) is a popular quadratic scoring rule to measure the performance of event probability predictions, which has been established in the framework of right censored data. Comparing to other misclassification metrics including ROC analysis, Hand (1997) suggested to prefer Brier score since the Brier score is based on the error from predictive probabilities is more relevant to the time-dependent outcome.

The Brier score is a strictly proper scoring rule, which means the score reaches its minimum if and only if the true underlying probability is specified. As such the scoring rule is capable to drive a probability model building towards the true underlying distribution. Gerds and Schumacher (2006) further developed this concept and simplified the term as Prediction Error (PE). Under the consistency established by Gerds and Schumacher (2006) for right censored data, the PE estimator is robust regardless of the specified model and thus can be used as a model-free metric to compare performances of different risk models. Due to different data structures and censoring mechanisms, the theory and method developed under right-censored data can not be directly applied to the interval-censored data without extra care. Thus it is necessary to construct the PE metric to evaluate models for interval-censored competing risks data. Inspired by Graw et al. (2009) and Cortese et al. (2013), we proposed a complete approach to estimating the PE under interval-censored competing risks setting, which is also suitable for interval-censored data without competing risks.

This dissertation is organized as follows. First we introduce the MLR model into the interval-censored competing risks setting in Chapter 2. We designed three simulation studies to investigate the performance of the proposed MLR model, including both correct modelling specification and misspecification settings. In Chapter 3 we propose a prediction accuracy metric in the framework of interval-censored competing risks data. The previous simulation studies in Chapter 2 are proceeded to examine the performance of our proposed method under both correct modelling specification and misspecification settings. We apply our proposed MLR model to a prospective cohort study of cognitive impairment in Section 2.4 and evaluate the performances of the proposed models and existing methods by PE in Section 3.4.

## 2.0 MULTINOMIAL LOGISTIC REGRESSION FOR INTERVAL-CENSORED COMPETING RISKS DATA

### 2.1 INTRODUCTION

The CIF is also termed as the cause-specific sub-distribution function, corresponds the cumulative probability of a cause-specific event under competing risks setting. The CIF is a marginal incidence function in the presence of competing event(s) and characterized the corresponding cumulative incidence over time. The prediction of CIF is of interest in many biomedical disciplines for its apparent interpretation in comparison with the subdistribution hazard. Direct CIF parametric modelling is a very attractive for the competing risks analysis.

However, the current methods for direct CIF modelling have some limitations in practice such as indirect parameter interpretations or prediction of CIF. Under right censored competing risks data, [Gerds et al. \(2012\)](#) reviewed the direct CIF parametric modelling based on the transformation models with different link functions. For example, the Fine-Gray model provides direct relationship between the CIF and covariates, but the interpretation of parameters is complicated, since in the definition of subdistribution hazard, competing events are kept in the risk set along with subjects who have not failed from any events. Similarly for the odds ratios which are based on the logit link function, the complement of CIF includes probability from both alive thus far and those with competing events. Thus, this odds ratio model is not obviously straightforward in practice either. Instead, [Gerds et al. \(2012\)](#) proposed a Multinomial Logistic Regression (MLR) as an alternative logit model. This MLR models the CIF directly and has both the inherent additivity constraint and

apparent parameter interpretations. Given limited availability of regression models for competing risks data, especially under the interval-censored setting, it is desirable to develop direct CIF modelling for interval-censored competing risks data. In this study we investigate and incorporate the MLR model (Gerds et al., 2012) into the interval-censored competing risks setting and extends its flexibility by using the cubic B-spline-based sieve method.

The remainder of this chapter is organized as follows. In Section 2.2, we introduce the MLR model with maximum likelihood inference. The cubic B-Spline-based sieve method is then incorporated and investigated to add flexibility to the proposed MLR model. In Section 2.3, we present three simulation studies to investigate the finite-sample performance of the proposed MLR model under different interval-censored competing risks data settings, including both correct modelling specification and misspecification settings. The proposed methods are then applied to a community-based study of cognitive impairment in aging population in Section 2.4. We conclude this chapter with a discussion in Section 2.5.

## 2.2 METHOD

### 2.2.1 Multinomial Logistic Regression (MLR) for the CIF

Under the competing risks setting, let  $T$  be time to first event, and  $\epsilon$  be the corresponding cause of failure.  $F_k(t|Z) = P_k(T \leq t, \epsilon = k|Z)$  denote the underlying true CIF for the event of cause  $k$  given covariates  $Z$ ,  $k = 1, \dots, K$ . The overall (all-cause) survival function  $S(t|Z) = 1 - \sum_{k=1}^K F_k(t|Z)$ . Let  $A_k(t)$  be a function of time. The Multinomial Logistic Regression (MLR) model proposed by Gerds et al. (2012) is a direct parameterization of the CIF in regression analysis. They proposed the following form for the CIF  $F_k(t|Z)$ :

$$F_k(t|Z) = \frac{\exp(A_k(t) + Z^T \beta_k)}{1 + \sum_{k=1}^K \exp(A_k(t) + Z^T \beta_k)}, \quad (2.2.1)$$

$$S(t|Z) = \frac{1}{1 + \sum_{k=1}^K \exp(A_k(t) + Z^T \beta_k)}, \quad (2.2.2)$$

$$\log \frac{F_k(t|Z)}{S(t|Z)} = A_k(t) + Z^T \beta_k, \quad (2.2.3)$$

where  $A_k(t)$  can be interpreted as the baseline log-odds of type  $k$  events.

This multinomial logistic model has the inherent additivity constraint among the predicted probabilities, i.e.  $F_1(t|Z) + \dots + F_k(t|Z) + S(t|Z) = 1$ . [Hudgens et al. \(2014\)](#) discussed the advantage and disadvantage of unconstrained and constrained likelihood estimation for direct CIF modelling. They suggested to use unconstrained likelihood estimation due to similar efficiency even though the sum of predicted CIFs could be greater than one. In contrast, the MLR model suggested by [Gerds et al. \(2012\)](#) has inherent additivity constraint in the model definition. From Equation 2.2.1, assuming a same function for all  $A_k(t)$  and only two type events  $k = 1, 2$  for simple illustration, the limit for event  $k$  CIF  $F_k(t|Z)$  is,

$$P(\epsilon = k|Z) = \lim_{t \rightarrow \infty} F_k(t|Z) = \frac{\exp(Z^T \beta_k)}{\exp(Z^T \beta_1) + \exp(Z^T \beta_2)}.$$

And one has

$$P(\epsilon = 1|Z) + P(\epsilon = 2|Z) = \sum_{k=1}^2 \lim_{t \rightarrow \infty} F_k(t|Z) = 1.$$

This constrained limit is originated from its definition and thus also inherent in estimation. The MLR model can be estimated by unconstrained likelihood and retains the additivity constraint without imposing additional constraints on the likelihood maximization. Consequently the MLEs have the nice asymptotic properties under regularity.

From the CIF expressed in Equation 2.2.1, the cause-specific hazard rate does not have clearly particular form. It is sensible since we focus on the direct CIF modelling and generally the effects of covariate act differently on the CIF and the cause-specific hazard function ([Gray, 1988](#)). However, from Equation 2.2.2, one can employ the relationship between the survival and hazard functions to get the overall cumulative hazard function,

$$H(t|Z) = -\log [S(t|Z)] = \log \left[ 1 + \sum_{k=1}^K \exp(A_k(t) + Z^T \beta_k) \right]. \quad (2.2.4)$$

From Equation 2.2.3,  $\exp(\beta_k)$  are ratios which are based on the odds between a cause-specific failure and survival from all-cause failure. Thus the exponential parameter can

be clearly interpreted as a cause-specific odds (of particular-failure vs. survival) ratio in practice. In this sense, the proposed MLR model is a proportional odds ratios model.  $A_k(t)$  represents a function to describe how a particular odds ratio on logarithm scale changes over the failure time  $T$ . Based on equation (2.2.4), it is reasonable to restrict  $A_k(t)$  as a non-decreasing function.

The function  $A_k(t)$  can hold linear or nonlinear forms to describe the relationship between odds ratio and time. The nonlinear form of  $A_k(t)$  can substantially extend the model flexibility. In our simulation study Section 2.3.2, a function of time  $A_k(t) = a_0 + a_{11}t + a_{12}t^{\frac{1}{2}}$  provided adequate fit event for the misspecification setting. Royston and Altman (1994) had developed the ordinary polynomial to the fractional polynomials for practical model building. Based on the method of fractional polynomials, the function  $A_k(t)$  has a general form of  $A_k(t) = a_0 + \sum_{h=1}^m a_{kh}t^{p_h}$ , where  $p_h$  is generally from a subset of  $p = \{-3, -2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ . For example,  $A_k(t) = a_0 + a_{11}t + a_{12}t^{\frac{1}{2}} + a_{13}t^{-1} + a_{14}t^{-1} \log(t)$ . Sauerbrei and Royston (1999) and Sauerbrei et al. (2006) extended its implements and offered software to find the best form of fractional polynomials. This fractional polynomial method is of great help to find the function  $A_k(t)$ . However, the Spline is a more robust approach to address a complex relationship regarding time. We incorporate this method to the MLR model in Section 2.2.4.

In our studies, we focused on the arbitrarily interval-censored lifetime data. Without loss of generality, we exemplify the interval-censored data as only the general interval-censored lifetime data (typical case 2 data without right or left censoring) for this study, i.e. time for an event of interest is censored only within two consecutive most-recent follow-ups as some interval  $(L_i, R_i)$ , where  $i = 1, 2, \dots, n$  denote subjects from a random sample with size of  $n$ .

### 2.2.2 Independence of Irrelevant Alternative (IIA) in the MLR model

The definition of multinomial logistic regression in Equations 2.2.1 and 2.2.2 implies the property of Independence of Irrelevant Alternative (IIA). Generally speaking, the relative odds depend on the time function and the events of interest.

$$\log \frac{F_k(t|Z)}{S(t|Z)} = A_k(t) + Z^T \beta_k.$$

Assuming the function of time  $A_k(t)$  is identical for all events, the odds between any pair of events only depend on the two events (for example,  $k_1$  and  $k_2$ ) of interest themselves. That is,

$$\frac{F_{k_1}(t|Z)}{F_{k_2}(t|Z)} = \exp(Z^T \beta_{k_1} - Z^T \beta_{k_2}),$$

or,

$$\log \frac{F_{k_1}(t|Z)}{F_{k_2}(t|Z)} = Z^T (\beta_{k_1} - \beta_{k_2}).$$

The odds between any pair of events are independent of the characteristics of any irrelevant events. Here it is equivalent to the claim that the MLR model is a proportional odds ratios model regarding the risk factor.

However, under the CIF modeling, it is not necessary to assume the independence of any pair of events. IIA doesn't imply independence of alternatives as a prerequisite for the MLR. In general, IIA is actively used in the discrete choice modeling regarding marketing or economic modeling, but not much in the biomedical research, to introduce or modify alternatives (i.e. different types of event).

### 2.2.3 Maximum Likelihood Estimation and Inference in the MLR Model

We use maximum likelihood estimation for the MLR model. For notational simplicity, let  $\theta$  be the parameter in  $A_k(t; \theta)$  for the function of time,  $\Omega = (\beta_1, \beta_2, \dots, \beta_K)$  for the covariate parameter, and  $\Theta = (\theta, \Omega)$ . Let  $\epsilon = 0$  denote censored status at the last followup, i.e. neither of two risk events is observed. Assuming non-informative censoring and only considering two mutually exclusive risk events  $\epsilon = 1, 2$  for simplification, the full likelihood function for interval-censored competing risks data is proportional to

$$\begin{aligned} L(\Theta) &= \prod_{i=1}^n \prod_{k=1}^2 \{F_k(R_i; \Theta; Z_i) - F_k(L_i; \Theta; Z_i)\}^{I(\epsilon=k)} \left\{ 1 - \sum_{k=1}^2 F_k(R_i; \Theta; Z_i) \right\}^{I(\epsilon=0)} \\ &= \prod_{i=1}^n \{F_1(R_i; \Theta; Z_i) - F_1(L_i; \Theta; Z_i)\}^{I(\epsilon=1)} \times \\ &\quad \{F_2(R_i; \Theta; Z_i) - F_2(L_i; \Theta; Z_i)\}^{I(\epsilon=2)} \times \{S(R_i; \Theta; Z_i)\}^{I(\epsilon=0)}, \end{aligned} \tag{2.2.5}$$

where  $I(\cdot)$  is indicator function,  $F_k(T; \Theta; Z)$  and  $S(T; \Theta; Z)$  are as defined in equations (2.2.1) and (2.2.2) respectively. [Hudgens et al. \(2014\)](#) discussed a full likelihood estimation in more details. As pointed out by [Jeong and Fine \(2007\)](#), unlike the cause-specific hazard regression model, using logarithm does not make the likelihood function factored into cause-specific parts individually in the direct CIF modelling. The log-likelihood function is given by

$$\begin{aligned} \log L(\Theta) = \sum_{i=1}^n \left\{ I(\epsilon = 1) \times \log \{F_1(R_i; \Theta; Z_i) - F_1(L_i; \Theta; Z_i)\} \right. \\ \left. + I(\epsilon = 2) \times \log \{F_2(R_i; \Theta; Z_i) - F_2(L_i; \Theta; Z_i)\} \right. \\ \left. + I(\epsilon = 0) \times \log \{S(R_i; \Theta; Z_i)\} \right\}. \end{aligned} \quad (2.2.6)$$

The maximum likelihood estimator (MLE)  $\hat{\Theta}$  can be found from the score function of equation (2.2.6). Taking the negative second derivatives of the log-likelihood function (2.2.6), one has the observed Fisher information matrix. The variance of  $\hat{\Theta}$ , denoted as  $\hat{\Sigma}_{\hat{\Theta}}$  is the average quantity based on the inverse of the observed Fisher information. For statistical inference, the MLE  $\hat{\Theta}$ , has asymptotically Normal distribution with mean  $\Theta$  and variance  $\hat{\Sigma}_{\hat{\Theta}}$ , i.e.,

$$\hat{\Theta} \sim AN(\Theta, \hat{\Sigma}_{\hat{\Theta}}). \quad (2.2.7)$$

Based on the invariance property of the MLE, the MLEs of CIF  $F_k(t; \Theta, z)$  and all-cause survival function  $S(t; \Theta, z)$  are  $F_k(t; \hat{\Theta}, z)$  and  $S(t; \hat{\Theta}, z)$  for given  $z$ , respectively. Using the delta method, the variance of  $F_k(t; \hat{\Theta}, z)$  is

$$\widehat{var}\{(F_k(t; \hat{\Theta}, z))\} = \left( \frac{\partial F_k(t; \Theta, z)}{\partial \Theta} \right) \Big|_{\Theta=\hat{\Theta}} \times \hat{\Sigma}_{\hat{\Theta}} \times \left( \frac{\partial F_k(t; \Theta, z)}{\partial \Theta} \right)^T \Big|_{\Theta=\hat{\Theta}}.$$

Under some regularity conditions,  $F_k(t; \hat{\Theta}, z)$  is asymptotically Normal with mean  $F_k(t; \Theta, z)$ , i.e.,

$$F_k(t; \hat{\Theta}, z) \sim AN \left( F_k(t; \Theta, z), \widehat{var}\{(F_k(t; \hat{\Theta}, z))\} \right). \quad (2.2.8)$$

As pointed out by [Jeong and Fine \(2007\)](#) and [Hudgens et al. \(2014\)](#), a pointwise 95% confidence interval for  $F_k(t; \Theta, z)$  is,  $F_k(t; \hat{\Theta}, z) \pm z_{0.975} \sqrt{\widehat{var}\{(F_k(t; \hat{\Theta}, z))\}}$ . The other practical inference including the Wald, likelihood ratio and Score tests can be derived accordingly. [Hudgens et al. \(2014\)](#) suggested a goodness of fit test for model assessment in direct CIF

modelling. In our study we previously assume all cause-specific CIFs have identical  $A_k(t)$  for simplification. As exemplification for practical inference, we can conduct a goodness of fit test to examine whether all  $A_k(t)$  are identical or not.

One can also derive the inference for the MLE  $S(t; \hat{\Theta}, z)$  similarly,

$$S(t; \hat{\Theta}, z) \sim AN \left( S(t; \Theta, z), \widehat{var}\{S(t; \hat{\Theta}, z)\} \right),$$

where

$$\widehat{var}\{S(t; \hat{\Theta}, z)\} = \left( \frac{\partial S(t; \Theta, z)}{\partial \Theta} \right) \Big|_{\Theta=\hat{\Theta}} \times \hat{\Sigma}_{\hat{\Theta}} \times \left( \frac{\partial S(t; \Theta, z)}{\partial \Theta} \right)^T \Big|_{\Theta=\hat{\Theta}}.$$

#### 2.2.4 Cubic B-Spline-based Sieve MLE

[Geman and Hwang \(1982\)](#) accomplished a sieve maximum likelihood estimation that uses a linear span of basis function in the likelihood as a substitute for unknown functions to approximate the infinite-dimensional parameter estimation. The space of polynomial spline functions has great power to approximate unknown functions ([Schumaker, 2007](#)). Spline-based sieve maximum likelihood method has been adapted into Cox proportional hazard model and Fine-Gray ([1999](#)) subdistribution hazards mode for flexibility while modelling the CIF. Specifically, [Zhang et al. \(2010\)](#) employed a monotone B-spline function to approximate the baseline cumulative hazard function in the Cox model for interval-censored data. They established the spline-based sieve semi-parametric maximum likelihood estimation method and studied the asymptotic properties. [Li \(2016\)](#) applied this spline-based semi-parametric maximum likelihood estimation method to the Fine-Gray model for interval-censored competing risks data. The approach is to approximate the log-scale baseline cumulative subdistribution hazard by using a monotone B-spline function. They recommend the spline-based estimator over the NPMLLE for reduced dimension, smoothed estimation and fast convergence rate if the degree of spline is properly specified. In the direct CIF parametric modeling, usually the relationship between the CIF and the time is complex in practice, which makes the model fitting very difficult. As for the MLR model, the function of time is crucial to predict the CIF. The linear or non-linear function of time  $A_k(t)$  is not flexible enough to address the complexity in practice. On the other hand, the space of polynomial spline functions has



good power and convenient computation in approximation (Schumaker, 2007). Thus it is necessary to represent the complexity of time function with a space of polynomial splines. In this study we incorporate cubic B-Spline-based sieve method into the MLR model to extend its flexibility for modelling interval-censored competing risks data.

The idea of the sieve method is to use a sequence of finite-dimensional parameter sub-space (finite and bounded) to approximate the infinite-dimensional parameter space. Specifically in the MLR model, the time function is assumed from a space of polynomial splines, and it can be well approximated by using a cubic B-Spline. According to the definition of a space of polynomial splines by Schumaker (2007) (Section 4.1, page 108), Zhang et al. (2010) described the space of polynomial splines under interval-censored setting. Here we follow the description of Zhang et al. (2010) for interval-censored competing risks data, and show how the method can be extended to the MLR setting. Suppose the bounds of time points observed from a collection of interval endpoints are  $[\mathfrak{d}_a, \mathfrak{d}_b]$ . Let  $\mathfrak{d}_a = d_0 < d_1 < \dots < d_{\mathcal{K}_n+1} = \mathfrak{d}_b$  be a partition of  $[\mathfrak{d}_a, \mathfrak{d}_b]$  with knots of  $\mathcal{K}_n$ . Denote the partitioned subintervals of  $[\mathfrak{d}_a, \mathfrak{d}_b]$  by  $I_{\mathcal{K}_t} = [d_t, d_{t+1}), t = 0, \dots, \mathcal{K}_n$ , where  $\mathcal{K}_n \approx n^v$  which satisfies  $\max_{1 \leq \mathcal{K} \leq \mathcal{K}_n+1} |d_{\mathcal{K}} - d_{\mathcal{K}-1}| = O(n^{-v})$ . Let the set of knots (partition points) be  $D_n = \{d_1, \dots, d_{\mathcal{K}_n}\}$ , and the order of polynomial is  $m$  (where  $m \geq 1$ ). Denote  $\mathcal{S}_n(D_n, \mathcal{K}_n, m)$  be the space of polynomial splines for the unknown time function in the MLR model. The spline functions  $s$  in  $\mathcal{S}_n(D_n, \mathcal{K}_n, m)$  fulfill the following conditions: (i)  $s$  is a polynomial function with order  $m$  (where  $m \leq \mathcal{K}_n$ ) in the subintervals  $I_{\mathcal{K}_t}$ ; and (ii) the  $m'$  times derivative of  $s$  exists and is continuous on  $[\mathfrak{d}_a, \mathfrak{d}_b]$ , where  $0 \leq m' \leq \max(m-2, 0)$ .

According to Schumaker (2007) (Section 4.2, Corollary 4.10), the space  $\mathcal{S}_n(D_n, \mathcal{K}_n, m)$  has a local basis, called B-Spline,  $\mathfrak{B}_n \equiv \{b_t, 1 \leq t \leq q_n\}$ , where  $q_n \equiv \mathcal{K}_n + m$ . In other words, the local bases or B-Splines can be constructed to approximate the space  $\mathcal{S}_n(D_n, \mathcal{K}_n, m)$ . As for the MLR model, the time function,  $A(t)$  is a non-decreasing function. We can define a monotone sub-space of polynomial spline for the function of time as:

$$\mathcal{M}_n(D_n, \mathcal{K}_n, m) = \{A_n : A_n(t) = \sum_{j=1}^{q_n} \alpha_j b_j(t) \in \mathcal{S}_n(D_n, \mathcal{K}_n, m), \alpha \in \Psi_n, t \in [\mathfrak{d}_a, \mathfrak{d}_b]\},$$

where  $\Psi_n = \{\alpha : \alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_{q_n}\}$ . Therefore the likelihood maximization over the covariate parameter domain  $\Omega$  and the monotone sub-space  $\mathcal{M}_n$  is equivalent to maximizing the likelihood over  $\Omega \times \Psi_n$  simultaneously. In other words, the maximum likelihood estimation and inference in the previous section is directly applied to the MLR model with a B-Spline for the function of time.

The  $i^{\text{th}}$  B-Spline basis functions of order  $j$ ,  $b_{i,j}$ , where  $i = 0, \dots, \mathcal{K}_n + 2m - 1$  and  $j = 0, \dots, m - 1$ , are defined on a recursive formula by [De Boor \(2001\)](#) as follows:

$$b_{i,0}(t) = \begin{cases} 1, & \text{if } d_i \leq t < d_{i+1} \\ 0, & \text{otherwise} \end{cases}.$$

$$b_{i,j+1}(t) = \omega_{i,j+1}(t)b_{i,j}(t) + [1 - \omega_{i+1,j+1}(t)]b_{i+1,j}(t),$$

where

$$\omega_{i,j+1}(t) = \begin{cases} \frac{t-d_i}{d_{i+j}-d_i}, & \text{if } d_{i+j} \neq d_i \\ 0, & \text{otherwise} \end{cases}.$$

The B-Spline is essentially a sequence of piecewise polynomials. The degrees ( $df = m - 1$ ) that controls the overall smoothness of the basis functions are usually recommended at quadratic ( $df = 2$ ) or cubic ( $df = 3$ ). The cubic provides adequate smoothness for regression spline in general ([Schumaker, 2007](#)). Thus we suggest a cubic B-Spline for the time function in the MLR model. Assuming  $\mathcal{K}$  inner knots are chosen, the cubic B-Spline function for the time function is given by

$$A(t) = \sum_{i=1}^{\mathcal{K}+3} \alpha_i b_{i,3}(t), t \in [\mathbf{d}_a, \mathbf{d}_b].$$

To construct a cubic B-Spline in application, it is also necessary to choose the number and position of knots. One can randomly assign both number and position of knots and then select the best by using the comparative Kullback-Leibler risk ([Shen, 1998](#)) or the information criterion for model building. But this approach leads to a great burden of computation. [Stone \(1986\)](#) pointed out the position of knots in a restricted cubic spline is not crucial in regression spline. [Harrell \(2015\)](#) urges to use fixed quantiles or percentiles to benefit from the empirical distribution such as reducing outliers influence. Therefore we

suggest to place equally spaced partition points on observed sample in practice. As for the number of knots, Stone (1986) claimed that 5 or less knots are usually enough for a restricted cubic spline. Rutherford et al. (2015) concluded the number of knots is not critical for the use of restricted cubic spline. Harrell (2015) recommended a selection among three, four and five knots. Use of the Akaike information criterion (AIC) was also suggested to select number of knots for B-Spline function. Regarding the optimal convergence rate, Zhang et al. (2010) and Li (2016) suggested to choose the largest integer below  $\sqrt[3]{N}$  or  $\sqrt[5]{N}$ , where  $N$  is the number of distinct time points observed from a collection of interval endpoints. We advocate to start the model building with these clues to achieve a balance between flexibility and some degree of global smoothness.

In the following simulation studies, a cubic B-Spline is used to approximate the function  $A_k(t)$  in the MLR model under different settings. It is shown that the cubic B-Spline method substantially improved the fitted MLR curves under misspecification settings.

## 2.3 SIMULATION STUDIES

We designed three simulation studies to investigate the performance of the proposed MLR model in different interval-censored competing risks data settings. In the first simulation study, we evaluated the performance of the proposed MLR model under correctly specified modelling. In other words, data samples are generated from a MLR model and then the MLR models are used to fit the data. The other two simulation studies were used to explore performances and properties of the MLR model under misspecification settings, where the MLR models are employed to fit the data simulated from the Fine-Gray model and Weibull distribution, respectively.

### 2.3.1 MLR Model Correctly Specified

In this section we simulated the interval-censored competing risks data based on a MLR model under three different parameter settings. Then the MLR models were fitted for the

simulated data to emulate the scenario of correct specification. In our studies, we assume the function of time,  $A_k(t) = a_0 + a_1 t$  is linear and identical for all functions in equations (2.2.1) and (2.2.2). Without loss of generality, only two type events  $k = 1, 2$  are considered. Let  $U \sim Uniform(0, 1)$ . Following Shi et al. (2013), one can derive the failure time  $T$  from the inverse function of CIF  $F_1(t|Z)$ ,  $T = F_1^{-1}(U)$ .

$$T = \frac{1}{a_1} \left\{ \log \frac{U}{(1-U) \exp(Z^T \beta_1) - U \exp(Z^T \beta_2)} - a_0 \right\},$$

where, the  $U$  has an upper limit which makes the CIF  $F_1(t|z)$  is invertible,

$$U < \frac{\exp(Z^T \beta_1)}{\exp(Z^T \beta_1) + \exp(Z^T \beta_2)}.$$

To simulate the failure time  $T$  from the above inverse CIF function, we first generate random uniform random variable  $U$ , and get failure time  $T$  if  $F_1(t|z)$  is invertible, and set  $\epsilon = 1$ . When  $U$  is out of the upper limit and thus  $F_1(t|z)$  is non-invertible, then we deduce the competing event  $\epsilon = 2$  is observed. Consequently we simulate the failure time  $T$  from the following conditional distribution:

$$P(T \leq t | \epsilon = 2; Z) = \frac{P(T \leq t, \epsilon = 2 | Z)}{P(\epsilon = 2 | Z)}.$$

Similar as for event of type 1, there is a limit for event 2 CIF  $F_2(t|Z)$ ,

$$C = P(\epsilon = 2 | Z) = \lim_{t \rightarrow \infty} F_2(t|Z) = \frac{\exp(Z^T \beta_2)}{\exp(Z^T \beta_1) + \exp(Z^T \beta_2)}.$$

Let  $V \sim Uniform(0, 1)$  and define conditional event 2 failure time  $T_{2.cond} = F_2^{-1}(V | \epsilon = 2; Z)$ . After some algebra, one has the inverse function of the conditional distribution  $P(T \leq t | \epsilon = 2; Z)$ ,

$$T_{2.cond} = \frac{1}{a_1} \left\{ \log \frac{V \times C}{(1 - V \times C) \exp(Z^T \beta_2) - V \times C \times \exp(Z^T \beta_1)} - a_0 \right\}.$$

Thus the event of type 2 failure time  $T_{2.cond}$  can be simulated from the above inverse CIF function after random  $V$  is generated.

We followed the approach proposed by Zhang (2009) and Gmez et al. (2009) to generate non-informative censoring interval regarding the simulated failure times. The algorithm is

first to generate two independent random variables  $U_1$  and  $U_2$  from uniform  $(0, m)$ , where  $m$  is constant. Then the  $i^{th}$  censored interval bounds for failure time  $T_i$  are,  $L_i = \max(T_i - U_{1,i}, T_i + U_{2,i} - m)$  and  $R_i = \min(T_i + U_{2,i}, T_i - U_{1,i} + m)$ .

In the detailed simulation setting, we designed two time-invariant predictors  $Z = (Z_1, Z_2)$ , where  $Z_1 \sim Unif(0, 1)$  and  $Z_2 \sim Bernoulli(p = 0.5)$ . The two corresponding covariate parameters are  $\beta_1 = (\beta_{11}, \beta_{12})$  for CIF  $F_1(t|Z)$ . Similarly  $\beta_2 = (\beta_{21}, \beta_{22})$  is for  $F_2(t|Z)$ . We designed and conducted three different parameter settings for examination. In the first setting, The parameter configurations were  $a_0 = -10, a_1 = 2, \beta_{11} = 0.25, \beta_{12} = 1, \beta_{21} = -0.25, \beta_{22} = 0.85$ . The median width of simulated censoring interval is 5% of the corresponding failure time. The resulting distribution of simulated competing events are 59% cause 1 and 41% cause 2. The second setting is mostly same as the first one except that the median width of simulated censoring interval is 20% of the corresponding failure time. In the third setting, the parameter configurations were  $a_0 = -10, a_1 = 2, \beta_{11} = 0.5, \beta_{12} = 0.25, \beta_{21} = -0.5, \beta_{22} = 0.25$ . The median width of simulated censoring interval is 20% of the corresponding failure time. The resulting distribution of simulated competing events are 62% cause 1 and 38% cause 2. Sample sizes of 200 and 500 were chosen for the simulation. The MLEs of parameters were estimated by SAS/STAT <sup>1</sup>(SAS Institute Inc., 2012). The simulation had repeated 2000 times and the results are summarized here.

In the table 2.1 we show the mean estimates, the model-based standard errors which are averages of the standard errors from the observed Fisher information matrices, the empirical standard errors which are the standard errors from repeated simulations, and the coverage rates of the estimated 95% Wald confidence interval for the true parameters. The parameter estimates and its coverage rate look very promising. The proposed MLR model presented a trend of more accurate parameter estimates as sample size increases. The standard errors had a trend of less variability along with larger sample size. It seems that the model-based standard errors tends little inflated in contrast to the empirical estimates. This phenomenon is not unusually in relative simulation studies, for example, Zeng et al. (2016) and Mao et al.

---

<sup>1</sup>The part of the output for this dissertation was generated using SAS software. Copyright [2012] SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

Table 2.1: MLR Parameter estimation accuracy from simulation.

Data	True Value	Mean Estimate		Modeled Std		Empirical Std		Coverage Rate		
		N=200	N=500	N=200	N=500	N=200	N=500	N=200	N=500	
Setting 1	$a_0$	-10	-10.12	-10.04	0.606	0.384	0.601	0.376	95.8%	96.1%
	$a_1$	2	2.02	2.01	0.120	0.075	0.122	0.074	95.3%	95.8%
	$\beta_{11}$	0.25	0.27	0.26	0.387	0.258	0.268	0.209	96.1%	96.0%
	$\beta_{12}$	1	1.01	1	0.274	0.173	0.268	0.171	95.9%	95.1%
	$\beta_{21}$	-0.25	-0.29	-0.27	0.411	0.275	0.294	0.225	96.2%	95.0%
	$\beta_{22}$	0.8	0.83	0.81	0.296	0.187	0.292	0.187	95.4%	95.2%
Setting 2	$a_0$	-10	-10.09	-9.98	0.640	0.405	0.707	0.541	94.6%	93.4%
	$a_1$	2	2.02	2.01	0.128	0.080	0.129	0.08	95.5%	95.4%
	$\beta_{11}$	0.25	0.27	0.26	0.396	0.266	0.271	0.212	96.7%	95.3%
	$\beta_{12}$	1	1.01	1	0.285	0.180	0.279	0.179	96.2%	95.1%
	$\beta_{21}$	-0.25	-0.29	-0.27	0.419	0.282	0.298	0.227	96.0%	94.7%
	$\beta_{22}$	0.8	0.83	0.81	0.307	0.193	0.302	0.196	95.3%	94.6%
Setting 3	$a_0$	-10	-10.11	-10.01	0.649	0.411	0.697	0.514	94.4%	93.4%
	$a_1$	2	2.03	2.01	0.128	0.080	0.13	0.081	94.9%	95.0%
	$\beta_{11}$	0.5	0.51	0.51	0.434	0.284	0.366	0.27	96.0%	96.4%
	$\beta_{12}$	0.25	0.29	0.26	0.270	0.172	0.229	0.162	97.3%	97.4%
	$\beta_{21}$	-0.5	-0.53	-0.51	0.476	0.312	0.398	0.292	95.7%	96.6%
	$\beta_{22}$	0.25	0.3	0.26	0.298	0.191	0.25	0.177	97.1%	97.4%

(2017). Nonetheless, the inflation is trivial and most likely due to noise from the interval censoring mechanism.

To illustrate the performance of estimated CIFs, we defined two emulated scenarios, "intervention (Int.)" and "control" using the simulated sample. The "intervention" is a setting with the continuous predictor  $Z_1 = 0.5$  and the binary predictor  $Z_2 = 1$ . The "control" is configured at  $Z_1 = 0.5$  and  $Z_2 = 0$ . The estimated and true CIFs for the data setting 2 with a sample size of 200 are plotted in Figures 2.1 and 2.2 for two causes respectively. The estimated CIFs from other settings also have a great match with the corresponding true CIFs (figures not shown here). Overall the simulation result shows the proposed MLR model performs very well for the interval-censored competing risks data.

The cubic B-Spline method is incorporated into the MLR model under this correctly specified setting to evaluate its flexibility for modelling interval-censored competing risks data. Three inner knots are chosen and equally spaced for the cubic B-Spline. The estimated CIFs based on a cubic B-Spline for the data setting 1 with a sample size of 200, are plotted in Figures 2.3 and 2.4 for two causes respectively. A great match is illustrated between the estimated CIFs and the corresponding true CIFs.

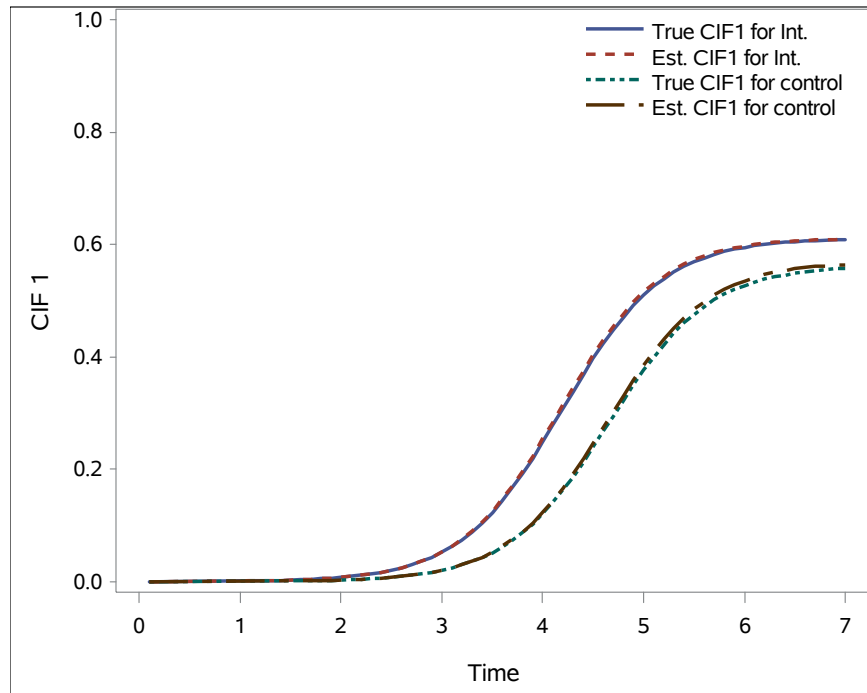


Figure 2.1: Estimated CIFs for cause-1 under correctly specified MLR.

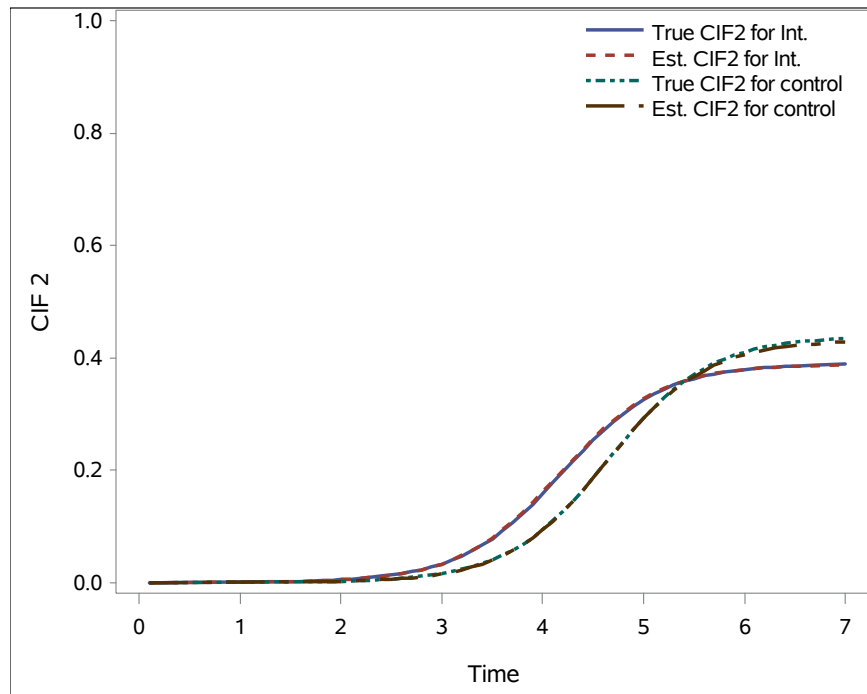


Figure 2.2: Estimated CIFs for cause-2 under correctly specified MLR.

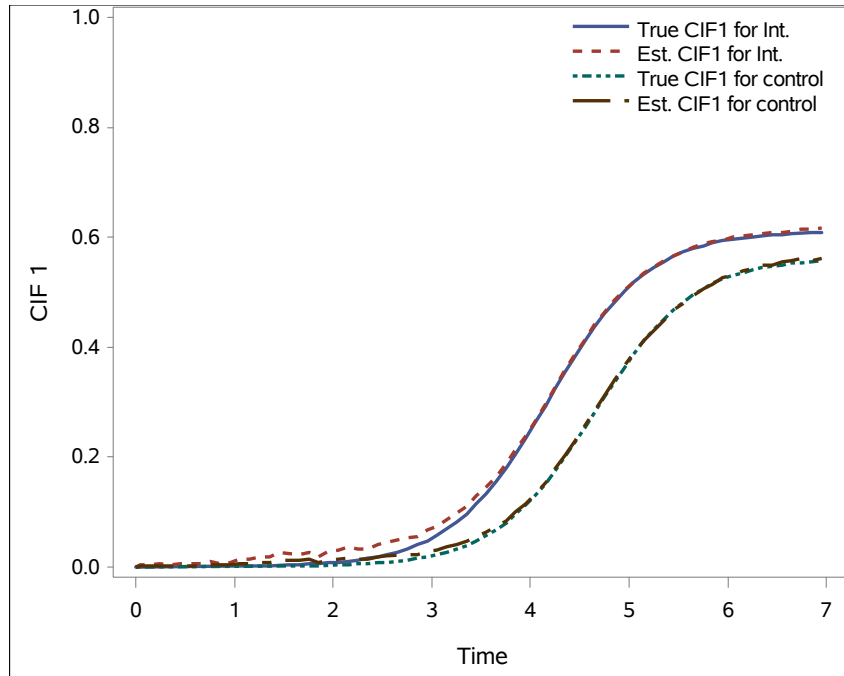


Figure 2.3: CIFs for cause-1 under correctly specified MLR with a cubic B-Spline (knots=3).

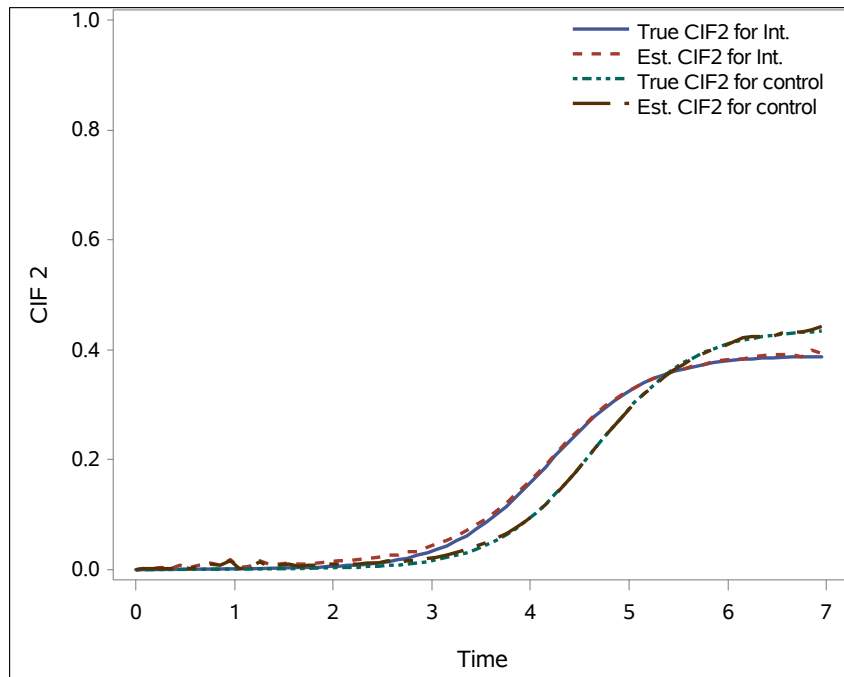


Figure 2.4: Estimated CIFs for cause-2 under correctly specified MLR with a cubic B-Spline (knots=3).



### 2.3.2 MLR Model Misspecified for Fine and Gray Model

In this section we investigate the situation where the MLR model is misspecified for the Fine and Gray model (Fine and Gray, 1999). More specifically, the interval-censored competing risks data was simulated from the Fine and Gray model. We then fitted a MLR model for the simulated data to emulate a misspecified setting.

We followed the approach of Fine and Gray (1999) to simulate the failure time and then generated the censoring interval times by using the same algorithm as in Section 2.3.1. Similar as the setting in Fine and Gray (1999), we designed two time-invariant independent predictors  $Z = (Z_1, Z_2)$ , where  $Z_1 \sim Unif(0, 1)$  and  $Z_2 \sim Bernoulli(p = 0.5)$ . The two corresponding predictor parameters are  $\beta_1 = (\beta_{11}, \beta_{12})$  for cause 1 CIF,  $\beta_2 = (\beta_{21}, \beta_{22})$  are parameters for the cause 2 CIF, and  $p$  is the asymptote of the cause 1 CIF when the covariates are set to zero. The parameter configurations were  $p = 0.3, \beta_{11} = 0.5, \beta_{12} = 0.5, \beta_{21} = -0.5, \beta_{22} = 0.5$  which are also the same as Fine and Gray (1999). The resulting distributions of simulated competing events are 45% cause 1 and 55% cause 2. The median width of the simulated censoring intervals is 5% of the corresponding failure time, while the mean width is 57%. The misspecified MLR model was estimated by the maximum likelihood method as in the previous section. With sample sizes of 200 and 500, the simulations were reiterated 2000 times.

Given misspecified setting and different parameter interpretation between the Fine-Gray and the MLR models, we opt to examine the estimated CIFs instead of the parameter estimates in order to investigate the MLR model assuming a simple linear function of  $A_k(t) = a_0 + a_1t$  for causes. Similar as in previous Section 2.3.1, we defined two emulated scenarios, "intervention (Int.)" and "control" using the simulated sample to illustrate the performance of estimated CIFs. The "intervention" corresponds to the continuous predictor  $Z_1 = 0.5$  and the binary predictor  $Z_2 = 1$ . The "control" is  $Z_1 = 0.5$  and  $Z_2 = 0$ . The estimated CIFs from a sample size of 200 and true CIFs are plotted in Figures 2.5 and 2.6 for two causes of events respectively. As expected, there is an obvious discrepancy between the estimated and true CIFs. The estimated CIFs from other settings present similar patterns in the estimated CIF figures (results not shown). To extend its flexibility, one can specify a polynomial

function of time instead of a linear and identical form,  $A_k(t) = a_0 + a_1t$  for all CIFs. For example, if we choose a function of time  $A_k(t) = a_0 + a_{11}t + a_{12}t^{\frac{1}{2}}$ , the predicted CIFs shows adequately much better fits (see Figures 2.7 and 2.8). Overall the simulation result shows the proposed MLR model needs some extra care to extend its flexibility, and a robust prediction accuracy metric is needed to detect misspecification and guide the model building for the interval-censored competing risks data.

The cubic B-Spline method is incorporated into the MLR model under this misspecification setting to evaluate its flexibility for modelling interval-censored competing risks data. Three inner knots are chosen and equally spaced for the cubic B-Spline. The estimated CIFs based on a cubic B-Spline with a sample size of 200, are plotted in Figures 2.9 and 2.10 for two causes respectively. A substantial improvement is illustrated between the estimated CIFs and the corresponding true CIFs for both competing risks. The cubic B-Spline method also shows better performance than those estimates from the model with the polynomial base function. Overall the simulation result shows the proposed MLR model with the B-Spline baseline is a strong competitor of the Fine and Gray model.

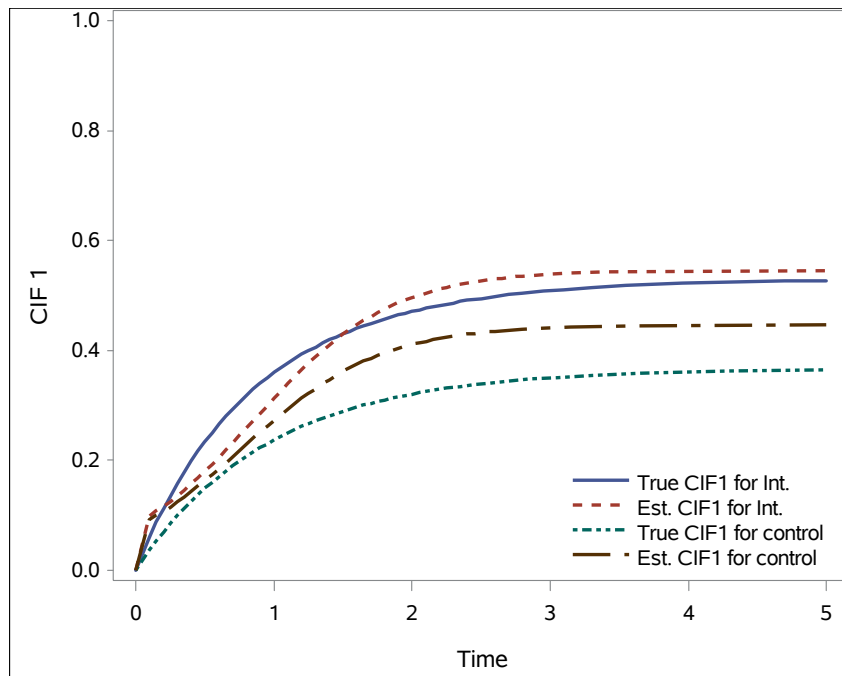


Figure 2.5: Estimated cause-1 CIFs from misspecified MLR with linear base; true CIFs from Fine and Gray.

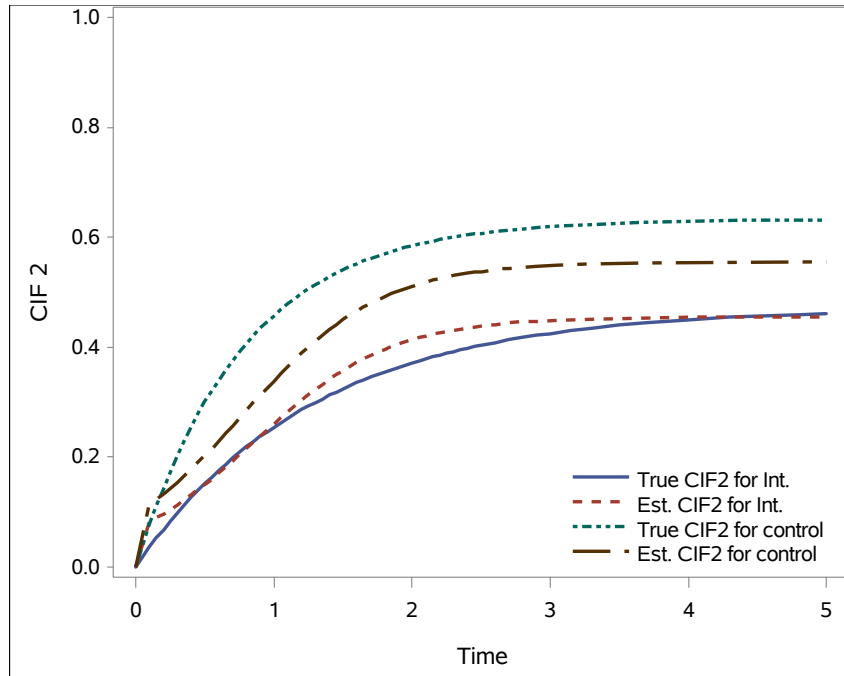


Figure 2.6: Estimated cause-2 CIFs from misspecified MLR with linear base; true CIFs from Fine and Gray.

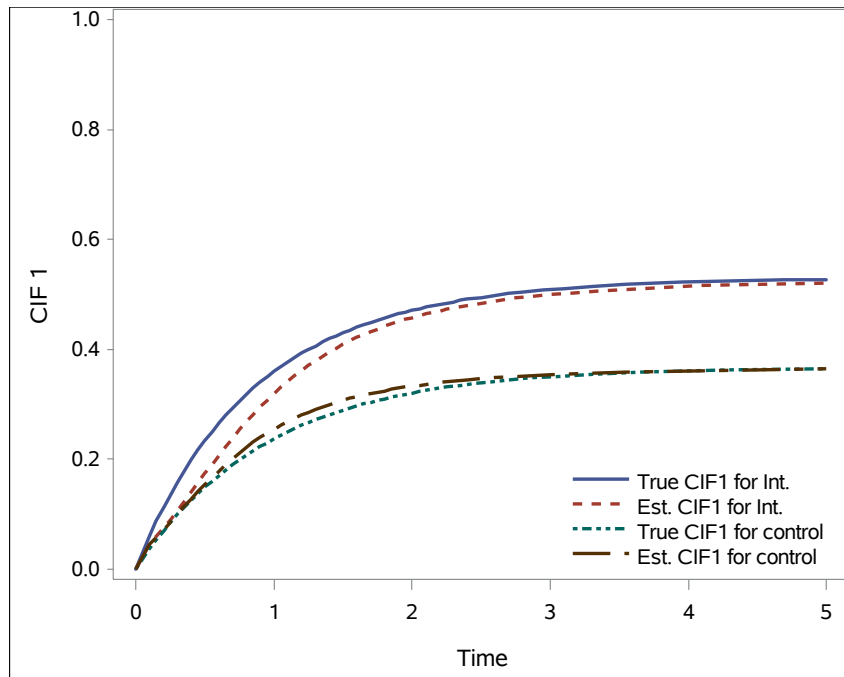


Figure 2.7: Estimated cause-1 CIFs from misspecified MLR with polynomial base; true CIFs from Fine and Gray

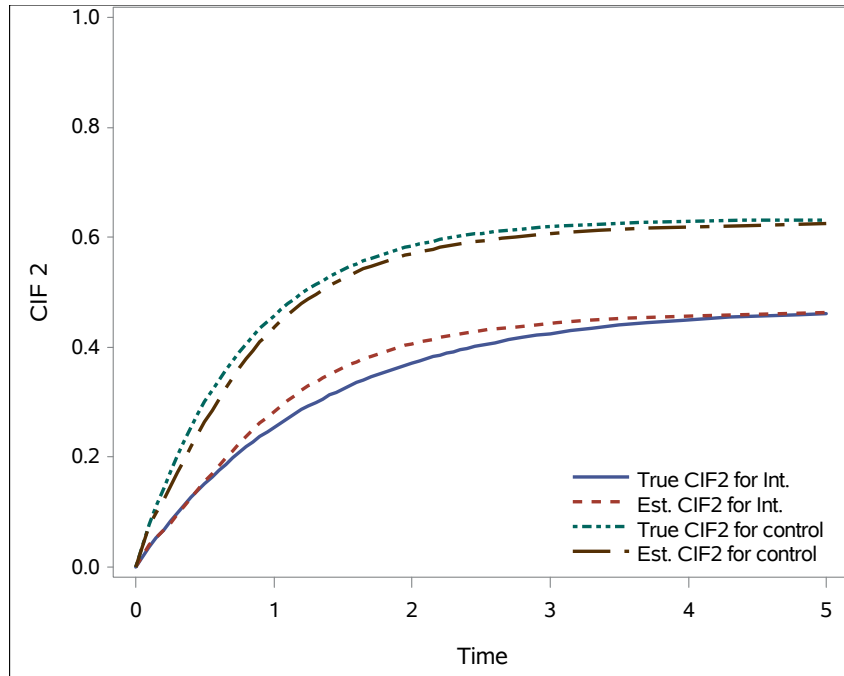


Figure 2.8: Estimated cause-2 CIFs from misspecified MLR with linear base; true CIFs from Fine and Gray

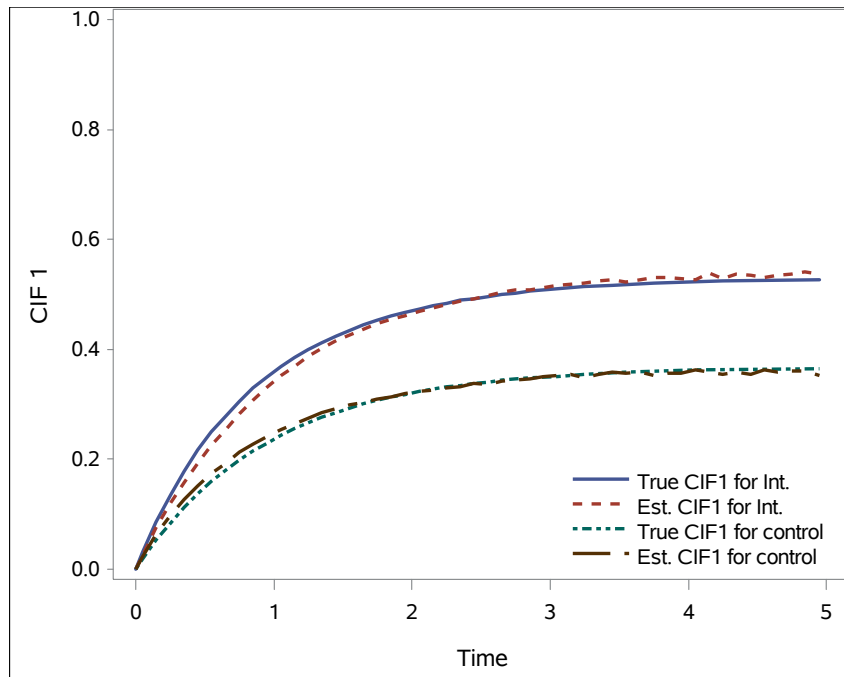


Figure 2.9: Estimated cause-1 CIFs from misspecified MLR with a cubic B-Spline (knots=3); true CIFs from Fine and Gray.

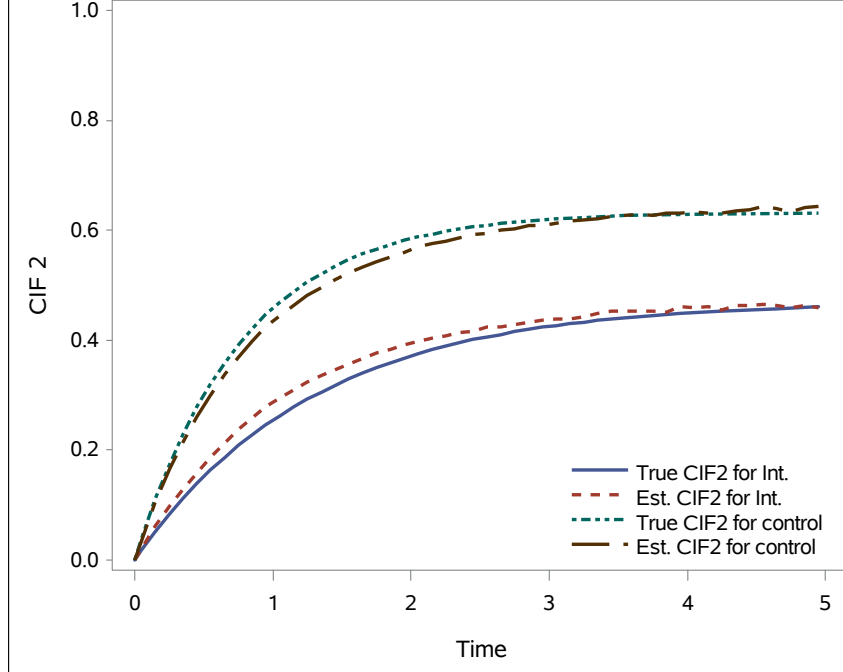


Figure 2.10: Estimated cause-1 CIFs from misspecified MLR with a cubic B-Spline (knots=3); true CIFs from Fine and Gray.

### 2.3.3 MLR Model Misspecified for Weibull Distributed Failure Time

In this section we investigate the situation where the MLR model is misspecified for the popular Weibull distribution. The interval-censored competing risks failure time was originally simulated from a Weibull distribution and then the censoring interval times were generated by using the same algorithm as in Section 2.3.1. The binomial distribution was used to randomly assign occurrence of the event of cause 1 or cause 2. A misspecified setting was emulated by fitting the MLR model with a linear baseline function for  $A_k(t)$  to the simulated data. The basic simulation strategy is same as in previous sections. To adapt the Weibull distribution to the framework of competing risks, we added a constant multiplier  $p_k$  into the Weibull distribution to set the constrained limit  $F_1(t|Z) + F_2(t|Z) + S(t|Z) = 1$  (Cheng, 2009).

$$F_k(t|Z) = p_k \times \left[ 1 - \exp \left( - \left( \frac{t}{b \times \exp(-Z^T \beta_k)} \right)^a \right) \right],$$

where the hazard rate function is,

$$\lambda_k(t|Z) = p_k a b^{-a} (\exp(-Z^T \beta_k))^{-a} t^{a-1}.$$

Failure time and cause indicator were generated from the inverse function of the adapted Weibull distribution, similarly as in the previous section.

$$T = b \times \exp(-Z^T \beta_k) \left[ -\log \left( 1 - \frac{U_w}{p_k} \right) \right]^{\frac{1}{a}},$$

where  $U_w \sim \text{Uniform}(0, p_k)$ .

Again two time-invariant independent predictors  $Z = (Z_1, Z_2)$  were considered, where  $Z_1 \sim \text{Unif}(0, 1)$  and  $Z_2 \sim \text{Bernoulli}(p = 0.5)$ . The two corresponding predictor parameters are  $\beta_1 = (\beta_{11}, \beta_{12})$  for the CIF of cause 1, and  $\beta_2 = (\beta_{21}, \beta_{22})$  are for the cause 2 CIF. The parameter configurations were  $b = 1, p_1 = 0.6, \beta_{11} = 0.5, \beta_{12} = 0.5, p_2 = 0.4, \beta_{21} = -0.5, \beta_{22} = 0.25$ . To extend the generality, we considered two scenarios for different shapes of Weibull distribution,  $a = 0.5$  and  $a = 2$ . The resulting distributions of simulated competing events are about 60% cause 1 and 40% cause 2 for both settings. For the setting with  $a = 0.5$ , the median width of simulated censoring intervals is 10% of the corresponding failure time. For the setting with  $a = 2$ , the simulated censoring interval width is 11% of failure time. The misspecified MLR model was estimated by maximum likelihood method as in previous section. The simulations were reiterated 2000 times with sample sizes of 200 and 500.

Similar as in previous Section 2.3.2, we need to examine the estimated CIFs instead of the parameter estimates in order to investigate the MLR model. We defined two emulated scenarios, “intervention (Int.)” and “control” using the simulated sample to illustrate the performance of estimated CIFs. The “intervention” corresponds to the continuous predictor  $Z_1 = 0.5$  and the binary predictor  $Z_2 = 1$ . The “control” is  $Z_1 = 0.5$  and  $Z_2 = 0$ . The estimated CIFs from a sample size of 200 using the MLR with a linear base are plotted in Figures 2.11 and 2.12, Figures 2.13 and 2.14, along with the true CIFs, for the two settings respectively. There is an apparent discrepancy between the estimated and true CIFs due to the misspecification. One can utilize a polynomial function of time to improve the fitting

as in previous section. The simulation results from a sample size of 500 present similar patterns in the estimated CIF figures (results are not shown). Overall we can draw similar conclusions as in the previous Section 2.3.2. That is to say, the proposed MLR model needs some extra treatment to extend its flexibility. On the other hand, a model-free prediction accuracy metric is needed to detect misspecification and gauge the model selection for the interval-censored competing risks data.

The cubic B-Spline method is incorporated into the MLR model under this misspecification setting to evaluate its flexibility for modelling interval-censored competing risks data. Three or four inner knots are chosen for the cubic B-Spline under two different scenarios respectively. The estimated CIFs based on a cubic B-Spline with a sample size of 200, are plotted in Figures 2.15 and 2.16, Figures 2.17 and 2.18 for the two different shape parameters respectively. A substantial enhancement is illustrated between the estimated CIFs and the corresponding true CIFs, for the Weibull distribution with shape parameter 0.5 or 2. One can conclude the cubic B-Spline method markedly improved the flexibility of the MLR model even under misspecification settings.

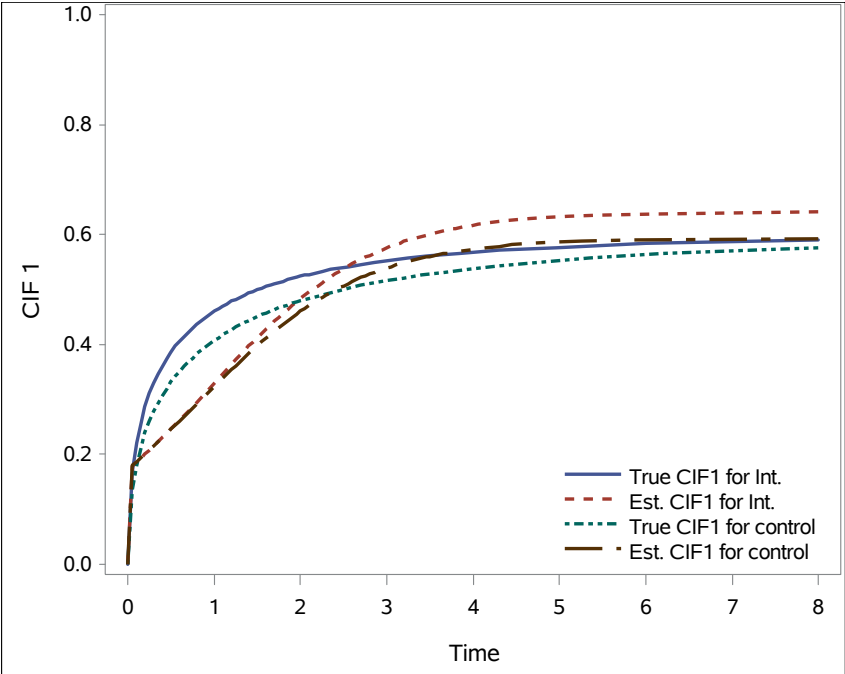


Figure 2.11: Estimated cause-1 CIFs from misspecified MLR with a linear base; true CIFs from Weibull ( $shape = 0.5$ ).

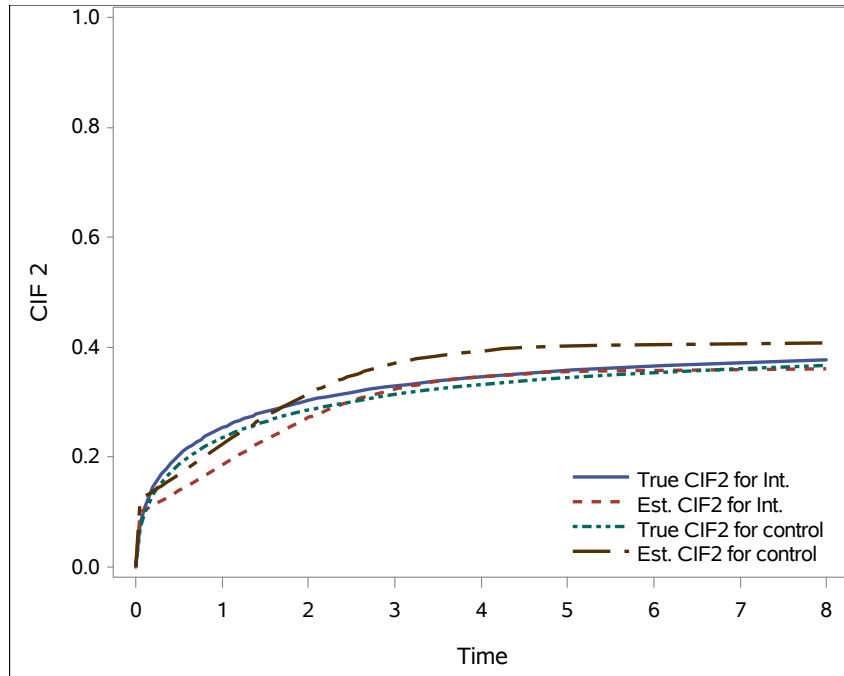


Figure 2.12: Estimated cause-2 CIFs from misspecified MLR with a linear base; true CIFs from Weibull ( $shape = 0.5$ ).

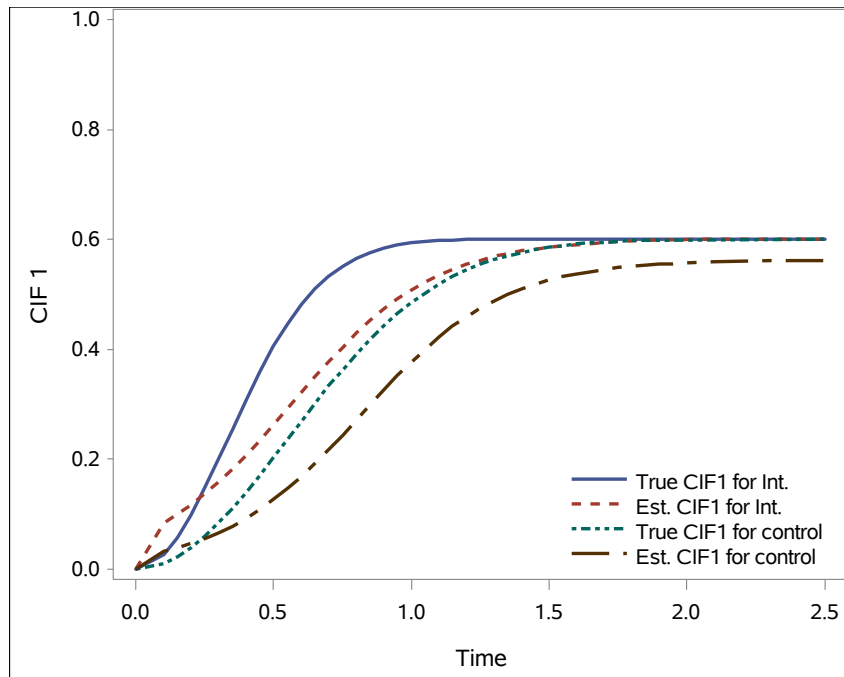


Figure 2.13: Estimated cause-1 CIFs from misspecified MLR with a linear base; true CIFs from Weibull ( $shape = 2$ ).



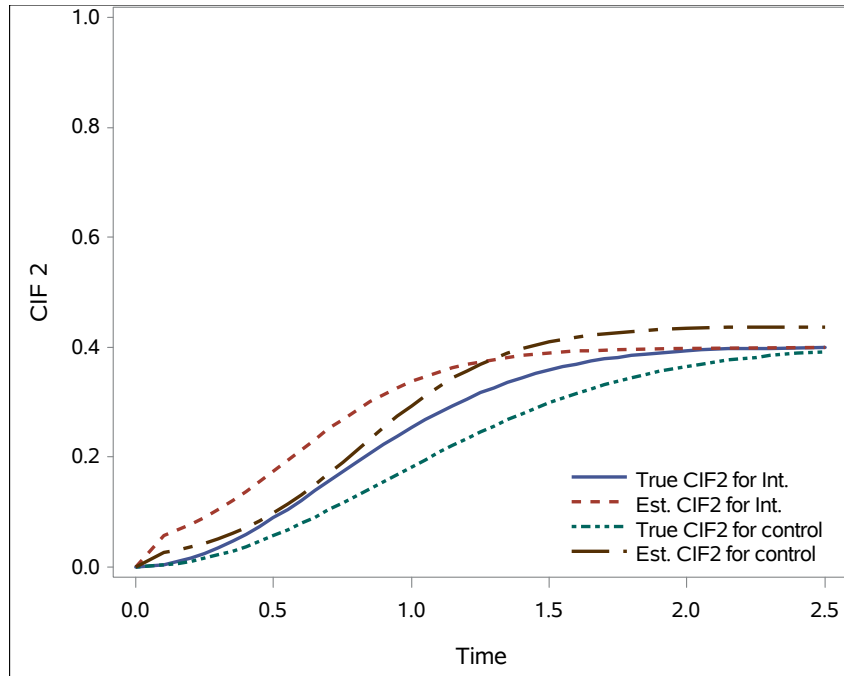


Figure 2.14: Estimated cause-2 CIFs from misspecified MLR with a linear base; true CIFs from Weibull ( $shape = 2$ ).

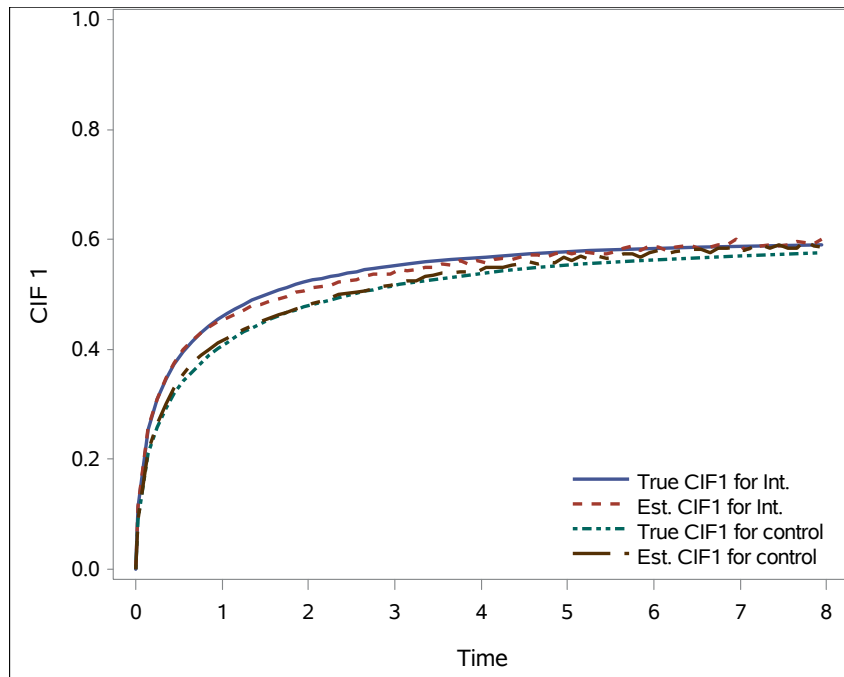


Figure 2.15: Estimated cause-1 CIFs from misspecified MLR with a cubic B-Spline (knots=3); true CIFs from Weibull ( $shape = 0.5$ ).

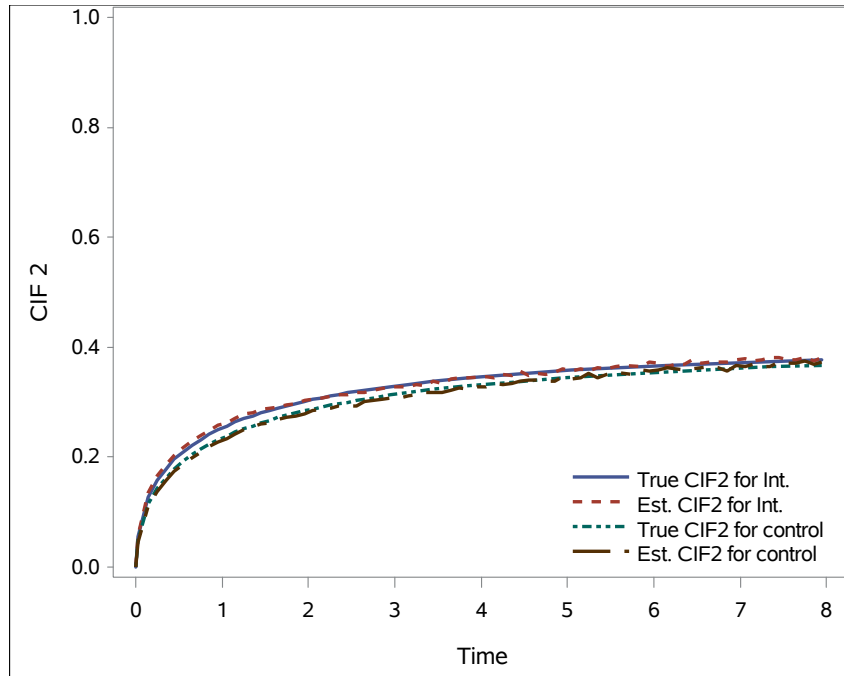


Figure 2.16: Estimated cause-2 CIFs from misspecified MLR with a cubic B-Spline (knots=3); true CIFs from Weibull ( $shape = 0.5$ ).

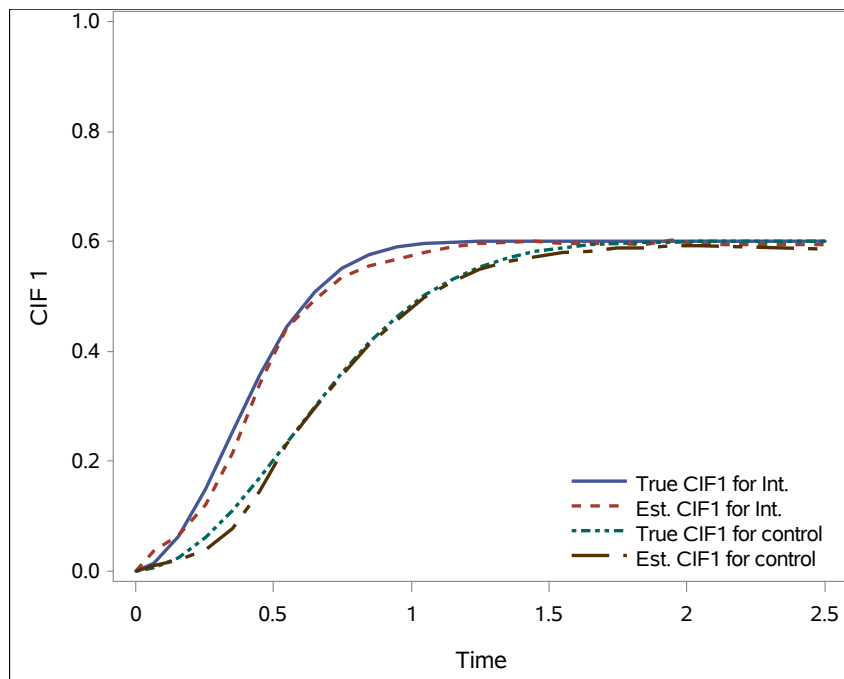


Figure 2.17: Estimated cause-1 CIFs from misspecified MLR with a cubic B-Spline (knots=4); true CIFs from Weibull ( $shape = 2$ ).

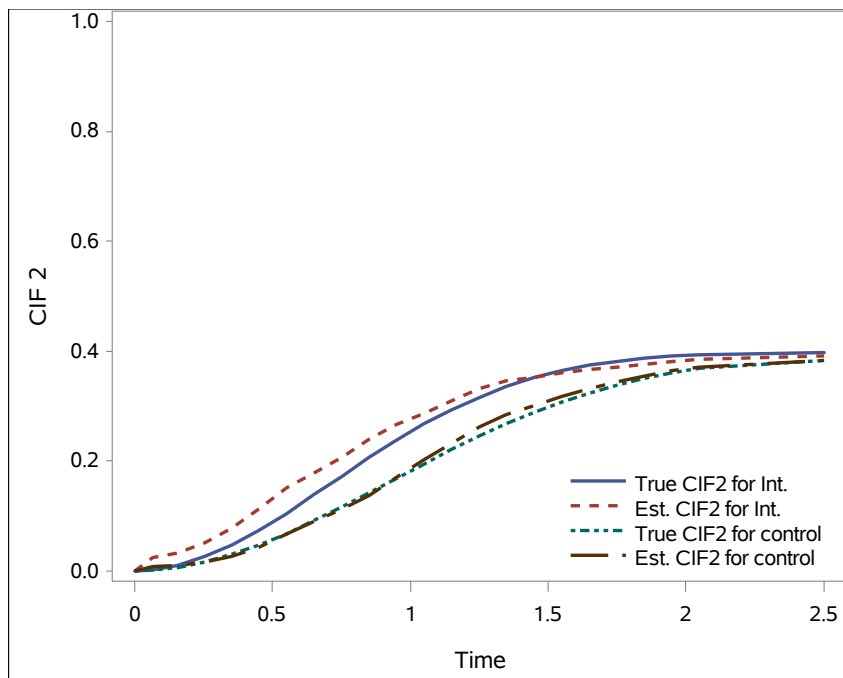


Figure 2.18: Estimated cause-2 CIFs from misspecified MLR with a cubic B-Spline (knots=4); true CIFs from Weibull ( $shape = 2$ ).

## 2.4 APPLICATION TO THE STUDY OF DEMENTIA EPIDEMIOLOGY

The Monongahela-Youghiogheny Healthy Aging Team (MYHAT) conducted a prospective cohort study of cognitive impairment in the small town community of southwestern Pennsylvania (Ganguli et al., 2009). This study was designed to examine the impact of subjective depressive symptoms for performance of cognitive domains. Based on initial screening, a total of 1982 eligible subjects were enrolled to be evaluated annually for 9 times. A summary of characteristics of this data were published by the MYHAT team (Ganguli et al., 2009; Hughes et al., 2015; McDade et al., 2016; Graziane et al., 2016). For this present analysis, we apply our proposed methods to study the effects of depressive symptoms (categorized into 3 groups: low, intermediate and high mCESD) on the cognitive impairment, while adjusting for other baseline covariates including age, gender and education (categorized into three levels:  $< 12$ ,  $= 12$  and  $> 12$  years)(Ganguli et al., 2009). The primary outcome was the progression of the mild cognitive impairment (MCI), which has cognitive dementia rating scale=0.5 or

higher, i.e. including dementia. The time to MCI was defined from the first time of evaluation to the time of MCI. We excluded those with cognitive dementia rating scale greater or equal to 0.5 at the baseline evaluation. The remaining 1282 total subjects were included for this present analysis, of which 27.8% had developed the MCI. The MCI was competing risk censored by death, where 16.7% subjects died without MCI progression. MCI events were only observed at annual evaluations and thus were actually interval-censored within two consecutive visits. Most subjects who died had exact dates of death. However, some of them did not, and thus their death times were also interval-censored. At the end of 9 years follow-up, 55.5% subjects were still alive without progression and their event times were censored at the date of last contact.

Firstly we utilized univariate analyses with a simple linear form of  $A_k(t)$  and observed the depressive symptoms and all baseline covariates, age, gender and education, have significant effects on the MCI. The estimated odds of MCI is 4.1 times higher for subjects with high depressive symptoms as compared to those with low depressive symptoms. In the same univariate analysis of the depressive symptoms, if we used a cubic B-Spline instead of the simple linear  $A_k(t)$  function, the estimated odds ratio of MCI is 2.66 (denoted as model 0 for the latter reference).

Secondly we applied the MLR model with a simple linear  $A_k(t)$  and adjusting for other baseline covariates (model I). It was observed that the high depressive symptoms are potential risk factor for both of the two competing risks after adjusting for other baseline covariates. The estimated odds of MCI is 2.4 times higher for subjects with high depressive symptoms compared to those with low depressive symptoms.

Thirdly, under the same covariates as in model I, a MLR model with a nonlinear, polynomial  $A_k(t) = a_0 + a_{11}t + a_{12}t^{\frac{1}{2}}$ , was fitted. The high depressive symptoms are still a potential risk factor for both of the two competing risks. The estimated odds ratio of MCI between high and low depressive symptoms has a similar value of 2.44 (model I-A). Under the same model using a cubic B-Spline instead of the polynomial  $A_k(t)$  function, the estimated odds ratio of MCI is 2.41 (model I-B). To depict the effect of depressive symptoms on the MCI, we present in Figure 2.19 the CIFs for a 77-year-old female with a high-school education.

The estimated CIFs of MCI for those with high depressive symptoms is obviously different from those with low depressive symptoms.

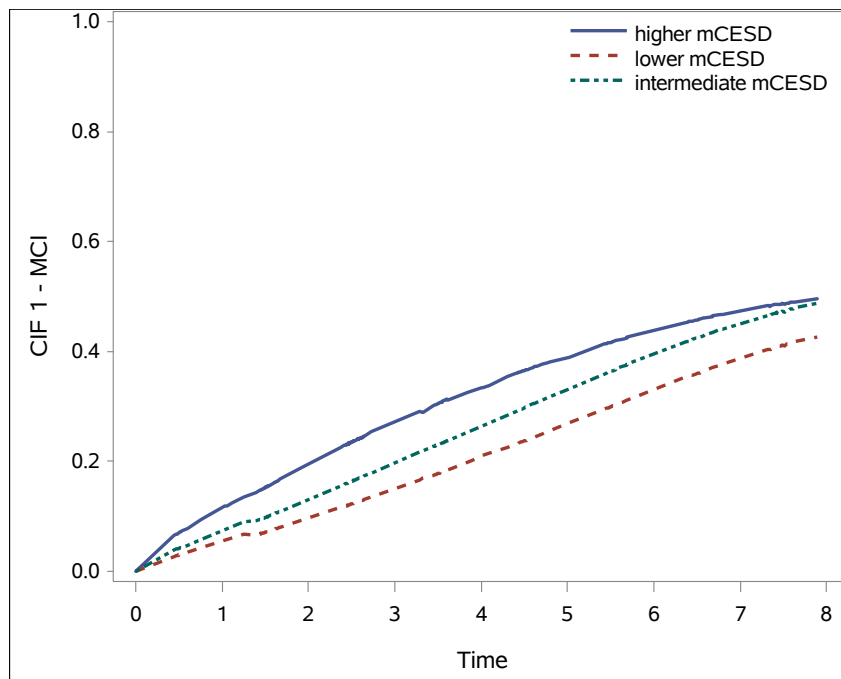


Figure 2.19: Estimated CIFs for the MCI from the MLR model with a cubic B-Spline (knots=6).

To further examine the effects of depressive symptoms on the MCI, we introduced more covariates and have built the model II-A with the polynomial  $A_k(t)$ . The selected covariates include age, gender, education, Florida Cognitive Activities score, IADL score, memory score, self-reported heart attack and stroke (Ganguli et al., 2009; Hughes et al., 2015; McDade et al., 2016; Graziane et al., 2016). The estimated odds of MCI is 2.2 times higher for subjects with high depressive symptoms compared to those with low depressive symptoms. However, the high depressive symptoms are potential risk factor only for the MCI, not for the competing risk event of death. Under the same model using a cubic B-Spline instead of the polynomial  $A_k(t)$  function, the estimated odds ratio of MCI is 2.5 (model II-B). Except for the univariate model using a simple linear form of  $A_k(t)$ , all the estimated odds ratios are closer one another regardless of these models were adjusted by different covariates. It shows the B-Spline method approximated the function of time in the MLR model very well such that the effect of depressive symptom could be derived consistently.

## 2.5 DISCUSSION

In this study we have investigated the MLR model for the direct CIF parametric regression under interval-censored competing risks settings. The maximum likelihood inference has been shown suitable for the proposed model and the simulation studies illustrated desirable finite-sample performance. This direct CIF model assures the additivity constraint and valid estimates on the predicted probabilities, which most popular methods have difficulty to fulfill. Moreover, it provides apparent parameter interpretation which is very appealing in practice, compared to other competitive models which based on the transformation link (e.g. logit) or subdistribution hazard ratios. With these desirable properties, this direct CIF model should be a powerful competitor of the current models.

The interval censoring mechanism and incomplete observation from competing risks are two challenges to analyze interval-censored competing risks data. This proposed method avoids to model the censoring distribution with additional assumptions and takes advantage of maximum likelihood inferences. The estimation of MLR model is based on a jointly maximum likelihood function from all causes, which derive the inference simultaneously for all CIFs. The simultaneous inference is capable to address the relative nature among mutually exclusive competing risks. On the other hand, unlike the cause-specific hazard regression model, the likelihood function cannot be factored into cause-specific components individually in the direct CIF modelling. Consequently, misspecification for any CIF may affect the validity of inferences for other CIFs (Jeong and Fine, 2007). This non-separated structure makes the MLR model more likely to suffer from misspecification. We need a robust prediction accuracy metric which is capable to examine misspecification under the interval-censored competing risks setting.

Our simulation studies show that there are noticeable discrepancies for the MLR model with a linear base function to fit the CIFs that are originated from the Fine-Gray model or Weibull distributions. As illustrated in simulation studies, a polynomial function of time can be used to improve the fit. More attractively, one can use the cubic B-Spline-based sieve method for the proposed MLR model to extend its flexibility, for which the maximum

likelihood inference is straightly applied. In the simulation studies of Section 2.3, it is shown that the cubic B-Spline method substantially improve the fitted MLR curves under misspecification settings. Theoretically it remains a challenge about how to decide the degrees, number and position of knots for a B-spline function. There is a trade off between flexibility and global smoothness. We have provided some suggestions for the model building to achieve a balance between flexibility and some degree of global smoothness.

In our study we assumed that causes of all events are observed and covariates are time-invariant only. However it is not unusually there are missing causes and time-varying explanatory variables in practice. It is a desired research topic for future work.

## 3.0 PREDICTION ACCURACY FOR MODELLING INTERVAL-CENSORED COMPETING RISKS DATA

### 3.1 INTRODUCTION

Parametric modelling of the CIF directly is a very attractive approach in competing risks research and has drawn attentions recently. Such risk probability models need robust metrics to evaluate and validate the prediction accuracy. Brier score ([Brier, 1950](#)) is a popular quadratic scoring rule to measure the performance of event probability predictions. Comparing to other misclassification metrics including the ROC analysis, [Hand \(1997\)](#) suggested that Brier score should be preferred, since the Brier score is based on the error from predictive probabilities and is more relevant to the time-dependent outcome. [Graf et al. \(1999\)](#) defined the mean squared error of probability prediction from a risk model and pointed out its equality to the expected Brier score. Furthermore, [Graf et al. \(1999\)](#) provided the empirical version of the expected Brier score and incorporated this quadratic score rule into right-censored data settings by using the Inverse of Probability of Censoring Weighting (IPCW) scheme. [Gerds and Schumacher \(2006\)](#) established the consistency of the IPCW based estimator which was proposed by [Graf et al. \(1999\)](#) under the right-censored data setting.

The properties of Brier score have been well investigated in the literature ([Selten, 1998](#); [Gneiting and Raftery, 2007](#)). The most important characteristic is that the Brier score is a strictly proper scoring rule, in the sense that the score reaches its minimum if and only if the true underlying probability is specified. As such the scoring rule is capable to drive a probability model building towards the true underlying distribution. Following the



definition of the expected Brier score (Graf et al., 1999) in terms of the mean squared error of probability prediction from a risk model, Gerds and Schumacher (2006) further developed the concept of the mean squared error of prediction, and simplified the term as Prediction Error (PE). Under the consistency established by Gerds and Schumacher (2006), the PE estimator is robust regardless of the specified model and thus can be used as a model-free metric to compare performances of different risk prediction models.

The landmark work of Gerds and Schumacher (2006) established the large sample properties of the PE estimator, and has literally promoted the PE in both theoretical and practical studies. Gerds and Schumacher (2007) further investigated the PE estimator empirically via resampling techniques and confirmed the performance of their proposed estimator. Schoop et al. (2008) developed and extended the PE to measure accuracy with longitudinal covariates for right-censored data. Binder et al. (2009) provided a boosting approach to apply the PE to competing risks data with high-dimensional covariates. Schoop et al. (2011) extended the PE to the framework of competing risks and introduced a consistent PE estimator accordingly. Blanche et al. (2015) proposed the definition and estimator of dynamic PE for joint models under longitudinal right-censored competing risks setting. Other than the approach using IPCW to address censoring distribution, Cortese et al. (2013) proposed to replace the censored event status with a Jackknife pseudo-value in the estimation of the PE under right-censored competing risks. However, to our best knowledge, prediction accuracy research for the interval-censored data is still lacking. It is necessary to construct the PE among models for data of interval-censored competing risks.

In this second part of our studies, the primary goal is to incorporate and investigate the PE as a prediction accuracy metric in the framework of interval-censored data. Inspired by Graw et al. (2009) and Cortese et al. (2013), we propose a novel solution to estimate the PE under the interval-censored competing risks setting, which is also suitable for interval-censored data setting without competing risks. The rest of this chapter is organized as follows. In Section 3.2, the PE is introduced and the method of estimation is provided. Its application on the examination of misspecification is revised. The simulation studies are followed in the Section 3.3. The previous three simulation studies in the last chapter(

Section 2.3) are considered again to examine the performance of our proposed PE metric. The proposed methods are then applied to a community-based study of cognitive impairment in aging population in Section 3.4. We conclude this chapter with discussions in Section 3.5.

## 3.2 METHOD

### 3.2.1 Definition

Let  $T$  be a time-to-event variable and  $F_k^*(t) = P_k(T \leq t, \epsilon = k)$  denote the underlying true margin CIF for the event from cause  $k$ , where we consider  $k = 1, 2$  for simplification. Let  $F_k(t|Z) = P_k(T \leq t, \epsilon = k|Z)$  denote the underlying true CIF for the event of cause  $k$  given covariates  $Z$ . We use  $H(z)$  to denote the marginal distribution of covariates  $Z$ . Let  $I(T \leq s, \epsilon = k)$  denote the event status of type  $k$  at any given time  $s$ , and define its predicted CIF given covariates  $Z$  is  $\hat{\pi}_k(s|Z) = \hat{P}_k(T \leq s, \epsilon = k|Z)$ . As [Graw et al. \(2009\)](#) pointed out, there is an important relationship among these terms:

$$\begin{aligned} F_k^*(t) &= P(T \leq t, \epsilon = k) = E[I(T \leq t, \epsilon = k)] \\ &= E_Z\{E_T[I(T \leq t, \epsilon = k)|Z]\} = E_Z[F_k(t|Z)]. \end{aligned} \tag{3.2.1}$$

[Graf et al. \(1999\)](#), [Gerds and Schumacher \(2006\)](#), [Schoop et al. \(2008\)](#) and [Schoop et al. \(2011\)](#) defined the prediction error (PE) as the expected quadratic loss of prediction:

$$\begin{aligned} PE_k(s) &= E[I(T \leq s, \epsilon = k) - \hat{\pi}_k(s|Z)]^2 \\ &= E_Z\{E_T\{[I(T \leq s, \epsilon = k) - \hat{\pi}_k(s|Z)]^2|Z\}\} \\ &= E\{I(T \leq s, \epsilon = k)[1 - 2\hat{\pi}_k(s|Z)] + \hat{\pi}_k(s|Z)^2\}. \end{aligned} \tag{3.2.2}$$

The prediction error is also termed as the mean squared error of prediction or expected Brier score in the literature. [Graf et al. \(1999\)](#) and [Schoop et al. \(2011\)](#) claimed the prediction

error can be decomposed into two components, “inseparability” (variance) and “imprecision” (bias), which had been previously defined and studied by [Hand \(1997\)](#) regarding Brier score.

$$\begin{aligned}
PE_k(s) &= E [F_k(s|Z)(1 - F_k(s|Z))] + E [F_k(s|Z) - \hat{\pi}_k(s|Z)]^2 \\
&= E [P(T \leq s, \epsilon = k|Z) (1 - P(T \leq s, \epsilon = k|Z))] \\
&\quad + E [P(T \leq s, \epsilon = k|Z) - \hat{\pi}_k(s|Z)]^2
\end{aligned} \tag{3.2.3}$$

The detailed illustration for decomposition is given below ([Graf et al., 1999](#); [Schoop et al., 2011](#)).

$$\begin{aligned}
PE_k(s) &= E [I(T \leq s, \epsilon = k) - F_k(s|Z) + F_k(s|Z) - \hat{\pi}_k(s|Z)]^2 \\
&= E [I(T \leq s, \epsilon = k) - F_k(s|Z)]^2 + E [F_k(s|Z) - \hat{\pi}_k(s|Z)]^2 \\
&\quad + E \{2 [I(T \leq s, \epsilon = k) - F_k(s|Z)] [F_k(s|Z) - \hat{\pi}_k(s|Z)]\} \\
&= E \{[I(T \leq s, \epsilon = k)]^2 + F_k(s|Z)^2 - 2I(T \leq s, \epsilon = k)F_k(s|Z) + E [F_k(s|Z) - \hat{\pi}_k(s|Z)]^2\} \\
&= E [F_k(s|Z)(1 - F_k(s|Z))] + E [F_k(s|Z) - \hat{\pi}_k(s|Z)]^2 \\
&= E [P(T \leq s, \epsilon = k|Z) (1 - P(T \leq s, \epsilon = k|Z))] \\
&\quad + E [P(T \leq s, \epsilon = k|Z) - \hat{\pi}_k(s|Z)]^2.
\end{aligned}$$

Consequently the PE which is defined as the expected quadratic loss function can measure both calibration and discrimination simultaneously. [Schoop et al. \(2011\)](#) claimed it is a strictly proper scoring rule. They also illustrated the bias of PE that was estimated from a misspecified model in simulation studies, while compared to the true underlying PE. [Gerds et al. \(2008\)](#) discussed two theoretical benchmarks, 25% and 33% for the PE as a scoring rule, which can also be derived from the above equation.

[Gerds and Schumacher \(2006\)](#) and [Schoop et al. \(2011\)](#) provided a consistent estimator for the PE while the subjects have identifiable event status at any given time point:

$$\widehat{PE}_k(s) = \frac{1}{n} \sum_{i=1}^n [I(T \leq s, \epsilon = k) - \hat{\pi}_k(s|z_i)]^2,$$

which is equivalent to

$$\widehat{PE}_k(s) = \frac{1}{n} \sum_{i=1}^n \{I(T \leq s, \epsilon = k) [1 - 2\hat{\pi}_k(s|z_i)] + \hat{\pi}_k(s|z_i)^2\}.$$

### 3.2.2 Estimating the Prediction Error

In survival analysis, subject's event status is often unknown or unidentifiable for a period due to censoring. These censored observations are primary difficulty for analysis. In the course of estimating the prediction error for a risk prognostic model, this challenge accompanies both the prediction of risk probabilities and identification of subjects' event status at some given time. For right-censored data, two typical approaches have been developed to address the censoring issue for the estimation of PE. The first approach is to ignore those observations with unidentifiable event status and use the Inverse of Probability of Censoring Weighting (IPCW) scheme to compensate the missingness (Robins, 1993; Graf et al., 1999; Gerds and Schumacher, 2006; Schoop et al., 2011). The second approach is to estimate the unidentifiable event status with a Jackknife pseudo-value (Graw et al., 2009; Cortese et al., 2013). Specifically, for right-censored competing risks data, Cortese et al. (2013) proposed to use a Jackknife pseudo-value as an estimate of the censored and unidentifiable event status. The Jackknife pseudo-values are based on the Aalen-Johnsen estimates of CIFs. Graw et al. (2009) and Cortese et al. (2013) have shown the consistency of the pseudo-value-based estimator  $\hat{J}_k^*(t)$  in right-censored competing risks,

$$E \left\{ \left[ \hat{J}_k^*(t) - I(T \leq s, \epsilon = k) \right] | history \right\} \xrightarrow{n \rightarrow \infty} 0.$$

For interval-censored data, subjects do not have actual event status observed for any time inside of the censoring interval, but the event status is known exactly at any time outside of the interval. The challenge of interval-censored data is that the actual event status which is required in the estimation of PE, is interval-censored and thus unidentifiable within the observed interval. Inspired by the previous work from Graw et al. (2009) and Cortese et al. (2013), a Jackknife pseudo-value can be likewise used as an estimate of censored event status for interval-censored competing risks data. The idea of a pseudo-value estimator provides a general solution to address the unidentifiable event status due to censoring, which avoids specifying the censoring mechanism directly. Thus it is suitable to be adapted to the setting of interval-censored competing risks. The basic strategy is first to construct the Jackknife pseudo-values which are based on the NPMLEs of CIFs. Then one can estimate the PE by

using the Jackknife pseudovalues or the bias-adjusted Jackknife estimator as substitutes for the unidentifiable event status at any given time inside of the observed censoring interval. Let  $\hat{F}_k(s)$  denote the NPMLE of the underlying true CIF for the event of cause  $k$  at a given time  $s$ . Let  $\hat{F}_k^{(i)}(s)$  be the “leave-1-out” NPMLE of the true CIF for the event of cause  $k$  at a given time  $s$ , without using data from the  $i$ -th subject. The Jackknife pseudovalues are defined as

$$\hat{J}_{\epsilon=k}^{(i)}(s) = n\hat{F}_k(s) - (n-1)\hat{F}_k^{(i)}(s).$$

The bias-adjusted Jackknife estimator is

$$\hat{J}_{\epsilon=k}(s) = \frac{1}{n} \sum_{i=1}^n \hat{J}_{\epsilon=k}^{(i)}(s).$$

The consistent pseudo-value-based estimator proposed by [Cortese et al. \(2013\)](#) can be adapted to estimate the PE under the interval-censored competing risks setting, while one can use either the bias-adjusted Jackknife estimator  $\hat{J}_{K=k}$  or the Jackknife pseudovalues  $\hat{J}_{K=k}^{(i)}$ :

$$\widehat{PE}_k(s) = \frac{1}{n} \sum_{i=1}^n \{ \hat{J}_{\epsilon=k}(s) [1 - 2\hat{\pi}_k(s|z_i)] + \hat{\pi}_k(s|z_i)^2 \}, \quad (3.2.4)$$

or,

$$\widehat{PE}_k(s) = \frac{1}{n} \sum_{i=1}^n \{ \hat{J}_{\epsilon=k}^{(i)}(s) [1 - 2\hat{\pi}_k(s|z_i)] + \hat{\pi}_k(s|z_i)^2 \}. \quad (3.2.5)$$

Assuming covariate-independent censoring, we can show the above PE estimator is consistent if the corresponding NPMLEs of CIF are consistent.

$$\begin{aligned} \left| E \left[ \widehat{PE}_k(s) \right] - PE_k(s) \right| &= \left| E \left\{ J_{\epsilon=k}^{(i)}(t) [1 - 2\hat{\pi}_k(s|Z)] - I(T \leq s, \epsilon = k) [1 - 2\hat{\pi}_k(s|Z)] \right\} \right| \\ &= \left| E \left\{ \left[ J_{\epsilon=k}^{(i)}(t) - I(T \leq s, \epsilon = k) \right] [1 - 2\hat{\pi}_k(s|Z)] \right\} \right| \\ &\leq \left| \int_Z \int_T \left\{ \left[ J_{\epsilon=k}^{(i)}(t) - I(T \leq s, \epsilon = k) \right] [1 - 2\hat{\pi}_k(s|Z)] \right\} dF_k(t|z) dH(z) \right| \\ &\leq \left| 3 \int_Z \int_T \left\{ \left[ J_{\epsilon=k}^{(i)}(t) - I(T \leq s, \epsilon = k) \right] \right\} dF_k(t|z) dH(z) \right| \\ &\xrightarrow{n \rightarrow \infty} F_k^*(s) - F_k^*(s) = 0. \end{aligned}$$

### Algorithm to estimate PE for interval-censored competing risks data

Below we provide algorithm by taking the event of cause  $k$  as an example:

- (1) Get the NPMLE of event  $k$  CIF,  $\hat{F}_k(t)$  and  $\hat{F}_k^{(i)}(t)$ , where  $\hat{F}_k^{(i)}(t)$  is the “leave-1-out” estimate of the underlying true CIF for the event of cause  $k$ . [Hudgens et al. \(2001\)](#) proposed a widely referenced NPMLE approach based on closed intervals which were derived from [Peto \(1973\)](#).
- (2) Construct smoothed NPMLEs via Spline or other methods such that the NPMLE is available for any given time.
- (3) Generate Jackknife Pseudovalues and the bias-adjusted Jackknife estimator for a given time  $s$ :

$$\hat{J}_{K=k}^{(i)}(s) = n\hat{F}_k(s) - (n-1)\hat{F}_k^{(i)}(s);$$

$$\hat{J}_{K=k}(s) = \frac{1}{n} \sum_{i=1}^n \hat{J}_{K=k}^{(i)}(s).$$

- (4) Estimating the PE accordingly,

$$\widehat{PE}_k(s) = \frac{1}{n} \sum_{i=1}^n \{ \tilde{I}(T \leq s, \epsilon = k) [1 - 2\hat{\pi}_k(s|z_i)] + \hat{\pi}_k(s|z_i)^2 \},$$

where  $I(T \leq s, \epsilon = k)$  is defined below:

- If the event status is known for a subject at given time  $s$ , then

$$\tilde{I}(T \leq s, \epsilon = k) = I(T \leq s, \epsilon = k);$$

- If the event status for a subject is unidentifiable at given time  $s$ , then we use the corresponding Jackknife pseudovalue or the bias-adjusted Jackknife estimator to estimate the censored event status

$$\tilde{I}(T \leq s, \epsilon = k) = \hat{J}_{\epsilon=k}^{(i)}(s),$$

or,

$$\tilde{I}(T \leq s, \epsilon = k) = \hat{J}_{\epsilon=k}(s).$$

### 3.3 SIMULATION STUDIES

In the previous chapter 2 we designed and conducted three simulation studies in Section 2.3 to investigate the performance of the proposed MLR model. In this section the three simulations are proceeded to examine the performance of our proposed PE metric in different interval-censored competing risks data settings. In the first simulation study, we utilized the proposed MLR model to evaluate the performance of the proposed PE estimator under correctly specified modelling. The other two simulation studies were used to examine the PE's performance and explore its properties under misspecification settings, where the MLR models were employed to fit the data that were simulated from the Fine-Gray model and Weibull distribution respectively.

To avoid over-fitting, we randomly re-assigned the estimated parameters to a simulated data sample to calculate the PE. That is, we first calculated the set of parameter estimates for each of the simulated data samples. Then a set of parameter estimates and a simulated data sample, were randomly matched together to calculate the PE estimates.

To investigate the PE performance under both correct and misspecified models, we defined and presented the following 3 PE-related quantities for the simulation studies.

- (1) The true PE. It is based on the PE definition Equation (3.2.2) with the use of the true event status  $N(t)$  and true CIF. It is denoted as “True N(t) and CIF ...” in the legend of figures.
- (2) The true-status-Est-CIF PE. We calculated this PE by using true event status  $N(t)$  and estimated CIF to evaluate the performance of estimated CIF alone. It is denoted as “True N(t) and Est. CIF ...” in the legend of figures.
- (3) The estimated PE. This is the primary quantity based on our proposed PE estimator from Eq. (3.2.4). The estimated PE was computed using our proposed algorithm which employs event status originated from the bias-adjusted Jackknife estimator. The estimated PE is used to evaluate the performance of both the estimated CIF and the Jackknife estimator together. It is denoted like “JK-avgPS and Est. CIF ...” in the legend of figures.

### 3.3.1 PE from Correctly Specified MLR Model

In this section we continue the previous simulation study in Section 2.3.1 (Page 16), which simulated the interval-censored competing risks data from a MLR model and then the MLR models were fitted to emulate the scenario of correct specification. Proceeded with the estimated CIFs from those fitted models in previous section, we calculated those 3 PE-related quantities according to the beginning introduction in Section 3.3.

Figure 3.1 plots those 3 PE-related quantities together for cause 1 under data setting 1 (Section 2.3.1). We observed all three curves coincided with each other, which is expected for this setting of correctly specified modelling. First of all, the true PE and the estimated PE match perfectly. This match is evident that the proposed PE method did catch this setting of correctly specified MLR modelling. It implied that our proposed PE estimator works for interval-censored competing risks data very well. Secondly, the true-status-Est-CIF PE curve and the estimated PE from our proposed method agreed very well. It produces evidence that the proposed Jackknife pseudo-value as substitute for interval-censored event status, is not only valid but also performs great. Thirdly, the true-status-Est-CIF PE and the true PE curve have identical match. It shows that the proposed PE method is capable to reflect this setting of correctly specified MLR modelling. Meantime, one can observe that the convergence of PE is completely related to the variance component in the decomposition Equation (3.2.3) under a setting of correctly specified modelling. Figure 3.2 shows those three PE curves for cause 2. Consistently we can draw the same conclusion for cause 2 as those for cause 1.

The PE curves from those MLR models with a cubic B-Spline incorporated are in Figures 3.3 and 3.4 for cause 1 and 2 respectively. As stated in the previous Section 2.3.1, three inner knots are chosen and equally spaced to form these cubic B-Splines. Both figures illustrated consistent results and also similar as those PE curves from the models with a linear function of time. It shows not only the capability of the proposed PE methods as prediction accuracy metric, but also additional evidence of a cubic B-Spline performance in the MLR model under this setting of correctly specified modelling. The same conclusions are also supported by results from other data settings (results not shown here).



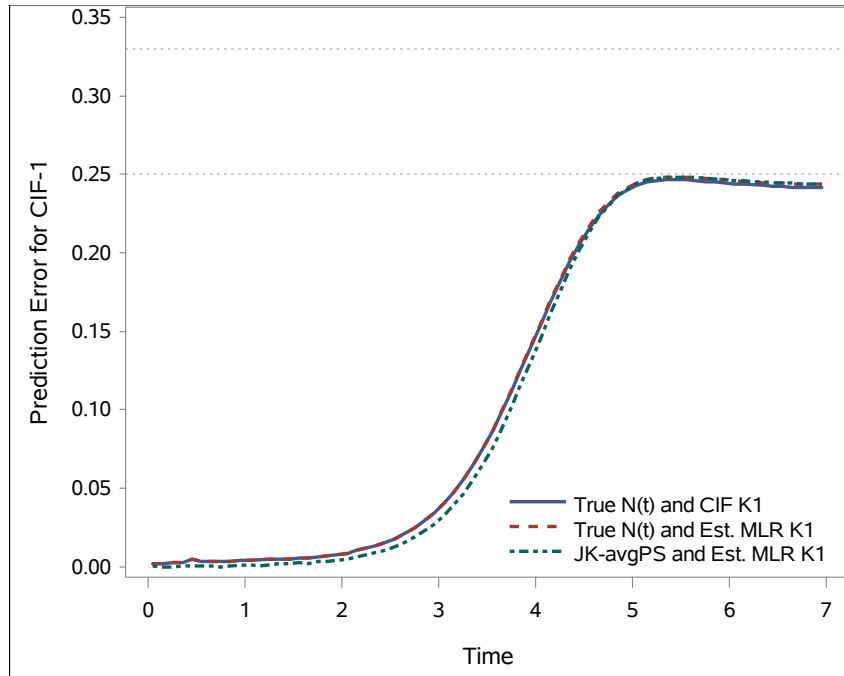


Figure 3.1: PE Estimates For MLR CIF-1.

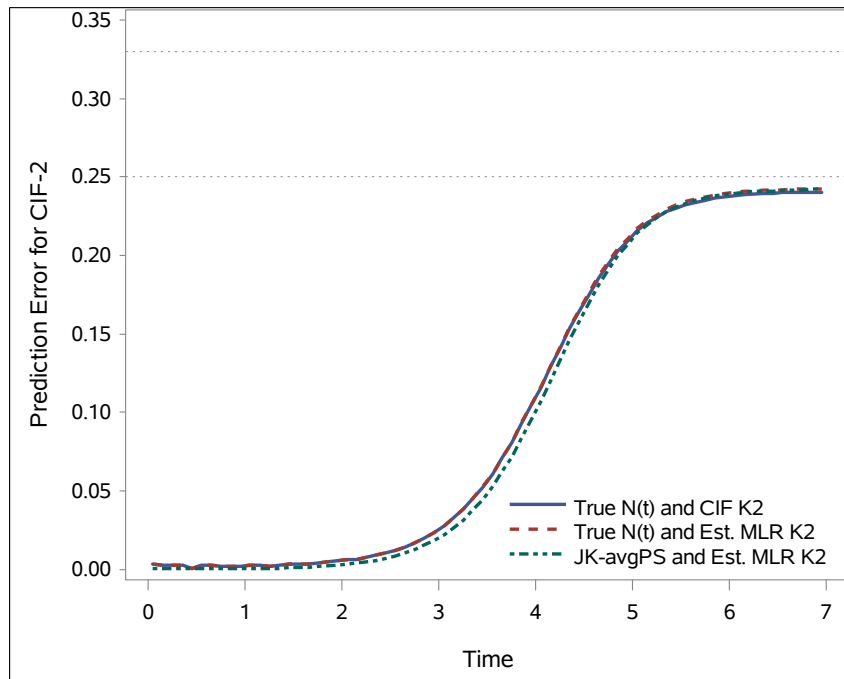


Figure 3.2: PE Estimates For MLR CIF-2.

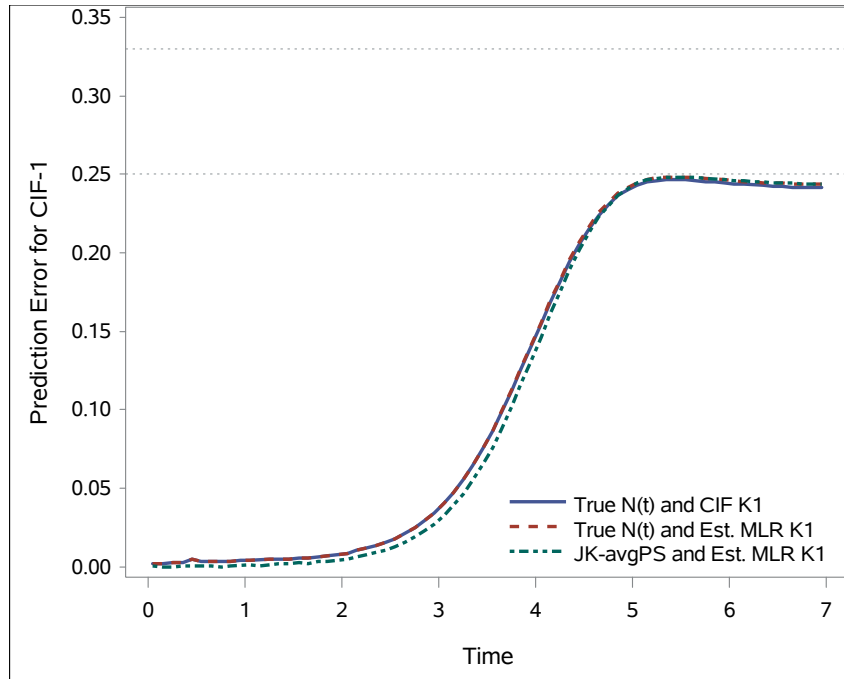


Figure 3.3: PE Estimates For MLR CIF-1 with a cubic B-Spline (knots=3).

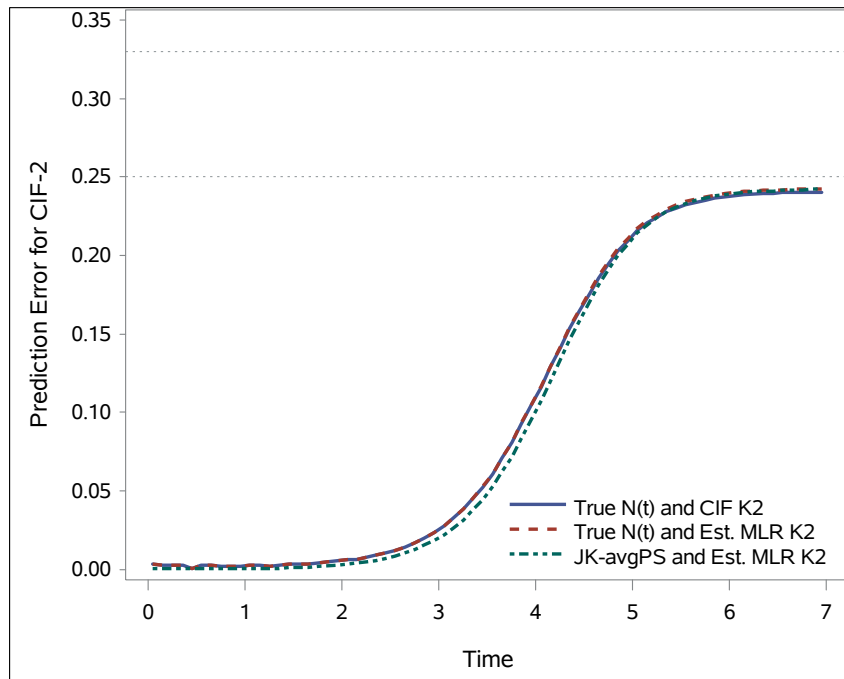


Figure 3.4: PE Estimates For MLR CIF-2 with a cubic B-Spline (knots=3).

### 3.3.2 PE from MLR Misspecified for Fine and Gray Model

In this section we proceed with the previous simulation study in Section 2.3.2 (Page 22), which simulated the interval-censored competing risks data from a Fine and Gray model first, and then the MLR models were fitted for the simulated data to represent the scenario of model misspecification setting. Continued with the estimated CIFs from those misspecified models in the previous section, the 3 PE-related quantities are calculated accordingly and illustrated in figures.

Figure 3.5 plots those 3 PE-related quantities together for cause 1. We observed the PE curves did not agree with each other as those patterns under the correctly specified modelling, which is as expected for this setting of misspecification. First of all, the PE using the true CIF and the estimated PE based on the estimated CIF had obvious discrepancy. It implied that our proposed PE estimator for interval-censored competing risks data is capable of examining modelling misspecification. Secondly, the true-status-Est-CIF PE didn't match the true PE well which shows the estimated CIF has apparent departure from the true CIF. It similarly shows the PE method has capability to examine model misspecification. Thirdly, the true-status-Est-CIF PE and the estimated PE from our proposed method matched very well. It provided extra evidence that our approach by replacing those interval-censored event status with Jackknife pseudovalues, is reliable and also performs well even under this setting of misspecification. Figure 3.6 illustrated those PE curves for cause 2. We can draw the consistent conclusions as those from cause 1.

In the previous Section 2.3.2 (Page 22), we chose a polynomial function of time  $A_k(t) = a_0 + a_{11}t + a_{12}t^{\frac{1}{2}}$  to add the flexibility. From these models with polynomial function of time, the corresponding CIF estimates show adequately better fits (see Figures 2.7 and 2.8). The Figures 3.7 and 3.8 shows the corresponding 3 PE-related quantities together for cause 1 and 2 respectively. It is obvious that these estimated PE quantities are much closer to the true quantities than those estimated quantities from the model without a polynomial term in the time function. To illustrate it clearly, we plotted the Figures 3.9 and 3.10, which includes the PE quantities from both the previous model with a simple linear time function and the one with a polynomial term in the time function together. It is apparently shown that the

estimated PE from the model with the polynomial term is steadily closer to the true PE curve than the one from the previous model with a simple linear time function. It is evident that the proposed PE is able to catch and represent each model fitting improvement among different model reliably.

Figures 3.11 and 3.12 are the PE curves from the misspecified MLR models with a cubic B-Spline incorporated. These figures displays obvious model improvement compared to those previous model with a simple linear time function. It suggests we can draw the same conclusions as previous. It also supports the cubic B-Spline method is very flexible and has best performance. In this case, the model with a polynomial function has very close performance with the cubic B-Spline method. However, compared with the performance of polynomial function and the process to find out such a suitable one, the B-Spline method is more flexible and easily implemented. Our simulation study regarding Weibull distribution in the next section showed an acceptable polynomial function is difficult to be found out for the misspecified MLR model.

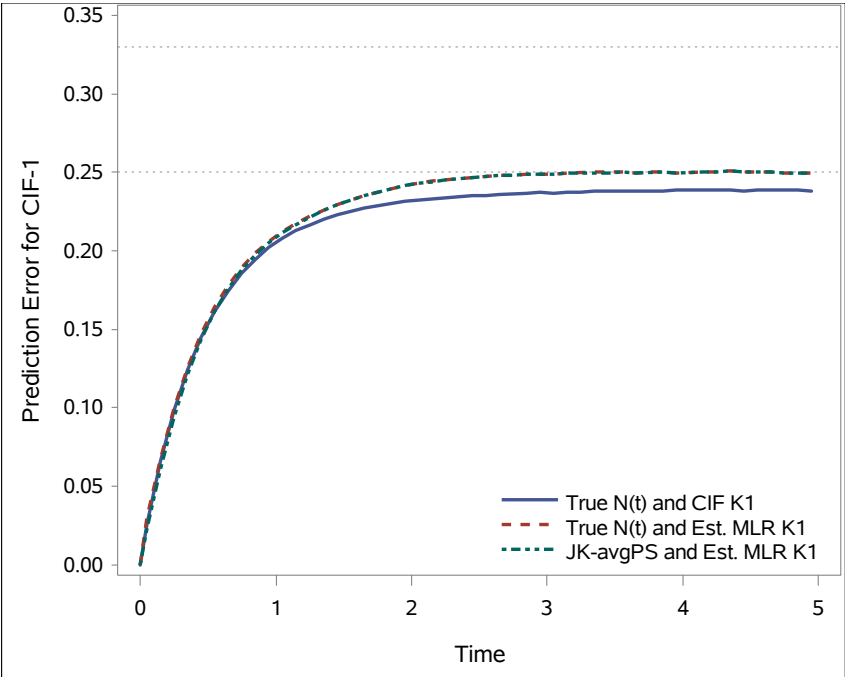


Figure 3.5: PE for the cause-1 CIF based on MLR model with a linear function; true data from the Fine and Gray model.

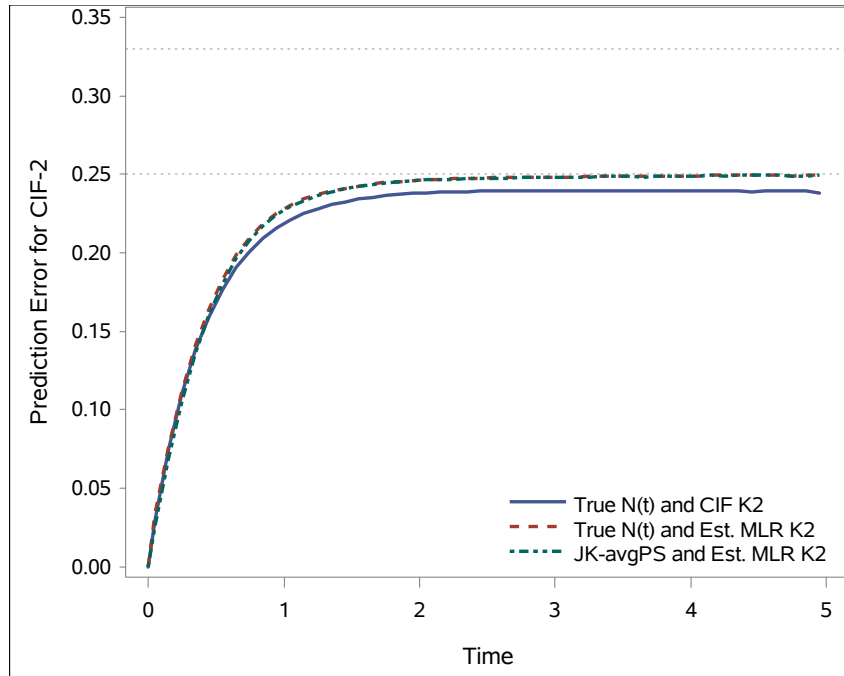


Figure 3.6: PE for the cause-2 CIF based on MLR model with a linear function; true data from the Fine and Gray model.

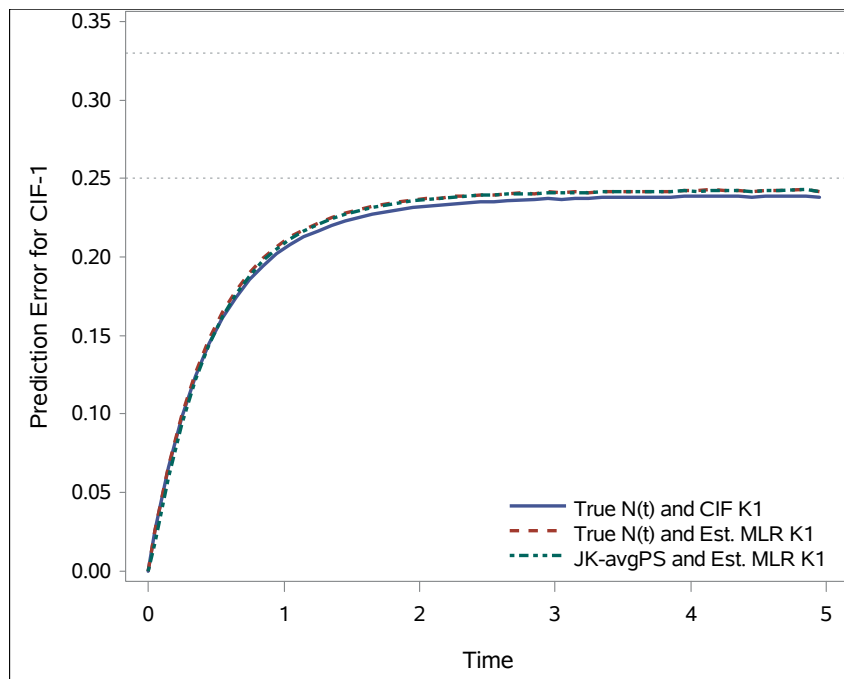


Figure 3.7: PE for the cause-1 CIF based on MLR model with a polynomial function; true data from the Fine and Gray model.

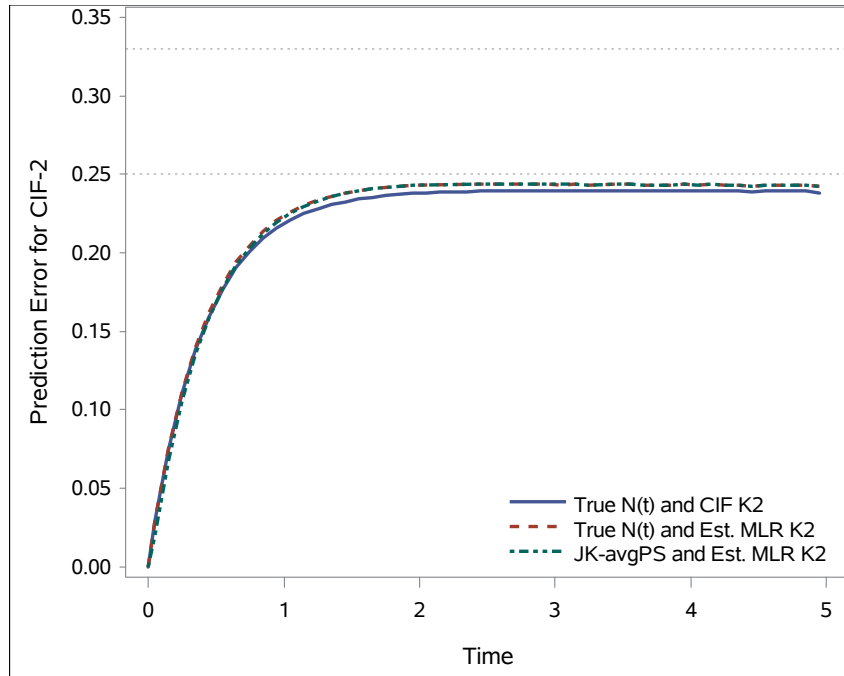


Figure 3.8: PE for the cause-2 CIF based on MLR model with a polynomial function; true data from the Fine and Gray model.

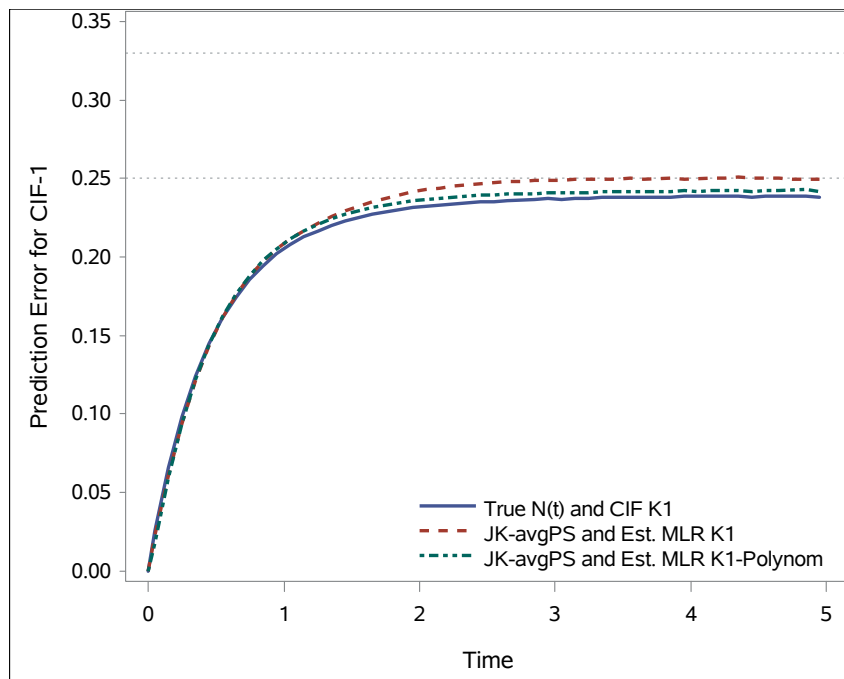


Figure 3.9: PE for the cause-1 CIF based on MLR model with or without a polynomial function; true data from the Fine and Gray model.

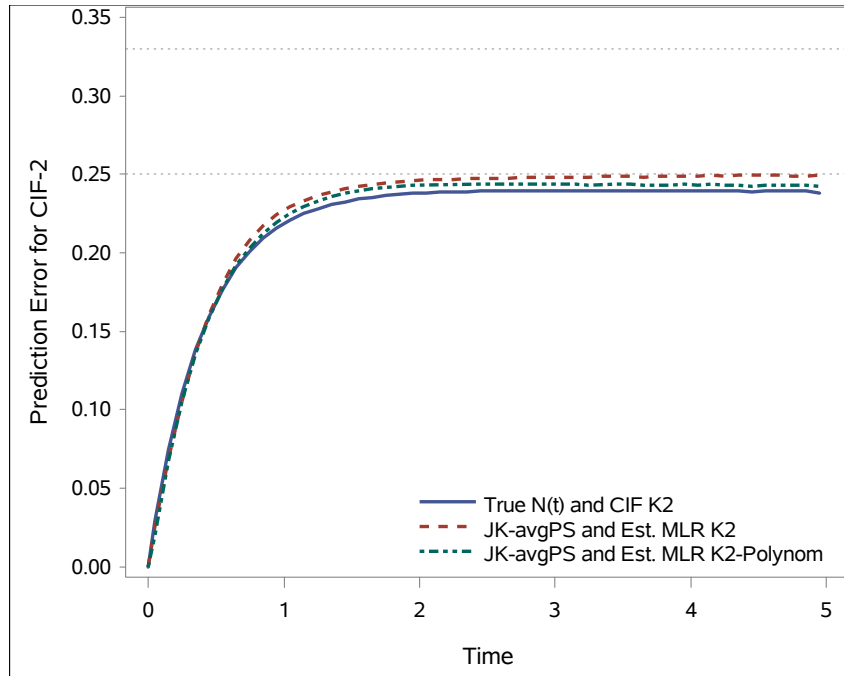


Figure 3.10: PE for the cause-2 CIF based on MLR model with or without a polynomial function; true data from the Fine and Gray model.

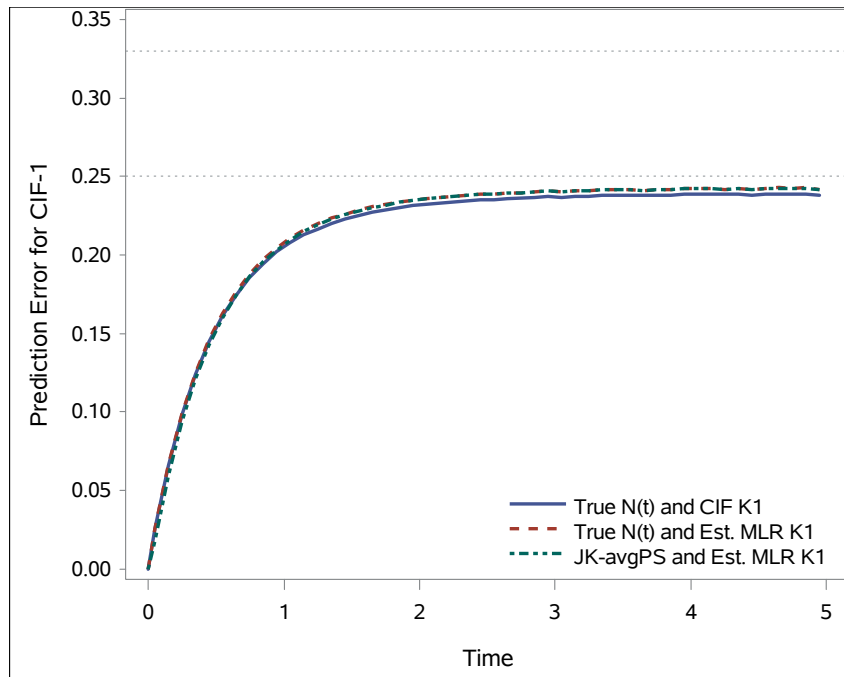


Figure 3.11: PE for the cause-1 CIF based on MLR model with cubic spline; true data from the Fine and Gray model.

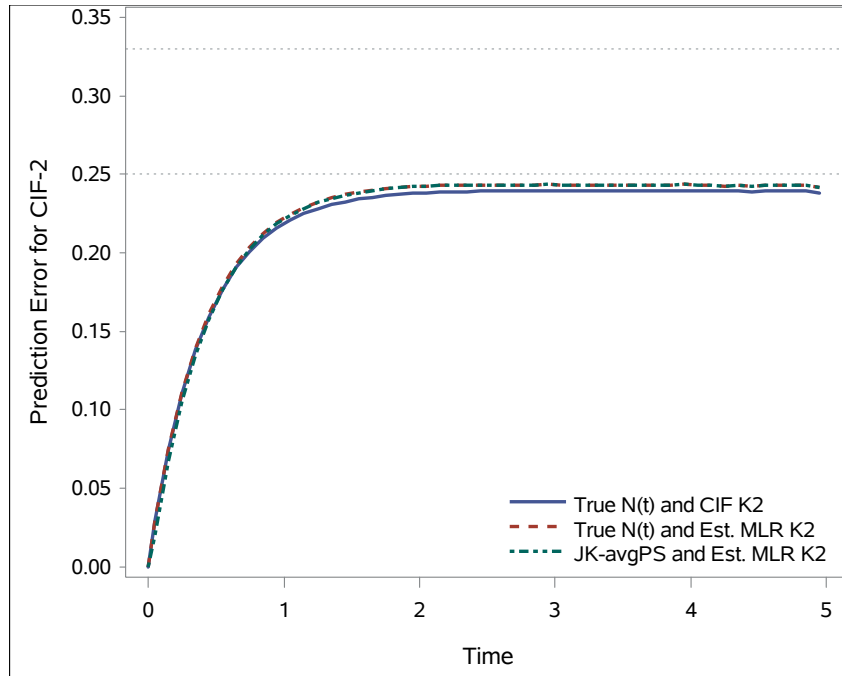


Figure 3.12: PE for the cause-2 CIF based on MLR model with cubic spline; true data from the Fine and Gray model.

### 3.3.3 PE from MLR Misspecified for Weibull Distributed Failure Time

In this section we continued the simulation study in Section 2.3.3 (Page 26), which simulated the interval-censored competing risks failure time from a Weibull distribution, and then the MLR models were used to fit the simulated data to emulate the scenario of misspecification. Continued with the predicted CIFs from those misspecified models, we estimated the PE quantities accordingly.

Figure 3.13 represents those 3 aforementioned PE-related quantities for the event of cause 1 while the failure time were simulated from Weibull distribution with  $shape = 0.5$ . We observed that the true PE curve did not agree with other estimated PE curves, which is as expected under this setting of misspecification. Similar as before, we observed that the true PE and the estimated PE had apparent discrepancy. It suggests that our proposed PE estimator for interval-censored competing risks data is capable of detecting model misspecification. The same results is shown from the departure between the true PE and the true-status-Est-CIF PE curves. The true-status-Est-CIF PE and the estimated PE from



our proposal match very well. It provides extra support that our approach by replacing those interval-censored event status with Jackknife pseudovalues, is reliable and functions consistently. For the event of cause 2, Figure 3.14 suggests to draw the same conclusions as these from cause 1.

The corresponding PE estimates from those MLR models with a cubic B-Spline incorporated are plotted in Figures 3.15 and 3.16 for the event of cause 1 and 2 respectively. These figures illustrates the estimated PE curves are much closer to the true PE curves than those from the models without a B-Spline. Thus it shows that the PE is able to detect model improvement and supports to draw consistent conclusions as those in the previous.

Figures 3.17 and 3.18 plot those 3 PE-related quantities from simulation using the Weibull distribution with  $shape = 2$ . The corresponding PE curves from the MLR models with a cubic B-Spline incorporated are in Figures 3.19 and 3.20. As discussed previously and illustrated here, the PE curves captured the departure and improvement of CIFs estimated from different models. We can conclude the same as the corresponding ones in the previous part with  $shape = 0.5$ .

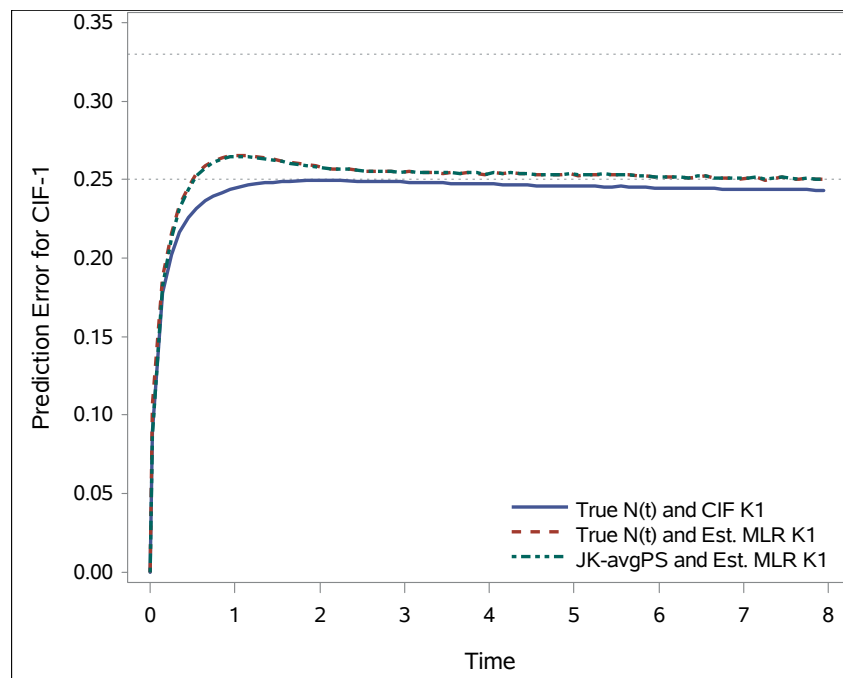


Figure 3.13: PE for the cause-1 CIF based on MLR model with a linear function; true data from Weibull distributed ( $shape = 0.5$ ).

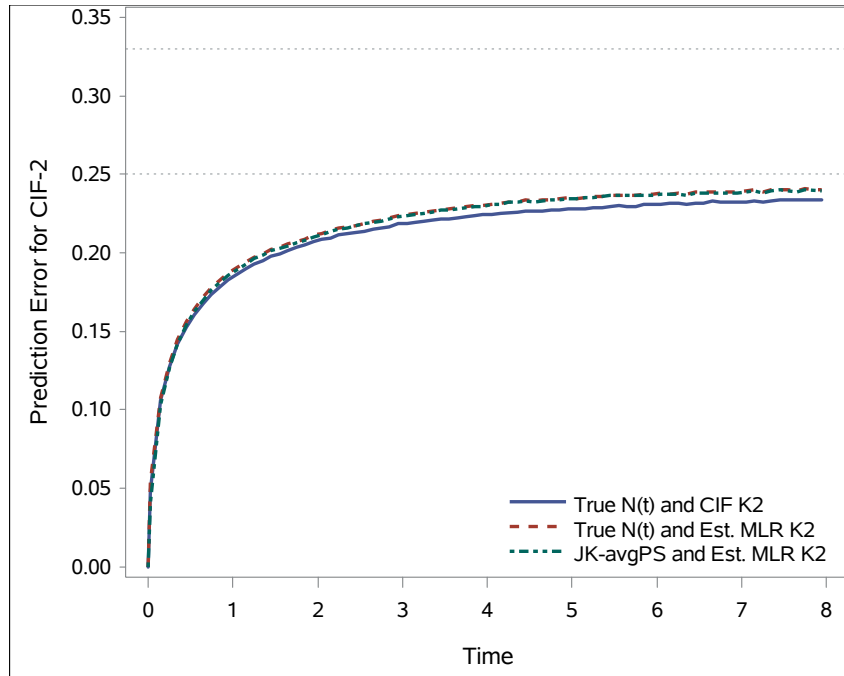


Figure 3.14: PE for the cause-2 CIF based on MLR model with a linear function; true data from Weibull distributed ( $shape = 0.5$ ).

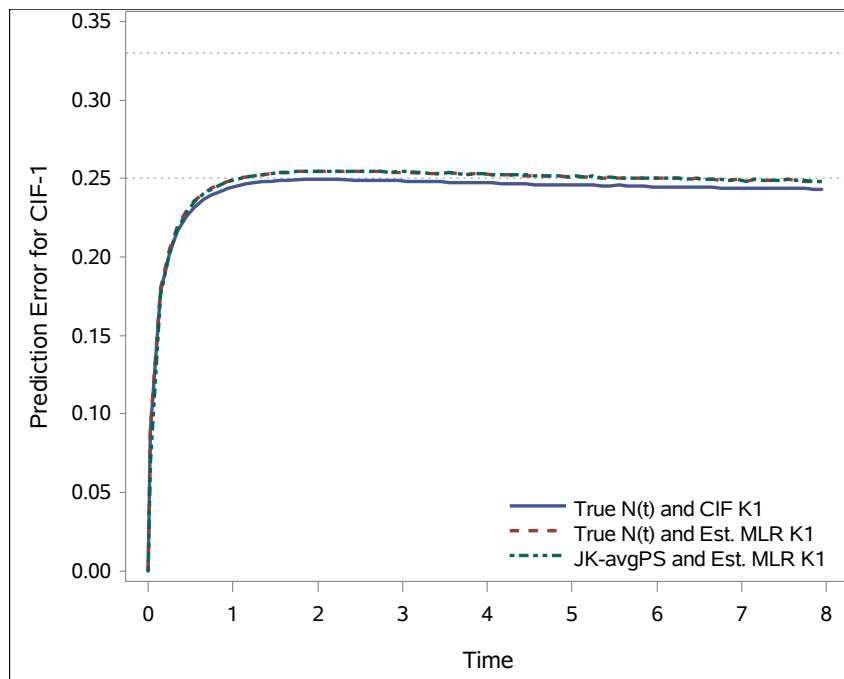


Figure 3.15: PE for the cause-1 CIF based on MLR model with a cubic B-spline; true data from Weibull distributed ( $shape = 0.5$ ).

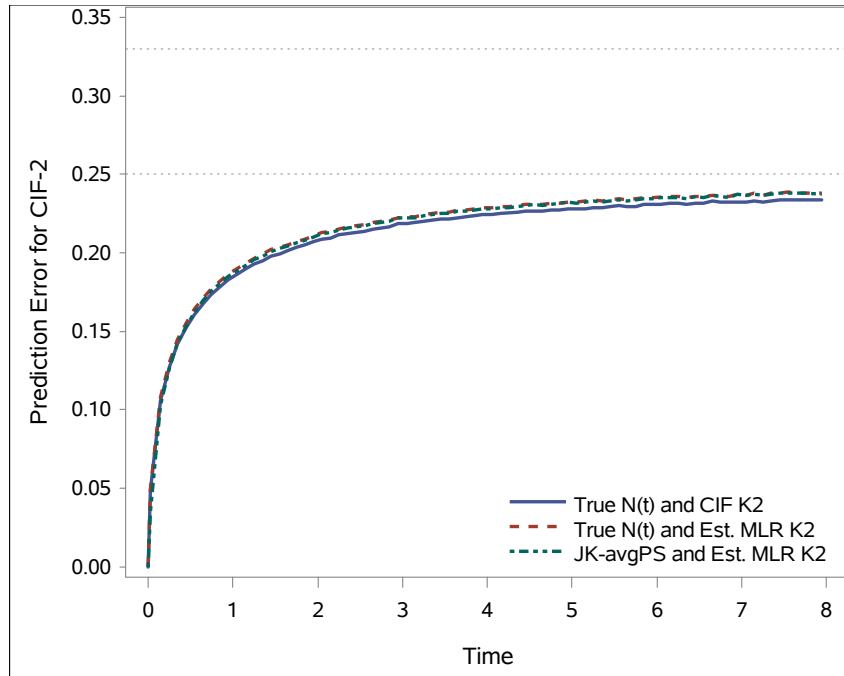


Figure 3.16: PE for the cause-2 CIF based on MLR model with a cubic B-spline; true data from Weibull distributed ( $shape = 0.5$ ).

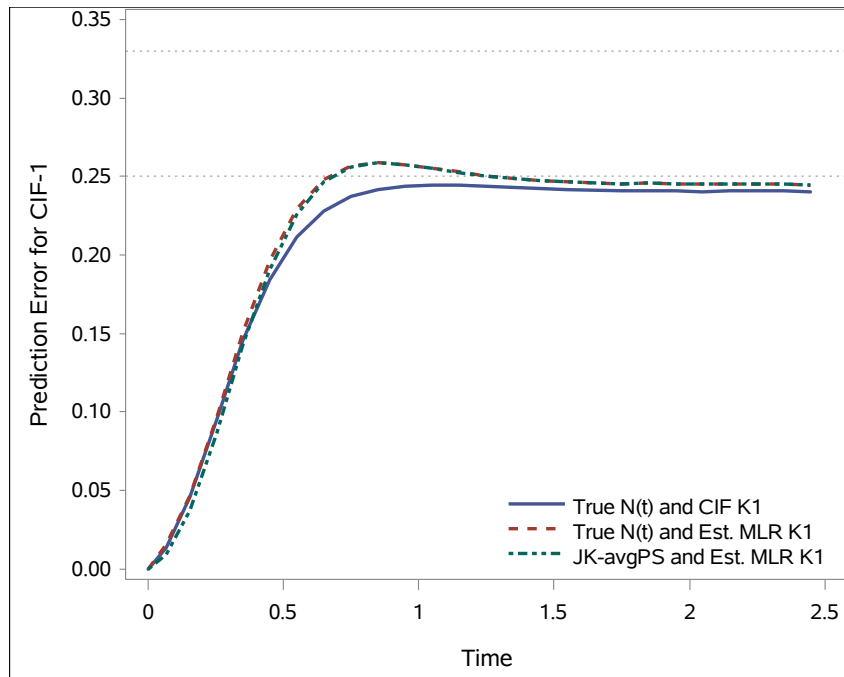


Figure 3.17: PE for the cause-1 CIF based on MLR model with a cubic B-spline; true data from Weibull distributed ( $shape = 2$ ).

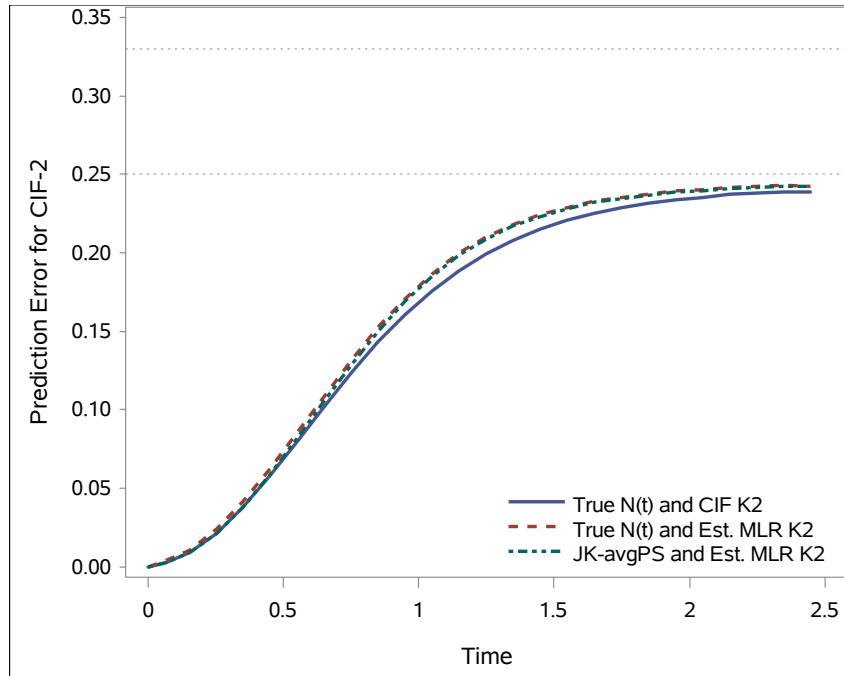


Figure 3.18: PE for the cause-2 CIF based on MLR model with a cubic B-spline; true data from Weibull distributed ( $shape = 2$ ).

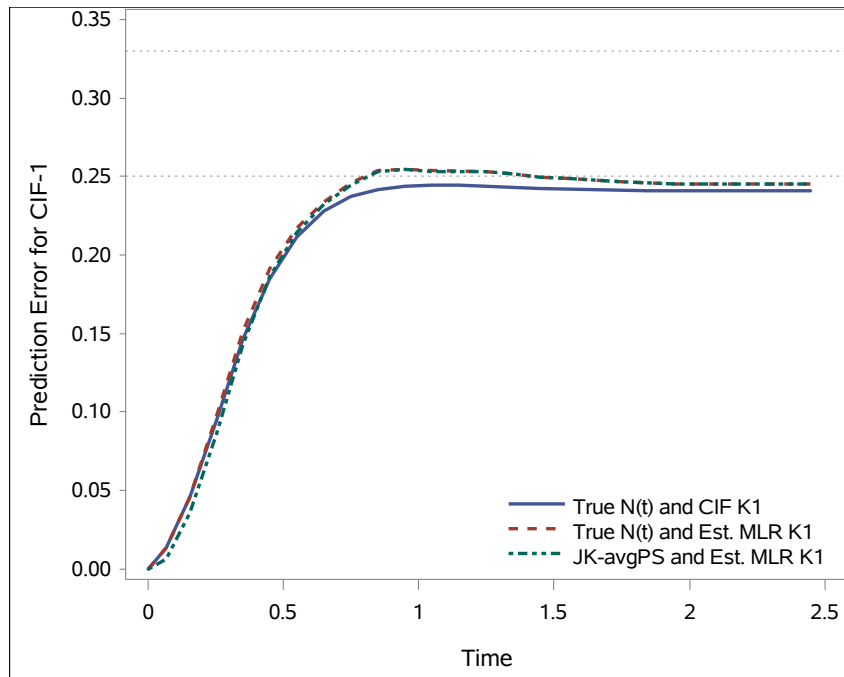


Figure 3.19: PE for the cause-1 CIF based on MLR model with a cubic B-spline; true data from Weibull distributed ( $shape = 2$ ).

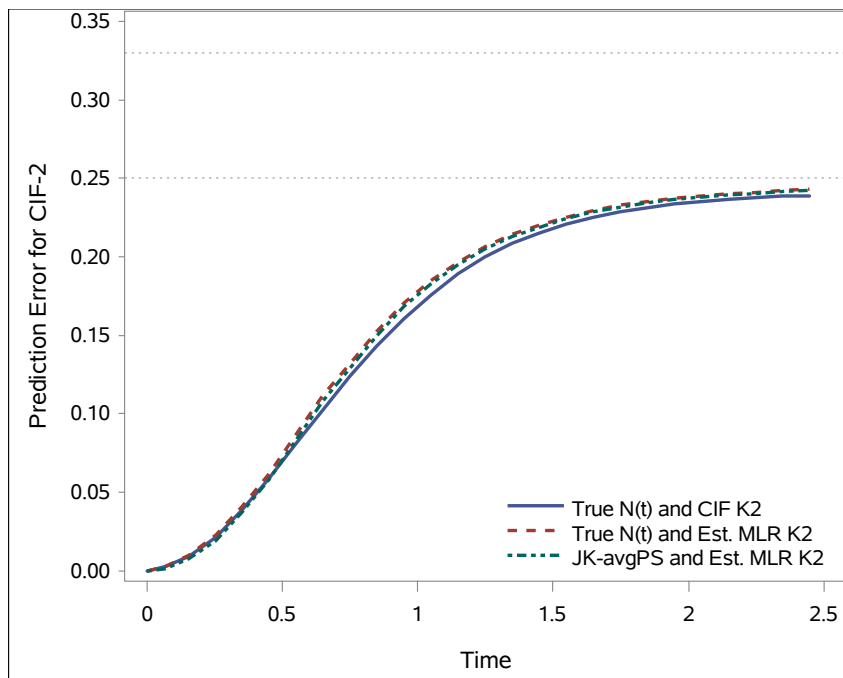


Figure 3.20: PE for the cause-2 CIF based on MLR model with a cubic B-spline; true data from Weibull distributed ( $shape = 2$ ).

### 3.4 APPLICATION TO THE STUDY OF DEMENTIA EPIDEMIOLOGY

We applied the proposed PE method to the MYHAT study ([Ganguli et al., 2009](#)) that was introduced in previous Section 2.4. Continued with the estimated CIFs from those fitted models in previous Section 2.4, we calculated the estimated PE accordingly.

Figure 3.21 presents the four estimated PE curves from the model I with a linear  $A_k(t)$ , and the three models(model 0, I-B, II-B) with a cubic B-Spline incorporated. We observed the PE curves from all above models are below the benchmark of 33%. The univariate model with a cubic B-Spline (model 0) performs better than the model I with a linear  $A_k(t)$  and adjusted for other covariates. Among the three models with a cubic B-Spline incorporated, the PE curves shows better performance along with more covariates included in the model. It is evident that the PE is able to catch each of the model improvement such as to compare performances of different models. The model II-B with a cubic B-Spline method has the best PE curve among all models. It suggests this overall model is the best fitted model for

the given data. It is also noticed that the PE estimates from the model II-B with a cubic B-Spline method are below the benchmark of 25% most of time. Thus it is still an acceptable model even under the most strict benchmark of 25% for interval-censored data.

The Figure 3.22 displays the PE curves from all the six models in the previous Section 2.4 together. Among the model I with a linear  $A_k(t)$  and the two models with a polynomial  $A_k(t)$  (model I-A, II-A), their PE curves get better while a polynomial function or more covariates are introduced. It again suggests that the PE is able to catch each of the model improvement among different models and capable to be robust metric for prediction models. The PE estimates from the models with a cubic B-Spline are obviously improved over the other corresponding models while controlled for same covariates. It implies the proposed MLR model with a cubic B-Spline method has great flexibility, which makes it as a strong candidate for competing risks data analysis.

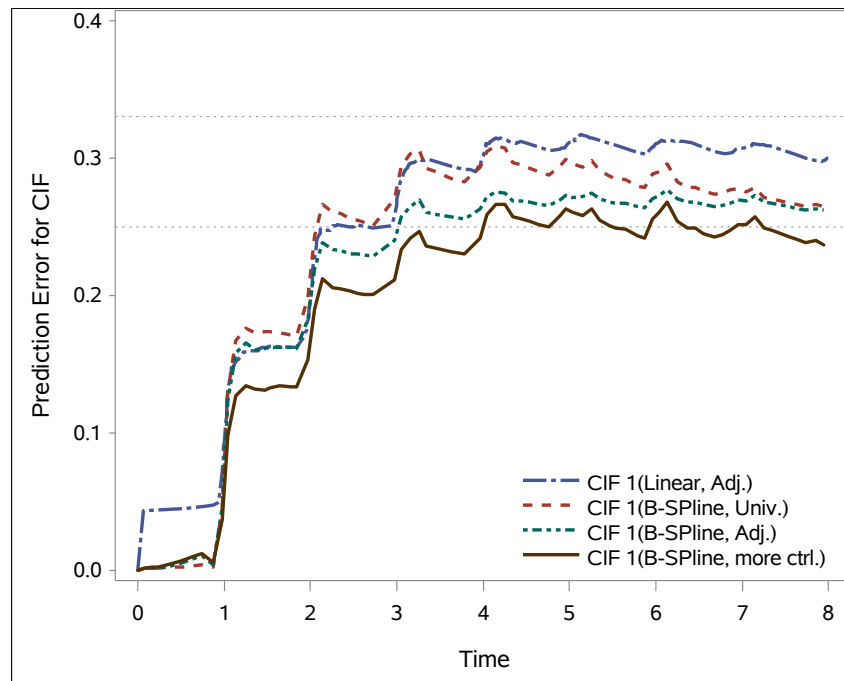


Figure 3.21: PE based on the different models with a B-Spline.

The Adj. denotes same covariates as in the model I; the more ctrl. denotes same covariates as in model II. The two vertical reference lines are benchmarks at 0.25 and 0.33.

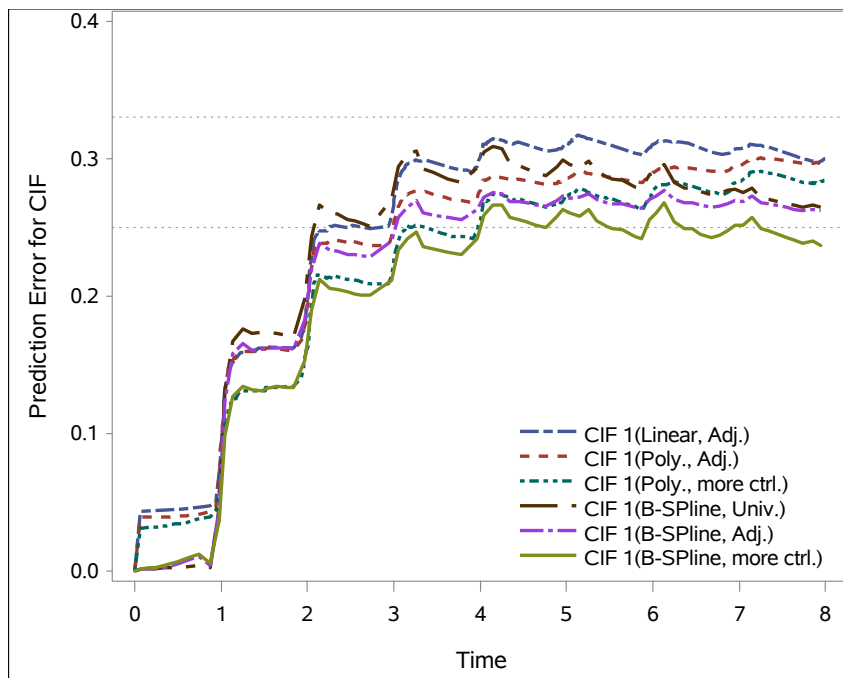


Figure 3.22: PE estimates comparison among all different models.

The Adj. denotes same covariates as in the model I; the more ctrl. denotes same covariates as in model II. The two vertical reference lines are benchmarks at 0.25 and 0.33.

### 3.5 DISCUSSION

In this study we have proposed the PE as a model-free prediction accuracy metric to evaluate direct CIF modelling under interval-censored competing risks setting. A complete solution which utilizes Jackknife pseudovalues to estimate the PE, has been constructed for interval-censored competing risks data. We investigated the consistency of the proposed PE estimator and illustrated its finite-sample performance in three simulation studies including both correct and misspecified models. The Jackknife approach has been shown great performance and reliability to estimate the incomplete event status within a censored interval. The proposed PE estimator has been clearly proved to have the capability to examine misspecification under the interval-censored competing risks setting. To the best of our knowledge, we believe this is the first strictly proper scoring rule for direct CIF modelling under interval-censored competing risks setting. A further research regarding the power of this assessment is needed and it is interesting topic for future.

For interval-censored competing risks data, the main challenge is how to deal with the incomplete observations, which are due to both the interval censoring mechanism or prevention from competing risks. The proposed method avoids modeling the censoring distribution and takes advantage of the Jackknife resampling technique. The estimation of PE utilizes Jackknife pseudovalues which are originally based on the smoothed NPMLE of CIFs. Thus the performance of proposed PE might be affected by those smoothed NPMLE of CIFs. This will be further studied in future.

The PE consists of the variance and bias components based on its decomposition Equation (3.2.3). If the variance part,  $E[F_k(s|Z)(1 - F_k(s|Z))]$  is dominant in the PE decomposition, then the bias part,  $E[F_k(s|Z) - \hat{\pi}_k(s|Z)]^2$  is relatively small and thus can not impact the PE markedly under misspecification. Our simulation studies illustrated there are only noticeable discrepancy between the true and estimated PE curves under some misspecification settings. [Schoop et al. \(2011\)](#) also pointed out the bias part is a small portion compared to the variance part in their simulation studies related to right-censored competing risks setting. This might be a potential limitation on the sensitivity of the PE metric to examine misspecification.

In our study, we proposed the PE metric for the interval-censored competing risks data. It can be readily applied to any censoring mechanism with or without competing risks. It has been assumed that causes of all events are observed and covariates are time-invariant only. However it is not unusual that there are missing causes and time-varying covariates in practice. It is a desired research topic for future work.



## BIBLIOGRAPHY

- Belot, A., M. Abrahamowicz, L. Remontet, and R. Giorgi (2010). Flexible modeling of competing risks in survival analysis. *Statistics in medicine* 29(23), 2453–2468.
- Binder, H., A. Allignol, M. Schumacher, and J. Beyersmann (2009). Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics* 25(7), 890–896.
- Blanche, P., C. ProustLima, L. Loubre, C. Berr, J. Dartigues, and H. JacqminGadda (2015). Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and timetoevent in presence of censoring and competing risks. *Biometrics* 71(1), 102–113.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review* 78(1), 1–3.
- Chen, D.-G. D., J. Sun, and K. E. Peace (2012). *Interval-censored time-to-event data: methods and applications*. CRC Press.
- Cheng, Y. (2009). Modeling cumulative incidences of dementia and dementia-free death using a novel three-parameter logistic function. *The International Journal of Biostatistics* 5(1).
- Cortese, G., T. A. Gerds, and P. K. Andersen (2013). Comparing predictions among competing risks models with timedependent covariates. *Statistics in medicine* 32(18), 3089–3101.
- De Boor, C. (2001). *A practical guide to splines*. Springer.
- Eriksson, F., J. Li, T. Scheike, and M. Zhang (2015). The proportional odds cumulative incidence model for competing risks. *Biometrics* 71(3), 687–695.
- Fine, J. P. and R. J. Gray (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association* 94(446), 496–509.
- Frydman, H. and J. Liu (2013). Nonparametric estimation of the cumulative intensities in an interval censored competing risks model. *Lifetime data analysis* 19(1), 79–99.
- Ganguli, M., B. Snitz, J. V. Bilt, and C. H. Chang (2009). How much do depressive symptoms affect cognition at the population level? the monongahelayoughioghenny healthy aging team (MYHAT) study. *International journal of geriatric psychiatry* 24(11), 1277–1284.

- Ge, M. and M.-H. Chen (2012). Bayesian inference of the fully specified subdistribution model for survival data with competing risks. *Lifetime data analysis* 18(3), 339.
- Geman, S. and C.-R. Hwang (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, 401–414.
- Gerds, T. A., T. Cai, and M. Schumacher (2008). The performance of risk prediction models. *Biometrical journal* 50(4), 457–479.
- Gerds, T. A., T. H. Scheike, and P. K. Andersen (2012). Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Statistics in medicine* 31(29), 3921–3930.
- Gerds, T. A. and M. Schumacher (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal* 48(6), 1029–1040.
- Gerds, T. A. and M. Schumacher (2007). Efron-type measures of prediction error for survival analysis. *Biometrics* 63(4), 1283–1287.
- Gmez, G., M. L. Calle, R. Oller, and K. Langohr (2009). Tutorial on methods for interval-censored data and their implementation in R. *Statistical Modelling* 9(4), 259–297.
- Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477), 359–378.
- Graf, E., C. Schmoor, W. Sauerbrei, and M. Schumacher (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine* 18(1718), 2529–2545.
- Graw, F., T. A. Gerds, and M. Schumacher (2009). On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis* 15(2), 241–255.
- Gray, R. J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of statistics*, 1141–1154.
- Graziane, J. A., J. C. Beer, B. E. Snitz, C.-C. H. Chang, and M. Ganguli (2016). Dual trajectories of depression and cognition: a longitudinal population-based study. *The American Journal of Geriatric Psychiatry* 24(5), 364–373.
- Groeneboom, P., M. H. Maathuis, and J. A. Wellner (2008a). Current status data with competing risks: consistency and rates of convergence of the MLE. *The Annals of Statistics*, 1031–1063.
- Groeneboom, P., M. H. Maathuis, and J. A. Wellner (2008b). Current status data with competing risks: Limiting distribution of the MLE. *Annals of statistics* 36(3), 1064.
- Hand, D. J. (1997). *Construction and assessment of classification rules*. Wiley.

- Harrell, F. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*, pp. 26–28. Springer.
- Heller, G. (2011). Proportional hazards regression with interval censored data using an inverse probability weight. *Lifetime data analysis* 17(3), 373–385.
- Hudgens, M. G., C. Li, and J. P. Fine (2014). Parametric likelihood inference for interval censored competing risks data. *Biometrics* 70(1), 1–9.
- Hudgens, M. G., G. A. Satten, and I. M. Longini (2001). Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation. *Biometrics* 57(1), 74–80.
- Hughes, T. F., J. T. Becker, C.-W. Lee, C.-C. H. Chang, and M. Ganguli (2015). Independent and combined effects of cognitive and physical activity on incident MCI. *Alzheimer's & Dementia* 11(11), 1377–1384.
- Jeong, J. and J. Fine (2006). Direct parametric inference for the cumulative incidence function. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 55(2), 187–200.
- Jeong, J.-H. and J. P. Fine (2007). Parametric regression on cumulative incidence function. *Biostatistics* 8(2), 184–196.
- Jewell, N. P. and J. D. Kalbfleisch (2004). Maximum likelihood estimation of ordered multinomial parameters. *Biostatistics* 5(2), 291–306.
- Jewell, N. P., M. van der Laan, and T. Henneman (2003). Nonparametric estimation from current status data with competing risks. *Biometrika* 90(1), 183–197.
- Kalbfleisch, J. D. and R. L. Prentice (2002). *Competing Risks and Multistate Models*, Chapter 8, pp. 247–277. John Wiley & Sons, Inc.
- Klein, J. P. and P. K. Andersen (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics* 61(1), 223–229.
- Li, C. (2016). The Fine-Gray model under interval censored competing risks data. *Journal of Multivariate Analysis* 143, 327–344.
- Lunn, M. and D. McNeil (1995). Applying Cox regression to competing risks. *Biometrics*, 524–532.
- Maathuis, M. H. (2006). *Nonparametric estimation for current status data with competing risks*. Doctoral dissertation, University of Washington.
- Maathuis, M. H. and M. G. Hudgens (2011). Nonparametric inference for competing risks current status data with continuous, discrete or grouped observation times. *Biometrika*, asq083.

- Mao, L. and D. Lin (2016). Efficient estimation of semiparametric transformation models for the cumulative incidence of competing risks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.
- Mao, L., D.-Y. Lin, and D. Zeng (2017). Semiparametric regression analysis of interval-censored competing risks data. *Biometrics Epub ahead of print*.
- McDade, E., Z. Sun, C.-W. Lee, B. Snitz, T. Hughes, C.-C. H. Chang, and M. Ganguli (2016). The association between pulse pressure change and cognition in late life: Age and where you start matters. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 4, 56–66.
- Peto, R. (1973). Experimental survival curves for interval-censored data. *Applied Statistics*, 86–91.
- Prentice, R. L., J. D. Kalbfleisch, A. V. Peterson Jr, N. Flournoy, V. Farewell, and N. Breslow (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 541–554.
- Ren, Y. (2015). *Proportional subdistribution hazards regression with interval-censored competing risks data*. Doctoral dissertation, University of Pittsburgh.
- Robins, J. M. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the Biopharmaceutical Section, American Statistical Association*, Volume 24, pp. 3.
- Royston, P. and D. G. Altman (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied statistics*, 429–467.
- Rutherford, M. J., M. J. Crowther, and P. C. Lambert (2015). The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *Journal of Statistical Computation and Simulation* 85(4), 777–793.
- SAS Institute Inc. (2012). *SAS/STAT Software, Version 9.4*. Cary, NC.
- Sauerbrei, W., C. Meier-Hirmer, A. Benner, and P. Royston (2006). Multivariable regression model building by using fractional polynomials: description of SAS, STATA and R programs. *Computational Statistics & Data Analysis* 50(12), 3464–3485.
- Sauerbrei, W. and P. Royston (1999). Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 162(1), 71–94.
- Scheike, T. H. and M.-J. Zhang (2008). Flexible competing risks regression modeling and goodness-of-fit. *Lifetime data analysis* 14(4), 464–483.

- Schoop, R., J. Beyersmann, M. Schumacher, and H. Binder (2011). Quantifying the predictive accuracy of timetoevent models in the presence of competing risks. *Biometrical Journal* 53(1), 88–112.
- Schoop, R., E. Graf, and M. Schumacher (2008). Quantifying the predictive performance of prognostic models for censored survival data with timedependent covariates. *Biometrics* 64(2), 603–610.
- Schumaker, L. (2007). *Spline Functions: Basic Theory*, pp. 108–109. Cambridge University Press.
- Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics* 1(1), 43–62.
- Shen, X. (1998). Propotional odds regression and sieve maximum likelihood estimation. *Biometrika* 85(1), 165–177.
- Shi, H., Y. Cheng, and J. Jeong (2013). Constrained parametric model for simultaneous inference of two cumulative incidence functions. *Biometrical Journal* 55(1), 82–96.
- Stone, C. J. (1986). [generalized additive models]: Comment. *Statistical Science* 1(3), 312–314.
- Sun, J., Y. Feng, and H. Zhao (2015). Simple estimation procedures for regression analysis of interval-censored failure time data under the proportional hazards model. *Lifetime data analysis* 21(1), 138–155.
- Sun, J. and J. Shen (2009). Efficient estimation for the proportional hazards model with competing risks and current status data. *Canadian Journal of Statistics* 37(4), 592–606.
- Tanner, M. A. and W. H. Wong (1987). An application of imputation to an estimation problem in grouped lifetime analysis. *Technometrics* 29(1), 23–32.
- Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 290–295.
- Wei, G. C. and M. A. Tanner (1991). Applications of multiple imputation to the analysis of censored regression data. *Biometrics*, 1297–1309.
- Werren, S. (2011). *Pseudo-likelihood methods for the analysis of interval censored data*. Thesis, ETH, Zurich.
- Zeng, D., L. Mao, and D. Lin (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika* 103(2), 253–271.

Zhang, Y., L. Hua, and J. Huang (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. (2009). Linear transformation models for interval-censored data prediction of survival probability and model checking. *Statistical Modelling* 9(4), 321–343.