ASSOCIATION ANALYSIS OF SUCCESSIVE EVENTS AND DIAGNOSTIC ACCURACY ANALYSIS FOR COMPETING RISKS DATA

by

Xiaotian Chen

B. S. Statistics, Zhejiang University, 2010

M. A. Statistics, University of Pittsburgh, 2014

Submitted to the Graduate Faculty of
the Kenneth P. Dietrich School of Arts and Sciences in partial
fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

THE KENNETH P. DIETRICH SCHOOL OF ARTS & SCIENCES

This dissertation was presented

by

Xiaotian Chen

It was defended on

June 25, 2015

and approved by

Yu Cheng, Associate Professor, Department of Statistics

Satish Iyengar, Professor, Department of Statistics

Leon J. Gleser, Professor, Department of Statistics

Chung-Chou H. Chang, Professor, Department of Biostatistics

Dissertation Director: Yu Cheng, Associate Professor, Department of Statistics

ASSOCIATION ANALYSIS OF SUCCESSIVE EVENTS AND DIAGNOSTIC ACCURACY ANALYSIS FOR COMPETING RISKS DATA

Xiaotian Chen, PhD

University of Pittsburgh, 2015

In this dissertation there are two overarching objectives to address the challenges in analyzing data from the Bipolar Disorder Center for Pennsylvanians (BDCP) study.

First, we aim to close a methodological gap in analyzing durations of successive events that are subject to induced dependent censoring as well as competing-risk censoring. In the BDCP study, some patients who managed to recover from their symptomatic entry later developed a new depressive or manic episode. It is of clinical interest to quantify the association between time to recovery and time to recurrence in patients with bipolar disorder. The estimation of the bivariate distribution of the gap times with independent censoring has been well studied. However, the existing methods cannot be applied to failure times censored by competing causes. Bivariate cumulative incidence function (CIF) has been used to describe the joint distribution of parallel event times that involve multiple causes. However, there is no method available for successive events with competing-risk censoring. Therefore, we extend the bivariate CIF to successive events data, and propose nonparametric estimators. Moreover, an odds ratio measure is proposed to describe the cause-specific dependence, leading to the development of a formal test for independence of successive events. The method is evaluated through simulations and also applied to the real dataset.

Next, motivated by another subgroup of subjects in the BDCP study who entered the study in a euthymic state, we investigate the Receiver Operating Characteristic (ROC) approach for a competing-risk censored outcome, when the diagnostic marker of interest, number of previous episodes, can be treated as censored observations. We propose two methods to estimate the discrimination measures such as sensitivity, specificity, positive and negative predictive values and the Area Under the Curve (AUC). We also develop cause-specific tests to compare two markers' discriminatory abilities in separating those subjects who will experience the cause-specific event by some time point from those who will not. The proposed estimators and tests have satisfactory performance in simulation studies. We also illustrate these methods through the analysis of the BDCP subsample.

Keywords: Bipolar disorder; Competing risks; Cumulative incidence function; Inverse probability weighting; Odds ratio; Successive events; AUC; Discrimination; ROC.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Time-to-event data are commonly encountered in biomedical, reliability, and finanical studies. When there are multiple interacting endpoints, the event of interest may not be observed, if the other competing events have occured first. This phenomenon is referred to as competing-risk censoring, and it is important to take into account the potential dependence between the event of interest and the competing events. As an example, in the Bipolar Disorder Center for Pennsylvanians (BDCP) study, the occurrences of depression and mania dependently censor each other, and in the analysis of depressive episodes, for instance, ignoring mania or treating it as independent censoring would lead to biased results. As detailed in this thesis, our work contains two parts, both dealing with competing-risk censored outcomes.

The first part addresses the challenge in analyzing successive data from a subsample of the BDCP study, where patients with symptomatic entry were followed up for recovery and subsequent new episodes. Only when a recovery occurred were we able to observe a new recurring episode. Moreover, the recurrence is subject to competing-risk censoring as either a depressive or manic episode was observed. Quantifying the association between successive events have been studied in the literature. However, the situation where the subsequent event is competing-risk censored has not yet been considered. Hence, we propose non-parametric association analyses in this successive event setting based on the bivariate cumulative incidence function (CIF) and the conditional CIF. We show through simulations that our methods perform well and we also apply the methods to the BDCP study.

In the second part of this thesis, we focus on time-dependent diagnostic accuracy analyses for a competing-risk censored outcome with a censored marker. The receiver operating characteristic (ROC) curve has been widely utilized to evaluate diagnostic accuracy of markers for a dichotomous outcome. The traditional ROC curve was extended in Heagerty et al. (2000) from a binary outcome to an event time, where diagnostic accuracy measures depend on time. It has also been generalized to competing-risk censored outcomes recently (Saha and Heagerty, 2010; Foucher et al., 2010). In this part, we consider another subsample of the BDCP study, where patients entered the study without any obvious symptoms (i.e., in a euthymic state). We are interested in evaluating the discrimination ability of the number of previously experienced episodes, as a marker, in separating those euthymic patients who developed a new episode quickly from those who relapsed late by a specific time point. However, the marker is subject to right censoring, as the number becomes difficult for patients to recall when it is extremely large. Therefore, we consider such a circumstance where both the marker and the outcome are censored and the outcome is also subject to competing-risk censoring. We will develop non-parametric discriminatory analyses and evaluate the proposed estimators and cause-specific tests for two-marker comparison by simulation. We will also apply the methods to the BDCP study.

2.0 ASSOCIATION ANALYSIS OF SUCCESSIVE EVENTS DATA IN THE PRESENCE OF COMPETING RISKS DATA

2.1 INTRODUCTION

Event time data are frequently encountered in practice. Multivariate survival times data arise when a study involves multiple interacting events–for example, time to recovery from symptomatic entry and time to relapse. It is often of scientific interest to quantify the association between successive event times. For successive events, only when the first gap time occurs does the second gap time have the chance to be observed. Unless the two duration periods are independent, the second gap time would be subject to censoring by a dependent variable that is related to the first duration time, which is usually referred to as induced dependent censoring. In many applications, the two consecutive events are dependent and their association is often quantified through their bivariate distribution.

A number of estimators have been proposed over the years to handle induced dependent censoring. Visser (1996) considered non-parametric estimation of the bivariate survival function in discrete time cases. Wang and Wells (1998) considered a more general situation and developed inverse probability weighting estimators for the bivariate survival function of gap times.Lin et al. (1999) and van der Laan et al. (2002) provided estimators for the multivariate gap time distribution function, while de Uña Álvarez and Meira-Machado (2008) constructed an estimator for the bivariate distribution function using an inverse probability weight different from what was used in Lin et al. (1999). Schaubel and Cai (2004) considered estimation for the conditional gap-time specific survival function. Huang and Louis (1998) focused on bivariate survival data with mark variables. Wang (1999) discussed bias in estimating the second gap time distribution under incident and prevalent cohorts. Lin and Ying (2001) developed nonparametric tests for gap time distributions of right-censored serial events. Huan and Wang (2005) studied the nonparametric estimation of bivariate recurrence times. Some other mechanisms of missing data such as left truncation (Chang and Tzeng, 2006; Shen, 2010) or interval censoring (Zhu and Wang, 2012) have also been studied.

However, none of these methods can be directly applied to the data from the Bipolar Disorder Center for Pennsylvanians (BDCP) study, where successive events involve multiple endpoints Fagiolini et al. (2009). Patients entered the BDCP study with a DSM-IV diagnosis of bipolar disorder, and they were followed up for a median duration of one year. Most of the patients managed to recover at some point from their symptomatic entry, and some of them later developed a specific new depressive or manic episode during the remaining follow-up period. It is of great scientific interest to examine the association between time to recovery and time to recurrence for each cause, as it will help identify high-risk patients and inform treatment decisions on how to better prevent or at least delay the recurrence of a new episode. However, the association analysis of time to recovery and time to relapse is complicated by the potential mutual dependence between depression and mania, while existing methods only consider one cause for each successive event.

We endeavor to quantify the association between the recovery time and the recurrence time since recovery for both causes. Multistate models, usually viewed as time-inhomogeneous Markov processes, are sometimes favorable for time-to-event data with multiple stages. Some of their appealing aspects have been discussed and compared with other classical models (Gill, 1992). If we refer to the entry of study as state 1, recovery as state 2, and subsequent new depressive or manic episodes as absorbing state 3 or state 4, then nonparametric estimation of the transition probabilities between states could be performed (Beyersmann et al., 2011). More specifically, the joint distribution of time to recovery and time to recurrence since study entry for both causes can be obtained using the Aalen-Johansen estimator (Aalen and Johansen, 1978), since these two time periods are both censored independently by the end of the study. The consistency and weak convergence of this estimator have been established when the underlying model is either Markovian or non-Markovian

(Datta and Satten, 2001). However, compared with the gap time between recovery and the subsequent new episode, the time since study entry until recurrence of depression or mania will be of less clinical interest. Therefore, in this chapter we focus on the joint distribution of recovery time and recurrence time since recovery for both causes.

Care must be taken when handling this competing-risk censoring problem in the presence of induced dependent censoring. Univariate competing risks data have been extensively studied (Kalbfleisch and Prentice, 2002; Klein and Moeschberger, 2003). However, any standard methods using the basic quantities such as a cumulative incidence function (CIF) or a survival function should account for possibly dependent risks or dependent censoring (Pepe, 1991; Pepe and Mori, 1993). With induced dependent censoring, the naive inference for the second gap time from a specific cause that ignores the first gap time will lead to biased results, because the analysis only contains those patients who were able to recover within a relatively short period of time. Hence, it cannot represent the entire population that scientific researchers intend to study. An appropriate analysis should include the first gap time as well.

Bivariate time-to-event data with competing-risk censoring have been studied during the past several years. Bandeen-Roche and Liang (2002) examined the conditional cause-specific hazard ratio for bivariate competing risks data through a semiparametric frailty model; Cheng and Fine (2008), Bandeen-Roche and Ning (2008) and Cheng et al. (2010) investigated nonparametric methods for this or an equivalent association measure. Shih and Albert (2010) proposed a bivariate model to study the association between the first event times and between failure types. Scheike et al. (2010) proposed a cross-odds ratio measure for the association between cause-specific failure times assuming a semiparametric model. Cheng et al. (2007) studied bivariate parallel competing risks data where they developed non-parametric inference on a bivariate cause-specific hazard function as well as a bivariate CIF. Sankaran et al. (2006) also considered the estimation for bivariate CIF. Cheng and Fine (2012) studied association models through CIFs and frailty models. However, none of these existing methodologies can be directly applied to the problem that we are dealing with, since they all assume that the outcomes are parallel instead of being successive. The circumstance where the second gap time is competing-risk censored and meanwhile subject to induced dependent censoring by the previous gap time has not been considered yet as far as we are aware.

Therefore, in this chapter we aim to close this methodological gap in analyzing the BDCP data, and propose non-parametric association analyses of two successive event times that explicitly take into account induced dependent censoring by the preceding event and competing-risk censoring between the two competing causes of failure in the subsequent event. Our analyses will focus on capturing the association between the time to recovery and the subsequent time to a new depressive or manic episode, using the bivariate CIF, since the univariate CIF is widely used in the competing risks literature, and the bivariate CIF enjoys the appealing probability interpretation analogous to the univariate CIF. We adapt and naturally extend the nonparametric estimation of the bivariate CIF proposed by Cheng et al. (2007) to appropriately deal with the successive events setting. Inverse probability weighting is incorporated in constructing the estimators for the bivariate CIF, where only the second component of the bivariate variable is subject to competing-risk censoring. We will also consider the conditional CIF, based on which we further construct a cause-specific odds ratio measure to describe association between gap times. These quantities are nonparametrically identifiable and can be very effectively adopted in analyzing the relationship between the time to recovery and the time to a subsequent relapse. In addition, a formal test can be conducted for independence between the two successive events.

The remainder of the chapter is organized as follows. In Section 2.2, we will detail the nonparametric estimation for the bivariate and conditional CIF, along with the cause-specific association measure and the tests. In Section 2.3, simulation results are presented where we evaluate the finite sample performance of our proposed estimators and the tests. Various dependence structures will be considered. Our estimators and tests are shown to perform very well. An application to the BDCP study is illustrated in Section 2.4. We conclude with some remarks in Section 2.5.

2.2 METHOD

In this chapter we consider the situation that each individual may experience two consecutive events. Let X denote the first time duration. The second time period is denoted by T, and without loss of generality, we assume that it is subject to two failure types, as multiple competing events can be grouped together into one. We observe X and T with the corresponding cause indicator $\epsilon = 1, 2$ for the second event. The censoring time C is assumed to be independent of X and T. Only when X is observed are we able to observe T. Clearly, the first event time X is independently censored by C. Unless X is independent of T, T is subject to dependent censoring, since T is censored by $(C - X)I(X < C)$. Let $\delta_1 = I(X \leq C), \delta_2 = I(X + T \leq C)$ be the censoring indicators. For each individual, we observe $(Y_1, Y_2, \eta_1, \eta_2)$, where $Y_1 = \min(X, C)$, $Y_2 = \min\{(X + T), C\}$, $\eta_1 = \delta_1$, and $\eta_2 = \delta_2 \cdot \epsilon$. The data consist of n independent and identically distributed copies of $(Y_1, Y_2, \eta_1, \eta_2)$, denoted as $\{(Y_{1i}, Y_{2i}, \eta_{1i}, \eta_{2i}), i = 1, \ldots, n\}$.

2.2.1 Estimating the bivariate CIF

We first extend the bivariate CIF to this successive events setting and define

$$\begin{aligned} F_l(x,t) &= P(X \leq x, T \leq t, \epsilon = l) && (2.2.1)\\ &= \int_0^x \int_0^t \lambda_l(u,v) S(u-, v-) du dv, && (2.2.2) \end{aligned}$$

where $l = 1, 2$ for the two competing causes of the subsequent event, and $S(u, v) = P(X > u, T > v)$ is the overall bivariate survival function. The bivariate cause-specific hazard (CSH) function is defined by

$$\lambda_l(x,t) = \lim_{\Delta x \to 0, \Delta t \to 0} \frac{1}{\Delta x \Delta t} P(x \leq X < x + \Delta x, t \leq T < t + \Delta t, \epsilon = l | X \geq x, T \geq t).$$

and the bivariate cumulative cause-specific hazard function is defined by

$$\Lambda_l(x,t) = \int_0^x \int_0^t \lambda_l(u,v) du dv.$$

The second duration time T is subject to two failure times while the first duration X is not. Hence the above bivariate CIF and CSH functions are special cases of the usual bivariate quantities that are used for bivariate parallel competing risks data.

For bivariate competing risks data there are typically two types of right censoring: bivariate censoring, where the two censoring times are parallel and often different for the two subjects in a pair, and univariate censoring, where the same censoring time is applied to the two individuals in a pair (e.g., two eyes from the same subject). In our case for the successive events, there is only one censoring time, i.e. the end of the study, and all the gap times could be potentially censored by this univariate censoring time. There have been some works on the estimation for the bivariate CIF for bivariate parallel competing risks data (Cheng et al., 2007; Sankaran et al., 2006), and the methods often require that the pair of censoring times be independent of the pair of event times regardless of the causes, even though the dependence structure between the two censoring times could be arbitrary. In the successive events setup, however, the second event time T can only become observable when the first event time X is observed. Hence,(X, T) as a pair would be subject to censoring by $(C, (C - X)I(X < C))$. Even though the administrative censoring C is assumed to be independent of both X and T, the censoring time for T, $(C - X)I(X < C)$, is likely to depend on T, due to the potential dependence between X and T. Therefore, the existing estimators cannot be directly applied here for the induced dependent censoring situation.

We now estimate the bivariate CIF $F_l(x, t)$, $l = 1, 2$ that is defined in (2.2.1). We will consider two approaches, under the current induced dependent censoring framework. One is in line with the nonparametric estimator proposed in Cheng et al. (2007), where we estimate the bivariate cumulative cause-specific hazard function and the overall bivariate survival function, and then plug them into the formula given in (2.2.2). The other approach directly utilizes the pairs where both time to recovery and time to a new specific episode were observed.

We start with defining the at-risk and the event processes. Since $\eta_{1i} = 1$ indicates $Y_{1i} = X_i$, for $u, v \geq 0$, we have

$$I(Y_{1i} \geq u, \eta_{1i} = 1, Y_{2i} - Y_{1i} \geq v) = I(X_i \geq u, T_i \geq v, C_i \geq X_i + v)$$

and

$$I(Y_{1i} \leq u, \eta_{1i} = 1, Y_{2i} - Y_{1i} \leq v, \eta_{2i} = l) = I(X_i \leq u, T_i \leq v, \epsilon_i = l, C_i \geq X_i + T_i),$$

where $l = 1, 2$ for cause 1 or 2, respectively. Thus, we have the following conditional expectations

$$E\{I(Y_{1i} \geq u, \eta_{1i} = 1, Y_{2i} - Y_{1i} \geq v)|X_i, T_i\} = I(X_i \geq u, T_i \geq v)P(C_i \geq X_i + v)$$

and

$$E\{I(Y_{1i} \leq u, \eta_{1i} = 1, Y_{2i} - Y_{1i} \leq v, \eta_{2i} = l)|X_i, T_i\} = I(X_i \leq u, T_i \leq v, \epsilon_i = l)P(C_i \geq X_i + T_i).$$

Let G denote the survival function of the censoring time variable C. Then

$$E\Big[E\{\frac{I(Y_{1i} \geq u, \eta_{1i} = 1, Y_{2i} - Y_{1i} \geq v)}{G((X_i + v)-)}|X_i, T_i\}\Big] = P(X_i \geq u, T_i \geq v)$$

and

$$E\Big[E\{\frac{I(Y_{1i} \leq u, \eta_{1i} = 1, Y_{2i} - Y_{1i} \leq v, \eta_{2i} = l)}{G((X_i + T_i)-)}|X_i, T_i\}\Big] = P(X_i \leq u, T_i \leq v, \epsilon_i = l). \tag{2.2.3}$$

Hence we define the empirical process $\mathbb{P}_n H(x,t) = \frac{1}{n}\sum_{i=1}^n \{I(Y_{1i} \geq x, \eta_{1i} = 1, Y_{2i} - Y_{1i} \geq t)/\hat{G}((Y_{1i} + t)-)\}$, where $\hat{G}$ is the Kaplan-Meier estimator of G calculated from $(Y_{2i}, 1 - \delta_{2i})$, and $\hat{G}((Y_{1i} + t)-)$ is the left-hand limit of $\hat{G}(Y_{1i} + t)$. Similarly, we define $\mathbb{P}_n N_l(x,t) = \frac{1}{n}\sum_{i=1}^n \{I(Y_{1i} \leq x, \eta_{1i} = 1, Y_{2i} - Y_{1i} \leq t, \eta_{2i} = l)/\hat{G}(Y_{2i}-)\}$ for cause l. Thus, $\mathbb{P}_n H(x,t)$ and $\mathbb{P}_n N_l(x,t)$ are the weighted numbers of individuals at risk at (x,t) and those who have had the events of interest by (x,t), respectively. Then the bivariate cumulative cause-specific hazard function can be estimated by

$$\hat{\Lambda}_l(x,t) = \int_0^x \int_0^t \mathbb{P}_n N_l(du, dv) \mathbb{P}_n^{-1} H(u,v). \tag{2.2.4}$$

A natural problem arises when the denominators involving $\hat{G}$ are zeros, which would occur if the largest observed value of Y_{2i} is the censoring time. In that case, the estimated survival function $\hat{G}$ evaluated at time larger or equal to Y_{2i} would be 0. However, the estimators are still valid if we adopt the convention that $0/0 = 0$, which guarantees that $\hat{\Lambda}_l$ is well defined.

Next, we consider the estimation for the overall bivariate survival function S. In order to construct the estimator for S, we first consider the quantity $K(x,t) = P(X \le x, T > t)$, which was proposed and utilized by Lin et al. (1999). They also proved the consistency and weak convergence for its non-parametric estimator $\hat{K}$ which will be introduced below. The estimator utilizes the pairs of subjects whose recovery times are observed. Note that de Uña Álvarez and Meira-Machado (2008) proposed an alternative approach to estimating K, where they only used the pairs whose recovery and recurrence times are both observed. We will use the estimator $\hat{K}$ from Lin et al. (1999), with $\hat{K}(u,v) = \frac{1}{n}\sum_{i=1}^{n}\{I(Y_{1i} \le u, Y_{2i} - Y_{1i} > v)/\hat{G}(Y_{1i} + v)\}$. Since $S(x,t) = K(\infty,t) - K(x,t)$, the corresponding estimator for S would be $\hat{S}(u,v) = \hat{K}(\infty,v) - \hat{K}(u,v)$. Hence we estimate $F_l(x,t)$ by

$$\hat{F}_l(x,t) = \int_0^x \int_0^t \hat{S}(u-,v-)\hat{\Lambda}_l(du,dv). \tag{2.2.5}$$

On the other hand, in an effort to estimate the bivariate CIF more directly, we notice that according to (2.2.3), $\frac{1}{n}\sum_{i=1}^{n} I(Y_{1i} \le x, \eta_{1i} = 1, Y_{2i} - Y_{1i} \le t, \eta_{2i} = l)/G(Y_{2i}-)$ would be an unbiased estimator for $F_l(x,t)$, if G were known. Since the Kaplan-Meier estimator $\hat{G}$ is an unbiased estimator for G, by Slutsky's theorem, $F_l(x,t)$ would be consistently estimated by $\mathbb{P}_n N_l(x,t) = \frac{1}{n}\sum_{i=1}^{n}\{I(Y_{1i} \le x, \eta_{1i} = 1, Y_{2i} - Y_{1i} \le t, \eta_{2i} = l)/\hat{G}(Y_{2i}-)\}$, which is the weighted empirical process that we defined earlier.

Interestingly, if we examine the form of $\hat{F}_l(x,t)$ in (2.2.5) more closely, after a few steps we can obtain

$$\begin{aligned}
&\hat{S}(u-,v-)\hat{\Lambda}_l(du,dv)\\
=&\frac{1}{n}\sum_{i=1}^{n}\frac{I(Y_{1i} \ge u, Y_{2i} - Y_{1i} \ge v, \eta_{1i} = 1)}{\hat{G}(Y_{1i} + v-)}\\
&\times\frac{\sum_{i=1}^{n} I(Y_{1i} = u, \eta_{1i} = 1, Y_{2i} - Y_{1i} = v, \eta_{2i} = l)/\hat{G}(Y_{2i}-)}{\sum_{i=1}^{n} I(Y_{1i} \ge u, \eta_{1i} = 1, Y_{2i} - Y_{1i} \ge v)/\hat{G}((Y_{1i} + v)-)}\\
=&\frac{1}{n}\sum_{i=1}^{n}\frac{I(Y_{1i} = u, \eta_{1i} = 1, Y_{2i} - Y_{1i} = v, \eta_{2i} = l)}{\hat{G}(Y_{2i}-)}.
\end{aligned}$$

Then $\hat{F}_l(x,t)$ in (2.2.5), as the integration of the above quantity, can be reduced to

$$\hat{F}_l(x,t) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(Y_{1i} \le x, \eta_{1i} = 1, Y_{2i} - Y_{1i} \le t, \eta_{2i} = l)}{\hat{G}(Y_{2i}-)}, \tag{2.2.6}$$

which is just $\mathbb{P}_n N_l(x,t)$. The two approaches we have described are analogous to the work of Sankaran et al. (2006), where they proposed the indirect and direct estimators of the bivariate CIF for a typical bivariate competing risks data. Because of the specific form of the estimator for the overall bivariate survival function that we have used in our successive events setting, our two approaches from these different perspectives turn out to be equivalent.

2.2.2 Estimating the conditional CIF of T

When interest is placed on the second gap time alone, one may want to follow the standard inference procedure that was developed for univariate competing risks data, by ignoring the previous gap time X that might have dependently censored the event of interest T. Suppose $m(\le n)$ subjects developed the first event within the follow-up period, and their subsequent event data $\{(Y_{2j} - Y_{1j}, \eta_{2j}), j = 1, \ldots, m\}$ would be used to obtain the naive non-parametric estimator for the univariate CIF $F_{l(T)}(t)$. Let $N_{l(T)}(t) = \sum_{j=1}^{m} I(Y_{2j} - Y_{1j} \le t, \eta_{2j} = l)$ and $H_T(t) = \sum_{j=1}^{m} I(Y_{2j} - Y_{1j} \ge t)$, where $l = 1, 2$. Then the naive non-parametric estimator of the CIF for T is

$$\hat{F}_{l(T)}(t) = \int_0^t \hat{S}_T(u-) H_T^{-1}(u) dN_{l(T)}(u) \tag{2.2.7}$$

where $\hat{S}_T(u) = \prod_{s \le u}\{1 - \frac{\Delta N_{1(T)}(s) + \Delta N_{2(T)}(s)}{H_T(s)}\}$ is the Kaplan-Meier estimator for the overall survival function of T based on the second event data only. However, the naive estimation method is only correct when independence between the two gap times holds. Therefore, without realizing the fact that the recovery time is of great relevance to the subsequent recurrence, the naive way of analyzing the recurrence time is not sufficient and may lead to biased results. Thus, we will study the distribution of the second gap time while accounting for the dependence between the first and the subsequent event time. The bivariate CIF

$F_l(x,t)$, $l = 1, 2$, is estimable when $x + t \leq R_C$, where R_C is the upper bound of the support of C. Hence it is possible to estimate the conditional CIF which is defined as

$$F_l(t|x) = P(T \leq t, \epsilon = l | X \leq x) = \frac{F_l(x,t)}{F_X(x)},$$

where $F_X(x)$ is the marginal distribution function of the first gap time X which can be estimated by $1 - \hat{S}_X(x)$, and $\hat{S}_X(x)$ is the Kaplan-Meier estimator. Then we naturally have the plug-in estimator for the conditional CIF:

$$\hat{F}_l(t|x) = \frac{\hat{F}_l(x,t)}{1 - \hat{S}_X(x)}. \tag{2.2.8}$$

The above conditional CIFs under the successive events setting can be meaningfully interpreted and utilized to describe the dependence between gap times. It quantifies the cumulative risk of the occurrence for a specific type of event, given the occurrence of the previous event by a certain time. When the two gap times are independent, the above conditional CIF estimator agrees with the naive univariate estimator $\hat{F}_{l(T)}(t)$, which will be shown in the simulation studies.

The estimators for the bivariate CIF and conditional CIF are both strongly consistent. Weak convergence can also be established. Details are discussed in the Appendix. The estimations for the variances of the estimators could be very complicated in their forms. Thus, we will adopt bootstrapping procedures, by taking resamples from the original data, and estimating the variance using the sample variance of the estimates from all the bootstrap samples. The pointwise confidence intervals can then be constructed based on asymptotic normality.

2.2.3 Measuring and testing dependence

In practice, it is often desired to formally assess the cause-specific association between gap times. We hence propose in this section a time-dependent association measure, as well as a summary measure, based on the conditional CIFs introduced above. Odds ratio is a common association measure for binary outcomes, which has been widely accepted by practitioners. Recently, for instance, Shih and Albert (2010) studied a model based on the event type using piecewise constant odds ratios for bivariate competing risks data; Scheike et al. (2010) proposed a cross odds ratio measure to describe the cause-specific association between event times. The odds ratio-type of association measure can be naturally extended to the successive events data with competing-risk censoring. Therefore, we consider the following cause-specific odds ratio

$$\phi_{(l)}(x,t|M) = \frac{P(T \le t, \epsilon = l|X \le x)/\big\{1 - P(T \le t, \epsilon = l|X \le x)\big\}}{P(T \le t, \epsilon = l|x < X \le M)/\big\{1 - P(T \le t, \epsilon = l|x < X \le M)\big\}}, \qquad (2.2.9)$$

where $l = 1, 2$ and M is a fixed constant that can be chosen as appropriate. In the BDCP study, for example, M can be a suitably large time by which most patients have recovered. $\phi_{(l)}(x,t|M)$ is the ratio of the odds of having a cause l successive event by t between those who had early occurrence of the first event and those who developed the first event later. When X and T are independent for cause l, $\phi_{(l)}(x,t|M) = 1$ for all $x < M$ and $t \le R_C - x$. $\phi_{(l)}(x,t|M) > 1$ indicates that time to the cause l successive event is positively associated with the first event time, while $\phi_{(l)}(x,t|M) < 1$ indicates negative association. The estimation of the quantity in the numerator of (2.2.9) follows naturally from the non-parametric methods proposed in the previous sections. Different from the odds ratio in Scheike et al. (2010), the denominator in our proposed odds ratio is based on the conditional CIF, instead of the marginal function. To estimate the denominator of the odds ratio in (2.2.9), we notice that

$$P(T \le t, \epsilon = l|x < X \le M) = \frac{P(x < X \le M, T \le t, \epsilon = l)}{F_X(M) - F_X(x)} = \frac{F_l(M,t) - F_l(x,t)}{F_X(M) - F_X(x)}.$$

We can plug in the bivariate estimator $\hat{F}_l$ and Kaplan-Meier estimator for X and hence obtain an appropriate estimator $\hat{\phi}_{(l)}(x,t|M)$ for the time dependent measure.

In addition, we can also combine the information across time by integrating the odds ratios weighted across time. Define the cause-specific global measure

$$\phi^{\star}_{(l)}(M) = \int_{\tau_{x_1}}^{\tau_{x_2}} \int_{\tau_{t_1}}^{\tau_{t_2}} W(u,v)\phi_{(l)}(u,v|M)dudv\Big\{\int_{\tau_{x_1}}^{\tau_{x_2}} \int_{\tau_{t_1}}^{\tau_{t_2}} W(u,v)dudv\Big\}^{-1},$$

where $0 < \tau_{x_1} < \tau_{x_2} < M$, such that the odds ratios are well defined on $[\tau_{x_1}, \tau_{x_2}] \times [\tau_{t_1}, \tau_{t_2}]$. The time periods can be properly selected such that the time dependent odds ratio is summarized over a region of research interest. $W(u,v)$ is a known weight function. Uniform weight is practically useful in interpretation as a simple average over time, though alternative weights can also be adopted to emphasize or deemphasize associations at specific time regions. The measure $\phi^{\star}_{(l)}(M)$ can be naturally estimated by

$$\hat{\phi^{\star}}_{(l)}(M) = \int_{\tau_{x_1}}^{\tau_{x_2}} \int_{\tau_{t_1}}^{\tau_{t_2}} \hat{W}(u,v)\hat{\phi}_{(l)}(u,v|M)dudv\Big\{\int_{\tau_{x_1}}^{\tau_{x_2}} \int_{\tau_{t_1}}^{\tau_{t_2}} \hat{W}(u,v)dudv\Big\}^{-1}.$$

Under the null of independence, $\phi^{\star}_{(l)}(M)$ equals 1. A Wald-type test can be constructed based on the asymptotic normality of the estimator $\hat{\phi^{\star}}_{(l)}(M)$:

$$Z_{\phi^{*}_{(l)}} = \frac{\hat{\phi^{\star}}_{(l)}(M) - 1}{BSE(\hat{\phi^{\star}}_{(l)}(M))},$$

where the standard error of the estimator is again obtained using the bootstrap method. The performances of the tests are evaluated in the following numerical studies.

2.3 SIMULATION STUDIES

In this section, the finite sample performances of the estimators and tests given before are assessed through numerical studies. First we evaluate the proposed estimators for the bivariate and conditional CIFs. The successive event times X and T were generated from the following Clayton copula model (Clayton, 1978):

$$P(X \leq x, T \leq t) = \Big[\big\{F_X(x)\big\}^{-\theta} + \big\{F_T(t)\big\}^{-\theta} - 1\Big]^{-\frac{1}{\theta}},$$

where $F_X(x) = P(X \leq x)$ is the cumulative distribution function (CDF) of X, $F_T(t) = P(T \leq t)$ is the marginal CDF of T, and θ is the association parameter which relates to Kendall's τ (Kendall, 1938), by $\tau = \frac{\theta}{\theta+2}$. We assume that the marginal distributions of X and T both follow a unit exponential distribution, and first consider a scenario by setting $\theta = -0.5$ which corresponds to $\tau = -1/3$. Two hundred pairs of uniform (0,1) marginal variates (U_1, U_2) were generated from the above Clayton model using R function *rcopula*, and X and T were obtained by inverting U_1 and U_2. The cause indicators associated with T were generated from a Bernoulli distribution with probability 0.7, and coded as 1 and 2. Thus, the bivariate CIFs of X and T have the form $F_1(x,t) = 0.7\big[\big\{F_X(x)\big\}^{-\theta} + \big\{F_T(t)\big\}^{-\theta} - 1\big]^{-\frac{1}{\theta}}$, and $F_2(x,t) = 0.3\big[\big\{F_X(x)\big\}^{-\theta} + \big\{F_T(t)\big\}^{-\theta} - 1\big]^{-\frac{1}{\theta}}$. We also consider the scenario that X and T are independent, where 200 pairs of (U_1, U_2) were simulated independently from the uniform(0,1) distribution. Then we followed the same steps as in the first scenario to get X, T and ϵ, assuming unit exponential marginals for X and T. We generated the censoring time C from an independent Uniform(0,4) distribution. The observed times are $Y_1 = \min(X, C)$, $Y_2 = \min\{(X+T), C\}$. The censoring indicators, δ_1, δ_2, are set to be 0 whenever $C < X$ or $C < X + T$, respectively, and $\eta_1 = \delta_1, \eta_2 = \delta_2 \cdot \epsilon$. The censoring proportions for both causes are about 25% for X and about 50% for T. For each scenario, 1000 data sets were generated.

The proposed non-parametric estimators for the bivariate CIFs $F_1(x,t)$ and $F_2(x,t)$ were calculated at $x, t = 0.5, 1.0, 1.5, 2.0$ for the 1000 datasets that we simulated for each scenario. The average of the estimates and the empirical standard error are reported in Table 2.1.

The bootstrap standard errors are obtained based on 250 bootstrap samples, and the mean bootstrap standard errors are also reported in Table 2.1, along with the coverage rates of the 95% asymptotic Wald confidence intervals. The upper panel corresponds to negatively correlated (X, T) from Clayton ($\tau = -\frac{1}{3}$) and the lower panel corresponds to the estimates from the independence case. Our proposed estimates have small biases for both dependent and independent cases. Their empirical standard errors and bootstrap standard errors agree well. In general, the coverage rates are close to the 95% nominal level, except that there is some under coverage for cause 2 at $x, t = 0.5$ for $\tau = -1/3$, which is likely due to the limited number of observed events at early times.

In Table 2.2 we show the estimates of the conditional CIFs as well as the naive univariate estimates for the successive event time $F_{l(T)}$, when (X, T) are discordant ($\tau = -\frac{1}{3}$). The $\hat{F}_l(t|x)$ is referred to as the IPW estimator, in contrast with the naive estimator for the marginal CIF $\hat{F}_{l(T)}(t)$. The latter are computed using the R function *cuminc* in the library cmprsk. Their standard errors can also be obtained by the R function *cuminc*. When t is fixed, the naive estimates would be invariant to the values of x taken by the previous event X. The proposed estimates for the conditional CIFs appear to be unbiased. The naive CIF estimates, however, are seriously biased and their coverage rates are very poor. For example, when $x, t = 0.5$, the naive estimate is more than twice the true value and the corresponding coverage rate is as low as 4.3%. When X and T are negatively correlated, for each t, the true values are increasing as x goes up. Hence we observe that the naive estimator overestimates when x is small and then underestimates the conditional CIF as x gets larger.

The results for independent X and T are summarized in Table 2.3. True values are now the same within each column for the conditional CIFs. The bias of the IPW estimates is still pretty small and the coverage rates are close to 95% for both causes. The naive estimates, as we have expected, agree with the IPW estimates when X and T are independent. Since the naive estimates are using all the data for T regardless of X, they have smaller standard errors than the conditional estimates, especially when x is small. The naive estimates also have excellent coverage rates.

In the second simulation, we evaluate the performance of the tests based on the odds

Table 2.1: Simulation results for the estimates for the bivariate CIFs $F_1(x,t)$ and $F_2(x,t)$ when (X,T) are from Clayton ($\tau = -\frac{1}{3}$) or independent. Ave is the average of the estimates. ESE is the empirical standard error of the estimator. BSE is the average of the bootstrap standard errors. Cov is the 95% coverage rate based on BSE.

$\tau = -\frac{1}{3}$					t					
			0.5		1.0		1.5		2.0	
			$\hat{F_1}$	$\hat{F_2}$	$\hat{F_1}$	$\hat{F_2}$	$\hat{F_1}$	$\hat{F_2}$	$\hat{F_1}$	$\hat{F_2}$
	0.5	True	0.045	0.019	0.125	0.053	0.181	0.078	0.217	0.093
		Ave	0.046	0.019	0.125	0.053	0.181	0.077	0.217	0.092
		ESE	0.017	0.011	0.027	0.017	0.031	0.021	0.035	0.025
		BSE	0.016	0.010	0.026	0.018	0.032	0.022	0.036	0.025
		Cov	0.935	0.837	0.940	0.917	0.946	0.932	0.937	0.923
	1.0	True	0.125	0.053	0.244	0.104	0.320	0.137	0.368	0.157
		Ave	0.125	0.053	0.243	0.104	0.320	0.137	0.368	0.157
		ESE	0.027	0.018	0.035	0.025	0.039	0.028	0.042	0.031
		BSE	0.026	0.018	0.035	0.025	0.040	0.029	0.043	0.032
x		Cov	0.931	0.919	0.947	0.941	0.951	0.946	0.962	0.942
	1.5	True	0.181	0.078	0.320	0.137	0.407	0.175	0.461	0.197
		Ave	0.183	0.078	0.321	0.136	0.408	0.174	0.462	0.196
		ESE	0.032	0.021	0.040	0.028	0.043	0.032	0.046	0.035
		BSE	0.032	0.022	0.040	0.029	0.044	0.033	0.046	0.036
		Cov	0.948	0.949	0.954	0.939	0.953	0.949	0.955	0.936
	2.0	True	0.217	0.093	0.368	0.158	0.461	0.197	0.517	0.222
		Ave	0.219	0.093	0.368	0.157	0.461	0.197	0.518	0.220
		ESE	0.035	0.024	0.043	0.030	0.046	0.035	0.049	0.038
		BSE	0.036	0.025	0.043	0.032	0.046	0.036	0.049	0.038
		Cov	0.951	0.946	0.948	0.948	0.948	0.951	0.957	0.946
Independent					t					
			0.5		1.0		1.5		2.0	
			$\hat{F_1}$	$\hat{F_2}$	$\hat{F_1}$	$\hat{F_2}$	$\hat{F_1}$	$\hat{F_2}$	$\hat{F_1}$	$\hat{F_2}$
	0.5	True	0.108	0.046	0.174	0.075	0.214	0.092	0.238	0.102
		Ave	0.107	0.047	0.173	0.076	0.214	0.092	0.237	0.102
		ESE	0.025	0.016	0.031	0.020	0.034	0.023	0.037	0.025
		BSE	0.023	0.016	0.029	0.020	0.033	0.023	0.035	0.025
		Cov	0.916	0.917	0.928	0.934	0.935	0.933	0.933	0.938
	1.0	True	0.174	0.075	0.279	0.120	0.344	0.147	0.382	0.164
		Ave	0.172	0.075	0.278	0.121	0.342	0.147	0.380	0.164
		ESE	0.031	0.021	0.038	0.026	0.040	0.030	0.044	0.032
		BSE	0.029	0.020	0.036	0.026	0.040	0.029	0.043	0.032
x		Cov	0.921	0.928	0.935	0.939	0.944	0.930	0.934	0.929
	1.5	True	0.214	0.092	0.344	0.147	0.422	0.181	0.470	0.202
		Ave	0.212	0.093	0.342	0.149	0.421	0.182	0.469	0.202
		ESE	0.035	0.023	0.042	0.030	0.043	0.034	0.047	0.036
		BSE	0.033	0.023	0.040	0.029	0.044	0.033	0.047	0.036
		Cov	0.917	0.937	0.940	0.930	0.942	0.936	0.946	0.933
	2.0	True	0.238	0.102	0.382	0.164	0.470	0.202	0.523	0.224
		Ave	0.236	0.103	0.380	0.166	0.469	0.201	0.521	0.224
		ESE	0.037	0.025	0.044	0.032	0.046	0.036	0.051	0.039
		BSE	0.035	0.025	0.043	0.032	0.047	0.036	0.050	0.039
		Cov	0.919	0.939	0.935	0.937	0.942	0.932	0.924	0.936

Table 2.2: Simulation results for the IPW estimator for the conditional CIFs $F_1(t|x)$ and $F_2(t|x)$ and the naive estimator when (X, T) are from Clayton ($\tau = -\frac{1}{3}$). Ave is the average of the estimates. ESE is the empirical standard error of the estimator. BSE(SE) is the average of the bootstrap standard errors or standard errors for the naive estimator. Cov is the 95% coverage rate based on BSE(SE).

			t							
			0.5		1.0		1.5		2.0	
$F_1(t\mid x)$			IPW	Naive	IPW	Naive	IPW	Naive	IPW	Naive
	0.5	True	0.115		0.317		0.460		0.552	
		Ave	0.116	0.240	0.317	0.413	0.459	0.524	0.551	0.593
		ESE	0.041	0.037	0.062	0.045	0.069	0.047	0.074	0.048
		BSE(SE)	0.039	0.036	0.061	0.044	0.070	0.047	0.076	0.049
		Cov	0.928	0.043	0.941	0.432	0.952	0.750	0.946	0.864
	1.0	True	0.198		0.386		0.507		0.582	
		Ave	0.198	0.240	0.385	0.413	0.505	0.524	0.581	0.593
		ESE	0.041	0.037	0.051	0.045	0.054	0.047	0.057	0.048
		BSE(SE)	0.040	0.036	0.051	0.044	0.056	0.047	0.059	0.049
x		Cov	0.939	0.790	0.945	0.894	0.952	0.934	0.956	0.950
	1.5	True	0.233		0.412		0.524		0.593	
		Ave	0.234	0.240	0.412	0.413	0.524	0.524	0.593	0.593
		ESE	0.040	0.037	0.048	0.045	0.051	0.047	0.053	0.048
		BSE(SE)	0.040	0.036	0.048	0.044	0.051	0.047	0.053	0.049
		Cov	0.943	0.949	0.942	0.949	0.940	0.939	0.950	0.955
	2.0	True	0.251		0.425		0.533		0.598	
		Ave	0.253	0.240	0.426	0.413	0.533	0.524	0.599	0.593
		ESE	0.040	0.037	0.047	0.045	0.050	0.047	0.052	0.048
		BSE(SE)	0.040	0.036	0.047	0.044	0.050	0.047	0.052	0.049
		Cov	0.951	0.927	0.945	0.931	0.941	0.931	0.951	0.959
			t							
			0.5		1.0		1.5		2.0	
$F_2(t\mid x)$			IPW	Naive	IPW	Naive	IPW	Naive	IPW	Naive
	0.5	True	0.049		0.136		0.197		0.237	
		Ave	0.049	0.102	0.134	0.175	0.195	0.223	0.233	0.252
		ESE	0.026	0.025	0.043	0.033	0.052	0.037	0.059	0.040
		BSE(SE)	0.025	0.026	0.043	0.034	0.053	0.038	0.059	0.042
		Cov	0.871	0.456	0.925	0.829	0.939	0.928	0.933	0.947
	1.0	True	0.085		0.165		0.217		0.249	
		Ave	0.084	0.102	0.164	0.175	0.217	0.223	0.248	0.252
		ESE	0.028	0.025	0.038	0.033	0.043	0.037	0.047	0.040
		BSE(SE)	0.028	0.026	0.038	0.034	0.044	0.038	0.048	0.042
x		Cov	0.923	0.929	0.935	0.955	0.945	0.963	0.948	0.951
	1.5	True	0.100		0.177		0.225		0.254	
		Ave	0.100	0.102	0.175	0.175	0.223	0.223	0.252	0.252
		ESE	0.027	0.025	0.035	0.033	0.040	0.037	0.043	0.040
		BSE(SE)	0.028	0.026	0.036	0.034	0.041	0.038	0.044	0.042
		Cov	0.949	0.949	0.941	0.950	0.951	0.955	0.943	0.946
	2.0	True	0.108		0.182		0.228		0.256	
		Ave	0.108	0.102	0.181	0.175	0.227	0.223	0.254	0.252
		ESE	0.028	0.025	0.035	0.033	0.039	0.037	0.043	0.040
		BSE(SE)	0.028	0.026	0.036	0.034	0.040	0.038	0.044	0.042
		Cov	0.941	0.928	0.944	0.939	0.952	0.952	0.945	0.940

Table 2.3: Simulation results for the IPW estimator for the conditional CIFs $F_1(t|x)$ and $F_2(t|x)$ and the naive estimator when (X,T) are independent. Ave is the average of the estimates. ESE is the empirical standard error of the estimator. BSE(SE) is the average of the bootstrap standard errors or standard errors for the naive estimator. Cov is the 95% coverage rate based on BSE(SE).

						t				
			0.5		1.0		1.5		2.0	
$F_1(t\vert x)$			IPW	Naive	IPW	Naive	IPW	Naive	IPW	Naive
	0.5	True	0.275		0.442		0.544		0.605	
		Ave	0.273	0.273	0.441	0.441	0.544	0.543	0.603	0.603
		ESE	0.058	0.040	0.068	0.047	0.071	0.047	0.076	0.050
		BSE(SE)	0.054	0.038	0.063	0.045	0.068	0.047	0.072	0.049
		Cov	0.924	0.932	0.932	0.934	0.935	0.948	0.940	0.942
	1.0	True	0.275		0.442		0.544		0.605	
		Ave	0.273	0.273	0.440	0.441	0.542	0.543	0.603	0.603
		ESE	0.046	0.040	0.054	0.047	0.056	0.047	0.060	0.050
		BSE(SE)	0.044	0.038	0.051	0.045	0.055	0.047	0.058	0.049
x		Cov	0.928	0.932	0.930	0.934	0.942	0.948	0.932	0.942
	1.5	True	0.275		0.442		0.544		0.605	
		Ave	0.273	0.273	0.440	0.441	0.542	0.543	0.603	0.603
		ESE	0.043	0.040	0.050	0.047	0.052	0.047	0.055	0.050
		BSE(SE)	0.040	0.038	0.048	0.045	0.051	0.047	0.054	0.049
		Cov	0.921	0.932	0.932	0.934	0.938	0.948	0.937	0.942
	2.0	True	0.275		0.442		0.544		0.605	
		Ave	0.273	0.273	0.441	0.441	0.543	0.543	0.604	0.603
		ESE	0.042	0.040	0.049	0.047	0.051	0.047	0.055	0.050
		BSE(SE)	0.040	0.038	0.047	0.045	0.051	0.047	0.054	0.049
		Cov	0.932	0.932	0.924	0.934	0.940	0.948	0.938	0.942
						t				
			0.5		1.0		1.5		2.0	
$F_2(t\vert x)$			IPW	Naive	IPW	Naive	IPW	Naive	IPW	Naive
	0.5	True	0.118		0.190		0.233		0.260	
		Ave	0.120	0.120	0.192	0.192	0.234	0.234	0.260	0.260
		ESE	0.038	0.027	0.049	0.034	0.055	0.039	0.060	0.042
		BSE(SE)	0.039	0.027	0.049	0.035	0.055	0.039	0.059	0.042
		Cov	0.926	0.948	0.934	0.948	0.942	0.935	0.935	0.945
	1.0	True	0.118		0.190		0.233		0.260	
		Ave	0.119	0.120	0.192	0.192	0.234	0.234	0.260	0.260
		ESE	0.032	0.027	0.040	0.034	0.045	0.039	0.049	0.042
		BSE(SE)	0.031	0.027	0.040	0.035	0.045	0.039	0.048	0.042
x		Cov	0.935	0.948	0.943	0.949	0.934	0.935	0.938	0.945
	1.5	True	0.118		0.190		0.233		0.260	
		Ave	0.120	0.120	0.192	0.192	0.234	0.234	0.260	0.260
		ESE	0.029	0.027	0.037	0.034	0.042	0.039	0.045	0.042
		BSE(SE)	0.029	0.027	0.037	0.035	0.042	0.039	0.045	0.042
		Cov	0.937	0.948	0.940	0.949	0.930	0.935	0.940	0.945
	2.0	True	0.118		0.190		0.233		0.260	
		Ave	0.120	0.120	0.192	0.192	0.233	0.234	0.259	0.260
		ESE	0.028	0.027	0.036	0.034	0.041	0.039	0.044	0.042
		BSE(SE)	0.029	0.027	0.036	0.035	0.041	0.039	0.045	0.042
		Cov	0.938	0.948	0.936	0.949	0.936	0.935	0.937	0.945

Table 2.4: Rejection rates of the 5% level test of independence

		Independent	$\tau = -\frac{4}{5}$	$\tau = -\frac{1}{3}$	$\tau = \frac{1}{3}$	$\tau = \frac{4}{5}$
$n = 200$	Cause 1	0.05	0.73	0.17	0.53	1.00
	Cause 2	0.05	0.43	0.11	0.25	0.77
$n = 400$	Cause 1	0.04	0.99	0.49	0.76	1.00
	Cause 2	0.06	0.72	0.19	0.43	1.00
$n = 600$	Cause 1	0.03	1.00	0.70	0.89	1.00
	Cause 2	0.05	0.87	0.29	0.47	1.00

ratio as proposed in section 2.2.3. The data sets were generated using the same strategy as previously. Weak or strong associations were considered corresponding to $\tau = -\frac{4}{5}, -\frac{1}{3}, \frac{1}{3}, \frac{4}{5}$. We also considered the independent case. Various sample sizes $n = 200, 400, 600$ were evaluated. In computing the odds ratio measures over time, we selected a fixed maximum time point M of 2.1 by which point around 90% of the first events have been observed. A uniform weight function $W(x, t) = I(0.5 \leq x \leq 1.5, 0.5 \leq t \leq 1.5)$ was adopted to avoid extremely early or late event times. The summary statistics $\hat{\phi}^{\star}_{(l)}(2.1), l = 1, 2$ were calculated and transformed to the natural log scale for each data set. The bootstrap standard errors of the $\log(\hat{\phi}^{\star}_{(l)}(2.1))$ were obtained based on 250 bootstrap samples. We rejected the null of independence if the ratio of $|\log(\hat{\phi}^{\star}_{(l)}(2.1))|$ to its bootstrap standard error exceeded 1.96.

The results of the rejection rates based on 1000 realizations are summarized in Table 2.4. Under the independence case, the empirical type I errors are all close to the significance level 5%. Under the alternatives, the power generally increases as the sample size goes up. There is more power under the alternative of positive association compared with negative association. We also note that tests for cause 1 association have more power than for cause 2, which can be explained by the larger number of observed events due to cause 1. Considerable power is present for a strong positive association ($\tau = \frac{4}{5}$) even when the sample size is moderate. When there is strong negative association ($\tau = -\frac{4}{5}$), the power approaches 1 as n reaches 600 for the cause 1 test.

Table 2.5: Estimated bivariate CIFs for the BDCP study (Standard errors are in the parentheses)

	Recurrence of depression (years)				
Recovery (years)	0.5	1.0	1.5	2.0	2.5
1.0	0.185 (0.026)	0.288 (0.033)	0.352 (0.037)	0.352 (0.037)	0.412 (0.070)
2.0	0.228 (0.031)	0.330 (0.035)	0.414 (0.042)	0.414 (0.042)	0.475 (0.073)
	Recurrence of mania/mixed state (years)				
Recovery (years)	0.5	1.0	1.5	2.0	2.5
1.0	0.044 (0.014)	0.064 (0.017)	0.079 (0.020)	0.116 (0.033)	0.116 (0.033)
2.0	0.049 (0.014)	0.088 (0.022)	0.104 (0.024)	0.140 (0.035)	0.140 (0.035)

2.4 THE BDCP STUDY

In this section, we apply our proposed non-parametric estimators to the data from the BDCP study. Bipolar disorder is a disabling, lifelong mental illness that has a strong likelihood of relapse. There has been substantial interest in investigating the association between the recovery and recurrence of bipolar disorder. Two hundred and ninety-nine eligible participants entered the BDCP study in a symptomatic state of bipolar disorder. During the median 1 year of follow-up, 221 patients managed to recover with a median recovery time of 20 weeks since study entry. Those patients continued staying under observation, and some of them experienced subsequent recurrence of new episodes. Among the 221 patients who recovered, we observed new episodes in 99 patients by the end of the study, with 78 depressive episodes and 21 manic or mixed episodes. Following the notation in the previous sections, we code the cause indicator $\epsilon = 1$ for depression and $\epsilon = 2$ for mania or a mixed state, since the cases of new manic or mixed episodes were relatively few in this study. Let X be the recovery time since study entry, and T be the time from recovery to a new episode. The estimates of the bivariate CIFs $F_1(x, t)$ and $F_2(x, t)$, evaluated at $x = 1, 2$ years and $t = 0.5, 1.0, 1.5, 2.0, 2.5$ years, are presented in Table 2.5. The standard errors are based on 3000 bootstrap resamples. For any fixed x and t, the estimated cause 1 CIFs were three or four times larger than the cause 2 CIFs, indicating that the probability of developing a new depressive episode is much higher than developing a manic or mixed episode.

In order to examine the association between the time to recovery and the subsequent recurrence, we now focus on the conditional CIFs, and compare the cumulative incidences of recurrence between those who recovered early and those who recovered late. For those patients who managed to recover early, the cumulative incidences of developing depression and a manic/mixed episode were estimated through $F_1(t|x)$ and $F_2(t|x)$, for some fixed value of x. As for those who took a longer time to recover, we estimated the quantity $P(T \leq t, \epsilon = l | x < X \leq M)$ for each cause using the method proposed in Section 2.2.3, where M was set to be 130 weeks almost covering the entire period of recovery. In Figure 3.1 and Figure 3.2, we present the estimated conditional cumulative incidences of new depressive episodes and manic/mixed episodes. The naive univariate CIF estimates $\hat{F}_{l(T)}(t), l = 1, 2$, regardless of recovery time, were also computed using the R function *cuminc* from the library cmprsk. Two recovery times were considered: $x = 24$ weeks in Figure 3.1 and $x = 36$ weeks in Figure 3.2, corresponding to almost the median recovery time and slightly late recovery time, respectively. For each x, we plot three curves: the conditional CIF for those who recovered earlier, the conditional CIF for those who recovered late, and the naive estimate of the marginal CIF, of a new depressive episode on the left panel and of a new manic/mixed episode on the right. At both $x = 24$ and $x = 36$ weeks, the three curves for new manic/mixed episodes appear to be close to each other. The association between time to recovery and time to a new manic/mixed episode may be fairly weak, or there simply were not enough new manic or mixed episodes before the end of the study to detect the association. For depressive episodes, however, the naive estimates fall between the two conditional CIFs, which suggests that the time to recurrence since recovery might be dependent on the recovery time, where the patients who took a longer time to recover seem to develop new depressive episodes more quickly, as compared to those who recovered early.

Next, we formally evaluate the associations by computing the odds ratio estimates $\hat{\phi}_{(l)}(x, t|M)$ and the 95% confidence intervals for each cause. The standard errors are based on 1000 bootstrap samples. The point estimates and 95% confidence intervals, as well as the 95% simultaneous confidence bands, are shown in Figure 2.3 and Figure 2.4, again for the recovery times of 24 and 36 weeks respectively. The confidence bands were calculated by

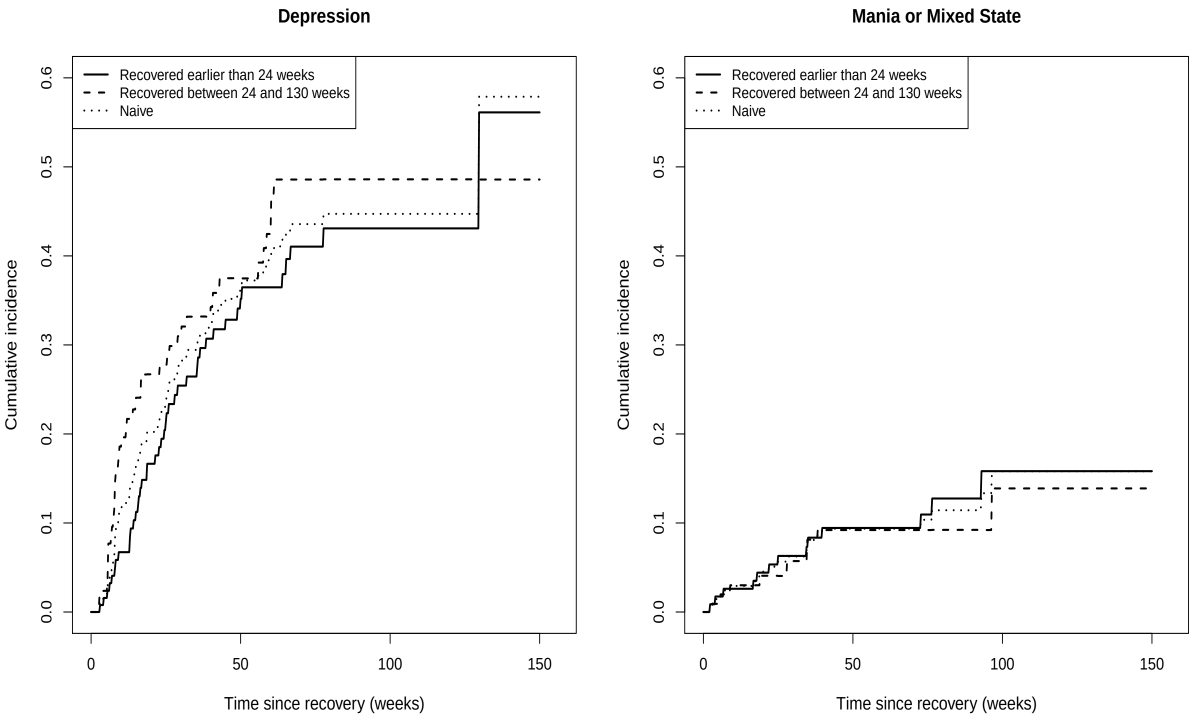

Figure 2.1: Estimates of the cumulative incidence of recurrence of depression and mania/mixed state for the patients who recover earlier or later than 24 weeks, along with the naive univariate CIF estimates.

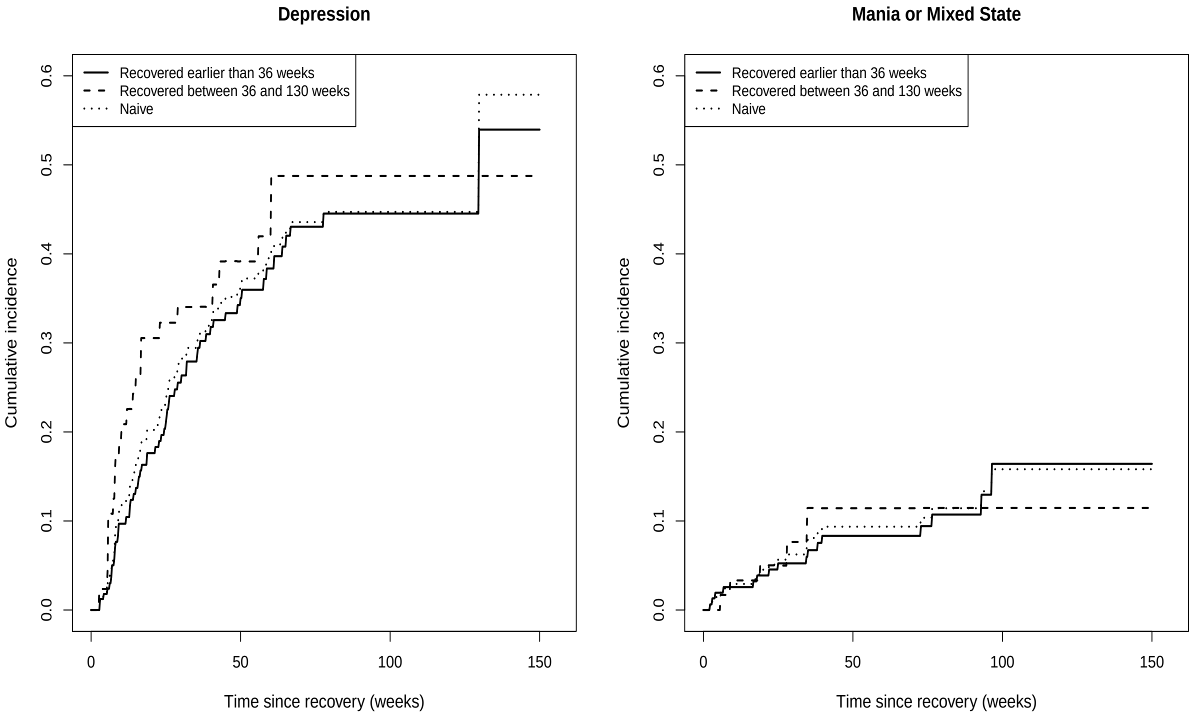

Figure 2.2: Estimates of the cumulative incidence of recurrence of depression and mania/mixed state for the patients who recover earlier or later than 36 weeks, along with the naive univariate CIF estimates.

the following procedure. We first took 1000 bootstrap samples, and computed $\hat{q}_{k(l)}$ for cause l as $\hat{q}_{k(l)} = \sup_{(x,t)} |\hat{\phi}_{k(l)}(x, t|M) - \hat{\phi}_{(l)}(x, t|M)|/\hat{\sigma}(x, t)$, where $\hat{\phi}_{k(l)}$ is the estimate from the kth bootstrap sample and $\hat{\sigma}$ is the bootstrap standard error. Let $\hat{q}_{(l)}$ be the empirical upper 0.95 percentile of all $\hat{q}_{k(l)}$. The cause-specific 95% confidence bands were then constructed as $\hat{\phi}_{(l)}(x, t|M) \pm \hat{q}_{(l)}\hat{\sigma}(x, t)$.

We note that although the point estimates for depression seldom go beyond 1, the confidence intervals are wide and always cover 1. For the other cause, however,we only detect a weaker association. Such a pattern can be explained by the fact that the episodes of mania and mixed states are fairly rare, and it is relatively difficult to observe a subsequent new episode after a long period of recovery given the limited follow-up period of the study. In the plots, simultaneous confidence bands are much wider than the pointwise confidence intervals, especially for mania/mixed state. Similar to the findings from the plots of conditional CIFs in the previous figures, the results imply that recurrence for depression is weakly associated with either average or relatively late recovery times.

We further conduct a formal test for independence using the weighted measure over a meaningful region of time. Only 7 patients managed to recover before 8 weeks and most recoveries occur by one year. Therefore, we focus on x between 8 weeks and one year. For the recurrence time t, we calculate over the period from 15 days to one year, since the first episode was observed at 15 days and there were very few recurrences of depression observed after one year following the recovery among the group of subjects who managed to recover late. We adopt a simple uniform weight on this region and compute $\hat{\phi}^{\star}_{(1)}(M) = 0.801$ for depression and $\hat{\phi}^{\star}_{(2)}(M) = 0.997$ for mania/mixed state. We also compute their 95% confidence intervals. The involved standard errors are based on 1000 bootstrap resamples. The confidence intervals are $(0.386, 1.664)$ and $(0.426, 2.336)$, both covering 1. The results again consistently reveal that there is only a somewhat weak association between recovery and recurrence for both causes.

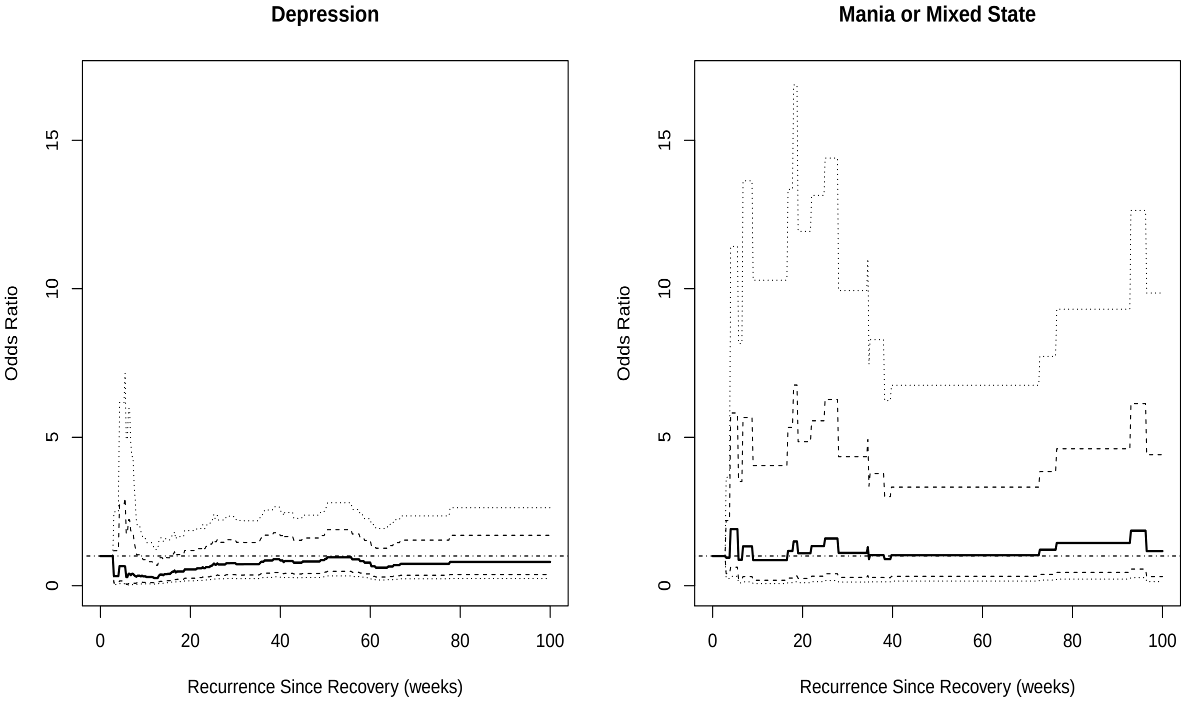

Figure 2.3: Estimated association using odds ratio $\hat{\phi}_{(l)}(24, t|M)$ for depression and mania/mixed state. (— , point estimate; - - - , 95% pointwise confidence interval; $\cdots$, 95% confidence band; horizontal line is odds ratio equal to 1)

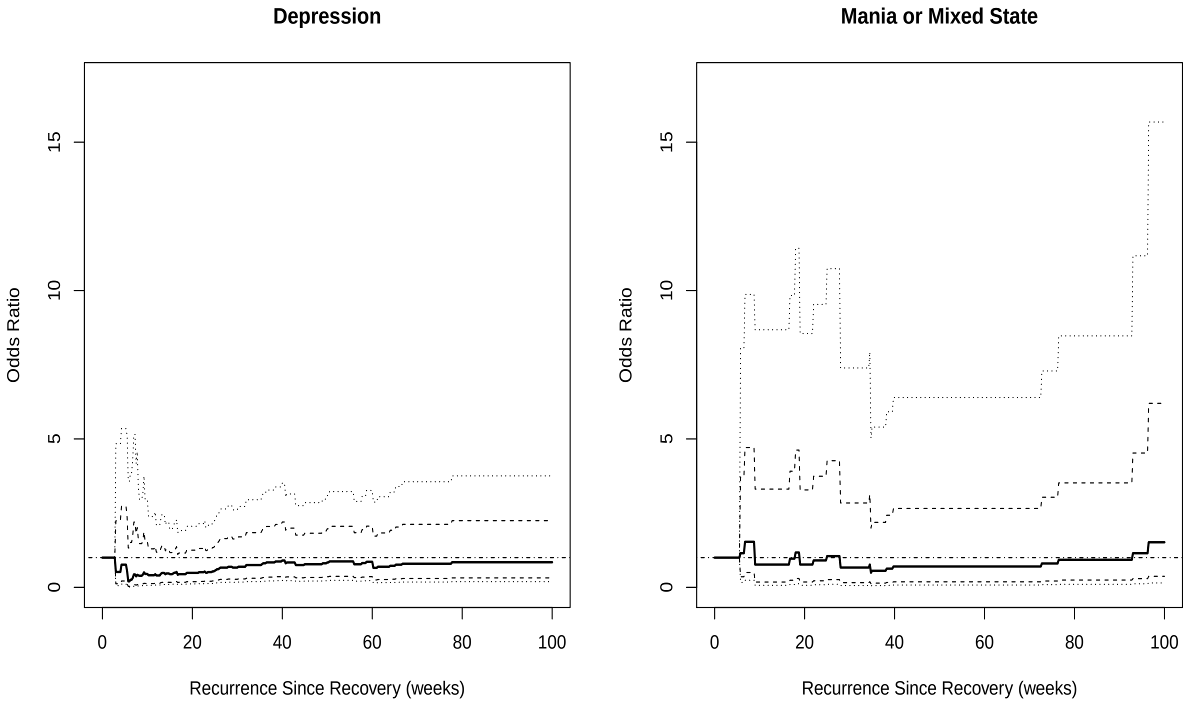

Figure 2.4: Estimated association using odds ratio $\hat{\phi}_{(l)}(36, t|M)$ for depression and mania/mixed state. (— , point estimate; - - - , 95% pointwise confidence interval; $\cdots$, 95% confidence band; horizontal line is odds ratio equal to 1)

2.5 REMARKS

There are other approaches to quantifying the time-varying association for general bivariate competing risks data (Cheng et al., 2007; Bandeen-Roche and Ning, 2008; Scheike et al., 2010). Quantifying the association usually requires the estimation of marginal distributions. However, in the successive events framework, the non-identifiability of the marginal distribution of the second gap time raises an issue for directly using those existing methods. Instead, we construct the association measure based on the conditional CIFs. Moreover, unlike the instantaneous measures considered in (Cheng et al., 2007; Bandeen-Roche and Ning, 2008), we adopt the odds ratio measure which is well accepted by practitioners. Based on the integrated odds ratio, we have developed a test for cause-specific dependence between gap times and that could be equivalently achieved by testing the Markov assumption if multistate models were utilized. The findings of the simulation studies suggest that our proposed test has sufficient power for moderate sample sizes when the failure cause is not very rare.

In our application to the BDCP study, we have been focusing on evaluating the associations at some specific time points as well as summarizing the overall dependence by calculating a simple average of the time-dependent odds ratio since the associations are fairly weak across time. However, in some applications, non-uniform weight functions may be preferred and some optimal weight function may be derived. Furthermore, our current work only considers competing causes for the subsequent event. The association framework may be extended to the case where the first event is also subject to competing-risk censoring, which will let the multistate model become more relevant. This is a topic of future research.

3.0 DIAGNOSTIC ACCURACY ANALYSIS FOR COMPETING RISKS OUTCOME WITH A CENSORED MARKER

3.1 INTRODUCTION

The receiver operating characteristic (ROC) curve is routinely used to evaluate the classification performance of a marker in medical studies. Methodologists have been interested in quantifying the classification and prediction accuracy of markers for time-dependent diagnostic outcomes in the past decades. As proposed by Heagerty et al. (2000), the traditional ROC analysis with binary responses was extended to the time-dependent framework with survival outcomes, where they considered a time-varying disease status for accuracy studies, which allows evaluation for the time-dependent diagnostic abilities of the markers to discriminate the subjects that are likely or unlikely to fail before a certain time point. The resulting time-dependent ROC curve, which is the plot of true positive rate (sensitivity) versus the false positive rate (one minus specificity) over time, has a significant role in diagnostic medicine as a number of associated accuracy measures provide direct clinical interpretations. Consequently, there has been extensive work exploring the use of ROC approaches for time-to-event outcomes with a continuous marker. For example, Zheng and Heagerty (2004), Cai et al. (2006) and Zheng and Heagerty (2007) studied the prognostic ability of longitudinal markers; Heagerty and Zheng (2005) proposed various definitions for accuracy measures and considered estimation based on Cox models; Cai and Cheng (2008) developed robust procedures for combining multiple makers using a composite score; Cai et al. (2011) utilized a meta-analysis approach for t-year survival prediction; the prognostic accuracy of markers in nested case-control studies was considered by Cai and Zheng (2011); Li and Ma (2011) investigated time-dependent ROC approaches under diverse censoring mechanisms.

In practice, time-to-event outcomes are often times subject to independent censoring, which has been properly handled by the existing methods. For the markers, a common assumption in most of these methods is that they can be fully observed in the sample. In the Bipolar Disorder Center for Pennsylvanians (BDCP) Study (Fagiolini et al., 2009), however, there are multiple endpoints such as depressive, manic, hypomanic and mixed episodes. When the primary focus is on the time to a cause-specific first episode, it will be subject to competing-risk censoring by other types of episodes. The number of previously experienced episodes, as a novel instrument marker at baseline, might be worth investigating for its discriminatory ability because of the historical information it carries for the subjects. Since this measure is self-reported, it may be difficult to recall if the subject has experienced a very large number of episodes. Hence, in practice it is more reasonable to censor these "too many to count" at some large value, e.g. 50. Therefore, it would be of clinical interest to study this situation where both the marker and the outcome are censored and the outcome is also subject to competing-risk censoring. To our best knowledge, no existing method is available for such a problem.

The previous literature on estimating the sensitivity and specificity with typical survival outcomes with independent censoring have focused on estimating the conditional probabilities directly (Uno et al., 2007) , or estimating the reverse conditional probabilities and then applying Bayes' theorem (Heagerty et al., 2000; Chambless and Diao, 2006; Zheng and Heagerty, 2007; Song and Zhou, 2008; Saha and Heagerty, 2010). See Blanche et al. (2013b) for a review and discussion. For example, the sensitivity for a threshold marker value at a specific time point can be estimated as the proportion of subjects with a higher marker value among all subjects who have failed from the event of interest by that time. On the other hand, one can first estimate the conditional cumulative distribution function of the event time at that time point, given that the marker value is higher than the threshold, and then use Bayes' theorem to obtain the reverse conditional probability. Some of these methods have been extended to handle an outcome that is subject to competing-risk censoring (Saha and Heagerty, 2010; Foucher et al., 2010). However, these methods can only work with a continuous marker without censoring.

To deal with the censoring for the marker, there has been some work on ROC-based inference for traditional binary outcomes in the presence of detection limit for markers (Perkins et al., 2007; Mumford et al., 2006; Perkins et al., 2009; Jafarzadeh et al., 2010; Perkins et al., 2011). In the time-dependent ROC framework, Cheng and Li (2015) proposed an estimating strategy that allows the markers to be censored. However, their method requires independent censoring for both the marker and the outcome, and hence cannot be applied to analyze the BDCP data where the event of interest is subject to competing-risk censoring.

In this chapter, we will develop estimation methods for time-dependent accuracy measures when the markers are subject to censoring and the outcome is subject to competing-risk censoring. Two estimators are proposed. The first is a simple plug-in estimator based on bivariate and univariate survival functions and cumulative incidence functions(CIF). The univariate Kaplan-Meier estimator and Dabrowska estimator (Dabrowska, 1988) of the bivariate survival function, along with the bivariate CIF estimator by Cheng et al. (2007) are utilized to construct the time-dependent accuracy measure estimators. Alternatively, we also consider an inverse probability weighting (IPW) method for estimation. In the presence of independent censoring as is typical for survival outcomes, the inverse probability weighting approach has been well adopted in the literature of time-varying diagnostic accuracy measure estimation. The subjects who have the event before a certain time are inversely weighted by the probability of being observed to compensate for those subjects who have been censored (Uno et al., 2007; Cai and Cheng, 2008; Hung and Chiang, 2010a,b); or the additional number of expected events is calculated for censored observations (Wolf et al., 2011). We will extend the IPW method to both the censored marker and the competing-risk censored outcomes.

The remainder of the chapter is organized as follows. In Section 3.2, we will give details on the two estimators that we have just described and also develop cause-specific tests for comparing the discriminatory abilities of two markers. Simulation studies will be presented and discussed in Section 3.3. The new methods will be applied to analyze the the Bipolar Disorder Center for Pennsylvanians (BDCP) Study in Section 3.4. The chapter concludes with some discussion in Section 3.5.

3.2 METHOD

Assume without loss of generality that there is only one competing event, as multiple competing events can be grouped together into one. As in a typical competing risks setting, we observe time to first event T and the corresponding cause indicator $\epsilon = 1$ or 2. We denote the continuous marker to be Y, and assume T and Y are random variables defined on proper measurable spaces. As is the convention, sensitivity is the probability of a positive test associated with a high marker value for subjects in the diseased group, and specificity is the probability of a negative test with a low marker value for subjects in the healthy group. For a specific event time t, the cause-specific time-dependent sensitivity using the marker Y at a decision threshold y is defined for cause k as

$$\texttt{se}_t^{(k)}(y) = P(Y \geq y | T \leq t, \epsilon = k), \tag{3.2.1}$$

The set $\{Y \geq y\}$ indicates that a positive diagnosis is made. Following to the terminology from Heagerty and Zheng (2005), the event $\{T \leq t, \epsilon = k\}$ is called the *cumulative* case, indicating that cause k failure occurs at or before time t and corresponds to a disease-present status. In the presence of competing-risk censoring, some consideration is required in defining the control group, since some subjects may have also failed from the competing event by time t. In this chapter, we define the cause-specific specificity as

$$\texttt{sp}_t^{(k)}(y) = P(Y < y | \{T \leq t, \epsilon = k\}^c), \tag{3.2.2}$$

where $\{T \leq t, \epsilon = k\}^c$, the complement of $\{T \leq t, \epsilon = k\}$, represents the disease-free status and is analogous to the concept of *dynamic* control group (Heagerty and Zheng, 2005). It includes the subjects who have not had any event by time t, as well as those who have failed at or before t due to the competing event. Among the existing literature, Zheng et al. (2012) and

Blanche et al. (2013a) considered this definition of disease-free group in constructing cause-specific specificity; Shi et al. (2014) assessed accuracy improvement in prediction models also based on this definition. In the meanwhile, there are other proposals for the disease-free group. Foucher et al. (2010) defined the specificity as $P(Y \leq y|T \geq t, \epsilon = k)$ which is less intuitive as $P(T \geq t, \epsilon = k)$ is nonparametrically nonidentifiable. Pepe et al. (2008) and Saha and Heagerty (2010) suggested using $P(Y \leq y|T > t)$ as the specificity, where the control group is free of any event. In this chapter, we will use $\{T \leq t, \epsilon = k\}^c$ as the definition for the healthy group for cause k and hence (3.2.2) as the cause-specific specificity. Note that the proposed estimating strategy in the following sections is also applicable to those alternative definitions.

The plot of the cause-specific true-positive classification rate $\mathtt{se}_t^{(k)}(y)$ versus the false-positive rate $1 - \mathtt{sp}_t^{(k)}(y)$ across all possible threshold y values yields the time-specific ROC curve. The ROC curve is a monotone increasing function in $(0, 1)$, given by

$$R_t^{(k)}(p) = \mathtt{se}_t^{(k)}\{[\mathtt{sp}_t^{(k)}]^{-1}(1-p)\}, 0 < p < 1$$

where $[\mathtt{sp}_t^{(k)}]^{-1}(p) = \inf\{y : \mathtt{sp}_t^{(k)}(y) \geq p\}$. The area under the ROC curve (AUC) calculated at a specific time, as a summary measure, is defined to be an integral of the ROC function:

$$A^{(k)}(t) = \int_0^1 R_t^{(k)}(p)\, d\, p,$$

which quantifies the discriminatory capacity of the marker. Values of AUC close to 1 indicate that the diagnostic marker has high discriminatory accuracy. It describes the probability that a randomly chosen person with cause k failure at or before t has higher maker value than that of a person randomly chosen from the control group with respect to cause k.

In many applications, no a priori time is specified. Under such circumstance, a summarized accuracy is then desired to characterize a marker's prognostic potential. The integrated AUC (IAUC) over the time range of interest can be used to measure the overall prognostic performance. We restrict attention to a specific follow-up period $[\tau_1, \tau_2]$. The lower bound of this interval is set to be $\tau_1 > 0$ as we are seldom interested in an event exactly at baseline $t = 0$ and the cause-specific sensitivity at baseline may not be be well defined. The upper bound is determined by the identifiability of the measures in the presence of censoring. The integrated summary measure is thus given by

$$S^{(k)} = \int_{\tau_1}^{\tau_2} A^{(k)}(t) \cdot d\,\omega(t),$$

where ω is some known weight measure.

Another set of time-dependent prognostic accuracy measures which may be central to clinical decision making are the positive and negative predictive values. They are defined for cause k as

$$\begin{aligned} \texttt{PPV}_t^{(k)}(y) &= P(T \leq t, \epsilon = k | Y \geq y) \\ \texttt{NPV}_t^{(k)}(y) &= P(\{T \leq t, \epsilon = k\}^c | Y < y), \end{aligned}$$

where we adopt consistent rules to identify the disease status of subjects.

3.2.1 Estimating Cause-specific Accuracy Measures

In this section, we develop two approaches for the estimation of the cause-specific accuracy measures. In the first method, we propose a plug-in estimator by utilizing the existing univariate and bivariate estimators for the CIF and survival function. The second estimator, however, is constructed based on an inverse probability weighting method.

As is defined before, let T_i denote the time to the event of clinical interest for the ith individual, which is subject to two distinct competing causes denoted as $\epsilon = 1, 2$. Due to the independent censoring imposed by (C_{Y_i}, C_{T_i}), for the marker Y_i and the time T_i, we observe $\mathcal{D} = \{(W_i, \xi_i, X_i, \delta_i), i = 1, \cdots, n\}$ for the cohort of n subjects followed prospectively, where $W_i = \min(Y_i, C_{Y_i}), X_i = \min(T_i, C_{T_i})$ and $\xi_i = I(Y_i \leq C_{Y_i})$, $\delta_i = \epsilon \cdot I(T_i \leq C_{T_i})$.

3.2.1.1 A plug-in estimator We can rewrite the cause-specific sensitivity $\texttt{se}_t^{(k)}(y)$ and specificity $\texttt{sp}_t^{(k)}(y)$ as

$$\begin{aligned} P(Y \geq y | T \leq t, \epsilon = k) &= \frac{P(T \leq t, \epsilon = k) - P(Y < y, T \leq t, \epsilon = k)}{P(T \leq t, \epsilon = k)} \\ &= \frac{F_T^{(k)}(t) - F_{Y,T}^{(k)}(y-, t)}{F_T^{(k)}(t)}, \end{aligned}$$

and

$$\begin{aligned} P(Y < y | \{T \leq t, \epsilon = k\}^c) &= \frac{P(Y < y) - P(Y < y, T \leq t, \epsilon = k)}{1 - P(T \leq t, \epsilon = k)} \\ &= \frac{F_Y(y-) - F_{Y,T}^{(k)}(y-, t)}{1 - F_T^{(k)}(t)}. \end{aligned}$$

Similarly, the positive and negative predictive values can be rewritten as

$$\texttt{PPV}_t^{(k)}(y) = \frac{F_T^{(k)}(t) - F_{Y,T}^{(k)}(y-, t)}{S_Y(y-)},$$

$$\texttt{NPV}_t^{(k)}(y) = \frac{F_Y(y-) - F_{Y,T}^{(k)}(y-, t)}{F_Y(y-)},$$

where $y-$ is the value just prior to y. $F_T^{(k)}$ is the marginal CIF of T for cause k. The marginal cumulative distribution function and survival function of Y are denoted as F_Y and S_Y. The bivariate CIF for cause k is defined as $F_{Y,T}^{(k)}(y, t) = P(Y \leq y, T \leq t, \epsilon = k)$. Only the outcome of interest T is subject to competing-risk censoring. We will show that the methods used in the literature of association analysis of bivariate competing risks data can be borrowed to estimate the cause-specific measures.

With censored markers, the estimator for the marginal cumulative distribution function $\hat{F}_Y(y)$ can be obtained by one minus the familiar Kaplan-Meier estimator for the survival function $\hat{S}_Y(y)$. The marginal CIF $F_T^{(k)}(t)$ will be estimated as the integral of the product of a Nelson-Aalen type estimator for the cause specific hazard function and the Kaplan-Meier estimator for the overall survival function $S_T(t) = P(T > t)$:

$$\hat{F}_T^{(k)}(t) = \int_0^t \hat{S}_T(u-) H_T^{-1}(u) \, dN_T^{(k)}(u),$$

where $N_T^{(k)}(t) = \sum_{i=1}^n I(X_i \leq t, \delta_i = k)$ and $H_T(t) = \sum_{i=1}^n I(X_i \geq t)$.

We adopt the nonparametric estimator provided by Cheng et al. (2007) to estimate the bivariate CIF, $F_{Y,T}^{(k)}(y, t)$. More specifically, we define the at-risk empirical process $H_{Y,T}(y, t) = \sum_{i=1}^n I(W_i \geq y, X_i \geq t)$ and the double-event empirical process $N_{Y,T}^{(k)}(y, t) = \sum_{i=1}^n I(W_i \leq y, \xi_i = 1, X_i \leq t, \delta_i = k)$. A plug-in estimator for the biviriate CIF is given by

$$\hat{F}_{Y,T}^{(k)}(y, t) = \int_0^y \int_0^t \hat{S}_{Y,T}(u-, v-) \hat{\Lambda}^{(k)}(du, dv)$$

based on the estimated bivariate cumulative cause-specific hazard function

$$\hat{\Lambda}^{(k)}(y,t) = \int_0^y \int_0^t N_{Y,T}^{(k)}(du,dv)\{H_{Y,T}(u,v)\}^{-1}$$

and $\hat{S}_{Y,T}$ is a Dabrowska estimator (Dabrowska, 1988) of the overall bivariate survivor function.

Next, the resulting estimates of the time-dependent measures can be obtained by plugging in the corresponding nonparametric estimators. The cause-specific sensitivity and specificity are estimated as

$$\hat{\texttt{se}}_t^{(k)}(y) = \frac{\hat{F}_T^{(k)}(t) - \hat{F}_{Y,T}^{(k)}(y-,t)}{\hat{F}_T^{(k)}(t)},$$

$$\hat{\texttt{sp}}_t^{(k)}(y) = \frac{\hat{F}_Y(y-) - \hat{F}_{Y,T}^{(k)}(y-,t)}{1-\hat{F}_T^{(k)}(t)}.$$

The prospective measures PPV and NPV are estimated as

$$\hat{\texttt{PPV}}_t^{(k)}(y) = \frac{\hat{F}_T^{(k)}(t) - \hat{F}_{Y,T}^{(k)}(y-,t)}{\hat{S}_Y(y-)},$$

$$\hat{\texttt{NPV}}_t^{(k)}(y) = \frac{\hat{F}_Y(y-) - \hat{F}_{Y,T}^{(k)}(y-,t)}{\hat{F}_Y(y-)}.$$

Subsequently the time-dependent ROC curve can be estimated for each cause by

$$\hat{R}_t^{(k)}(p) = \hat{\texttt{se}}_t^{(k)}\{[\hat{\texttt{sp}}_t^{(k)}]^{-1}(1-p)\}.$$

The time-dependent AUC can be estimated by $\hat{A}^{(k)}(t) = \int_0^1 \hat{R}_t^{(k)}(p)\,dp$. Finally, the cause-specific integrated AUC can be estimated by

$$\hat{S}^{(k)} = \int_{\tau_1}^{\tau_2} \hat{A}^{(k)}(t) \cdot d\,\omega(t).$$

3.2.1.2 An inverse probability weighting estimator Now we consider another way to estimate the accuracy measures. Since the cause-specific sensitivity and specificity can also be written as

$$\texttt{se}_t^{(k)}(y) = \frac{P(Y \geq y, T \leq t, \epsilon = k)}{P(T \leq t, \epsilon = k)},$$

and

$$\begin{aligned}\texttt{sp}_t^{(k)}(y) &= \frac{P(Y < y, \{T \leq t, \epsilon = k\}^c)}{P(\{T \leq t, \epsilon = k\}^c)} \\ &= \frac{P(Y < y, T > t) + P(Y < y, T \leq t, \epsilon = 3 - k)}{P(T > t) + P(T \leq t, \epsilon = 3 - k)},\end{aligned}$$

where $k = 1, 2$, by adopting inverse probability weighting method, an alternative estimator for the sensitivity, $\widetilde{\texttt{se}}_t^{(k)}(y)$, would be

$$\frac{\sum_{i=1}^n \{[\hat{G}(y-, X_i-)]^{-1} I(W_i \geq y, X_i \leq t, \delta_i = k)\}}{\sum_{i=1}^n \{[\hat{S}_{C_T}(X_i-)]^{-1} I(X_i \leq t, \delta_i = k)\}},$$

and the specificity can be estimated by $\widetilde{\texttt{sp}}_t^{(k)}(y)$, which is

$$\frac{\sum_{i=1}^n \{[\hat{G}(W_i-, t)]^{-1} I(W_i < y, \xi_i = 1, X_i > t) + [\hat{G}(W_i-, X_i-)]^{-1} I(W_i < y, \xi_i = 1, X_i \leq t, \delta_i = 3 - k)\}}{\sum_{i=1}^n \{[\hat{S}_{C_T}(t)]^{-1} I(X_i > t) + [\hat{S}_{C_T}(X_i-)]^{-1} I(X_i \leq t, \delta_i = 3 - k)\}},$$

where $\hat{G}$ is the Dabrowska estimator for the bivariate survival distribution of the censoring times C_Y and C_T. We use $\hat{S}_{C_T}$ and $\hat{S}_{C_Y}$ to denote the Kaplan-Meier estimators for the univariate survival functions of C_T and C_Y, respectively. These estimators are calculated based on $\{(W_i, 1 - \xi_i, X_i, I(\delta_i = 0)), i = 1, \cdots, n\}$.

The prospective measures PPV and NPV would be estimated as $\widetilde{\texttt{PPV}}_t^{(k)}(y)$ and $\widetilde{\texttt{NPV}}_t^{(k)}(y)$, respectively, which are

$$\frac{\sum_{i=1}^n \{[\hat{G}(y-, X_i-)]^{-1} I(W_i \geq y, X_i \leq t, \delta_i = k)\}}{\sum_{i=1}^n \{[\hat{S}_{C_Y}(y-)]^{-1} I(W_i \geq y)\}},$$

and

$$\frac{\sum_{i=1}^n \{[\hat{G}(W_i-, t)]^{-1} I(W_i < y, \xi_i = 1, X_i > t) + [\hat{G}(W_i-, X_i-)]^{-1} I(W_i < y, \xi_i = 1, X_i \leq t, \delta_i = 3 - k)\}}{\sum_{i=1}^n \{[\hat{S}_{C_Y}(W_i-)]^{-1} I(W_i < y, \xi_i = 1)\}}.$$

The corresponding ROC curve is estimated as $\widetilde{R}_t^{(k)}(p) = \widetilde{\texttt{se}}_t^{(k)}\{[\widetilde{\texttt{sp}}_t^{(k)}]^{-1}(1 - p)\}$. The AUC estimate is $\widetilde{A}^{(k)}(t) = \int_0^1 \widetilde{R}_t^{(k)}(p)\, dp$. The IAUC is estimated by $\widetilde{S}^{(k)} = \int_{\tau_1}^{\tau_2} \widetilde{A}^{(k)}(t) \cdot d\,\omega(t)$.

3.2.2 Inference

We study the large sample properties of the proposed cause-specific estimators for the accuracy measures. The asymptotic normality can be established using empirical process techniques; details are given in the Appendix. Therefore, statistical inference based on the normal approximation naturally follows from such theoretical results. However, the covariance structures may involve very complicated forms that are difficult to estimate. We employ the bootstrap approach for the inference in this work.

We can construct the pointwise asymptotic confidence intervals for the time-dependent accuracy measures. Specifically, we may take random samples with replacement repeatedly from the data $\mathcal{D} = \{(W_i, \xi_i, X_i, \delta_i), i = 1, \cdots, n\}$ to obtain B bootstrap samples $\mathcal{D}^b$ where $b = 1, \cdots, B$. We then estimate the accuracy measures from each bootstrap sample $\mathcal{D}^b$. The standard deviations of the proposed accuracy estimators are attained by calculating the sample standard deviations of the B bootstrap estimates respectively. The asymptotic confidence intervals for the accuracy measures can be constructed based on asymptotic normality. For instance, a $(1-\alpha)100\%$ confidence interval for $\mathtt{se}_t^{(k)}(y)$ can be obtained by $\hat{\mathtt{se}}_t^{(k)}(y) \pm z_{\alpha/2}\hat{SD}(\hat{\mathtt{se}}_t^{(k)}(y))$, where $\hat{SD}(\hat{\mathtt{se}}_t^{(k)}(y))$ is the estimate of the standard deviation based on the above bootstrap method and $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

In clinical studies, interests may lie on the comparison of two markers in their diagnostic potentials. Therefore, cause-specific tests can be conducted for each type of event. When a specific time point t is of scientific interest, for example, a year free of depression in the bipolar disorder study, it is worthwhile to test the following hypotheses for cause k:

$$H_0 : A_1^{(k)}(t) = A_2^{(k)}(t), v.s. H_1 : A_1^{(k)}(t) \neq A_2^{(k)}(t),$$

where $A_1^{(k)}(t)$ and $A_2^{(k)}(t)$ are the AUCs for the two markers at time t. We may construct a Wald test based on the bootstrap procedures. Note that the data now contain two markers Y_1, Y_2, and we observe $\mathcal{D} = \{(W_{1i}, \xi_{1i}, W_{2i}, \xi_{2i}, X_i, \delta_i), i = 1, \cdots, n\}$. Based on the plug-in (CFK) estimator and the inverse probability weighting (IPW) estimator, we can calculate the test statistics

$$Z_{CFK}^{(k)}(t) = \frac{\hat{A}_1^{(k)}(t) - \hat{A}_2^{(k)}(t)}{\hat{SD}(\hat{A}_1^{(k)}(t) - \hat{A}_2^{(k)}(t))},$$

and

$$Z_{IPW}^{(k)}(t) = \frac{\tilde{A}_1^{(k)}(t) - \tilde{A}_2^{(k)}(t)}{\tilde{SD}(\tilde{A}_1^{(k)}(t) - \tilde{A}_2^{(k)}(t))},$$

where $\hat{SD}(\hat{A}_1^{(k)}(t) - \hat{A}_2^{(k)}(t))$ and $\tilde{SD}(\tilde{A}_1^{(k)}(t) - \tilde{A}_2^{(k)}(t))$ are the bootstrap estimates of the standard deviation. Significance at level α is claimed if the absolute value of the test statistic is greater than $z_{\alpha/2}$.

In addition, a comparison for the overall discriminatory capacity of two markers across a range of time is also very likely to be of interest. Suppose $S_1^{(k)}$ and $S_2^{(k)}$ are the integrated summary measure for the common cause of failure k over the time period $[\tau_1, \tau_2]$. We can test the hypothesis

$$H_0 : S_1^{(k)} = S_2^{(k)}, v.s. H_1 : S_1^{(k)} \neq S_2^{(k)}.$$

Based on bootstrap procedures for calculating the standard error and in a similar fashion to the time-specific test, the test statistics can be constructed as

$$Z_{CFK}^{\star(k)} = \frac{\hat{S}_1^{(k)} - \hat{S}_2^{(k)}}{\hat{SD}(\hat{S}_1^{(k)} - \hat{S}_2^{(k)})}.$$

and

$$Z_{IPW}^{\star(k)} = \frac{\tilde{S}_1^{(k)} - \tilde{S}_2^{(k)}}{\tilde{SD}(\tilde{S}_1^{(k)} - \tilde{S}_2^{(k)})}.$$

We will reject the null hypothesis and conclude that the overall discriminatory abilities of the two markers are significantly different for cause k failure if the absolute value of the test statistic exceeds $z_{\alpha/2}$.

3.3 SIMULATION

In this section, we illustrate the finite-sample performance of the proposed estimators for accuracy measures and the tests for comparing two markers' discriminatory capabilities through simulation studies. First, we evaluate estimators for the cause-specific sensitivity

$\mathtt{se}_t^{(k)}(y)$, specificity $\mathtt{sp}_t^{(k)}(y)$, positive predictive value $\mathtt{PPV}_t^{(k)}(y)$, the negative predictive value $\mathtt{NPV}_t^{(k)}(y)$, as well as the AUC $A^{(k)}(t)$ and the IAUC $S^{(k)}$. Similar strategies are borrowed from Cheng and Li (2015) to simulate the datasets. We generate the marker values Y as $\min(|2Z|, 7)$, where Z is a standard normal random variable. We assume there are two causes of failures and the marker is correlated with the time to the first event through a log-linear model $\log(T) = \beta_0 + \beta_1 Y + \sigma e$, where e is from standard normal distribution. We set $\beta_0 = 1$, $\beta_1 = -0.6$ and $\sigma = 0.3$. The cause indicator ϵ is generated from a Bernoulli distribution, taking values 1 or 2 with probability of 0.6 and 0.4. The censoring variables C_Y and C_T are both generated independently from uniform distributions. Two scenarios of censoring rates are considered: $C_Y \sim Uniform(0, 5)$ and $C_T \sim Uniform(0.5, 5)$ imposing about 30% and 20% censoring on the marker Y and the event time T, respectively; while $C_Y \sim Uniform(0, 3)$ and $C_T \sim Uniform(0.5, 2.5)$ leading to around 50% censoring on Y and 40% censoring on T. A sample of 300 markers and event times is generated in each of the 2000 simulated datasets.

In table 3.1 are the simulation results for the scenario of 30% censoring on Y and 20% censoring on T. For both the CFK estimator and the IPW estimator, we present the average of the estimates, the empirical and the bootstrap standard errors, together with the 95% coverage rate based on the bootstrap error. We take 250 bootstrap samples to calculate the standard errors of the estimates. The estimated accuracy measures are evaluated at $y = 1.5, 2.0$ and $t = 1.0, 2.0$. The estimated IAUC is computed using a uniform weight between the 10% and 90% quantiles of T. We note that, generally, all estimators have small bias; the bootstrap standard errors agree with the empirical standard errors; the coverage rates are close to the nominal level for both causes. However, the IPW estimates have slightly larger bias and standard errors compared with the CFK estimates. Table 3.2 summarized the results in a heavier censoring scenario, where 50% of the marker values and 40% event times are censored. The estimators perform well in terms of bias and coverage rates, although their perfomance in general is worse than that under the light-censoring case due to fewer events under heavier censoring. It is worth noting that, when the accuracy measures approach very close to one, the IPW estimators tend to outperform the CFK estimators, giving better interval estimates.

Table 3.1: Simulation results for the estimates of $\texttt{se}_t^{(k)}(y)$, $\texttt{sp}_t^{(k)}(y)$, $\texttt{PPV}_t^{(k)}(y)$, $\texttt{NPV}_t^{(k)}(y)$, $A^{(k)}(t)$ and $S^{(k)}$, when censoring rates are 30% on markers and 20% on outcomes. Ave is the average of the estimates. ESE is the empirical standard error of the estimator. BSE is the average of the bootstrap standard errors. Cov is the 95% coverage rate based on BSE.

Cause 1	True	Ave		ESE		BSE		Cov	
		CFK	IPW	CFK	IPW	CFK	IPW	CFK	IPW
$SE_{1.0}(1.5)$	0.900	0.904	0.894	0.042	0.061	0.040	0.057	0.893	0.903
$SE_{1.0}(2.0)$	0.720	0.728	0.725	0.069	0.083	0.068	0.081	0.932	0.927
$SE_{2.0}(1.5)$	0.575	0.581	0.577	0.055	0.053	0.055	0.055	0.946	0.952
$SE_{2.0}(2.0)$	0.404	0.414	0.406	0.061	0.055	0.061	0.055	0.948	0.949
$SP_{1.0}(1.5)$	0.697	0.694	0.778	0.036	0.144	0.036	0.143	0.950	0.839
$SP_{1.0}(2.0)$	0.818	0.816	0.860	0.036	0.120	0.036	0.116	0.948	0.984
$SP_{2.0}(1.5)$	0.655	0.657	0.746	0.056	0.138	0.055	0.135	0.949	0.847
$SP_{2.0}(2.0)$	0.760	0.764	0.835	0.060	0.122	0.059	0.113	0.948	0.826
$PPV_{1.0}(1.5)$	0.500	0.499	0.497	0.047	0.051	0.047	0.052	0.953	0.950
$PPV_{1.0}(2.0)$	0.571	0.578	0.570	0.065	0.066	0.065	0.067	0.951	0.950
$PPV_{2.0}(1.5)$	0.599	0.613	0.601	0.056	0.052	0.055	0.053	0.945	0.950
$PPV_{2.0}(2.0)$	0.602	0.638	0.602	0.077	0.066	0.076	0.067	0.931	0.942
$NPV_{1.0}(1.5)$	0.954	0.956	0.952	0.020	0.077	0.019	0.066	0.890	0.971
$NPV_{1.0}(2.0)$	0.897	0.899	0.914	0.027	0.097	0.027	0.091	0.924	0.998
$NPV_{2.0}(1.5)$	0.632	0.635	0.723	0.052	0.142	0.051	0.138	0.948	0.854
$NPV_{2.0}(2.0)$	0.587	0.591	0.666	0.049	0.136	0.048	0.133	0.939	0.885
$AUC_{1.0}$	0.858	0.859	0.864	0.040	0.099	0.038	0.094	0.926	0.996
$AUC_{2.0}$	0.671	0.679	0.738	0.059	0.113	0.056	0.107	0.921	0.880
$IAUC$	0.764	0.776	0.792	0.041	0.094	0.039	0.093	0.905	0.940
Cause 2		CFK	IPW	CFK	IPW	CFK	IPW	CFK	IPW
$SE_{1.0}(1.5)$	0.899	0.903	0.882	0.053	0.076	0.049	0.072	0.863	0.908
$SE_{1.0}(2.0)$	0.718	0.727	0.719	0.085	0.107	0.083	0.101	0.913	0.916
$SE_{2.0}(1.5)$	0.574	0.578	0.573	0.067	0.071	0.067	0.070	0.938	0.942
$SE_{2.0}(2.0)$	0.402	0.414	0.404	0.074	0.071	0.074	0.070	0.938	0.939
$SP_{1.0}(1.5)$	0.636	0.633	0.723	0.035	0.154	0.035	0.153	0.945	0.831
$SP_{1.0}(2.0)$	0.763	0.760	0.817	0.035	0.131	0.035	0.129	0.951	0.929
$SP_{2.0}(1.5)$	0.601	0.600	0.678	0.044	0.133	0.044	0.130	0.948	0.881
$SP_{2.0}(2.0)$	0.721	0.722	0.789	0.046	0.119	0.045	0.112	0.944	0.863
$PPV_{1.0}(1.5)$	0.332	0.331	0.330	0.043	0.048	0.043	0.049	0.938	0.954
$PPV_{1.0}(2.0)$	0.378	0.386	0.378	0.058	0.064	0.058	0.065	0.948	0.952
$PPV_{2.0}(1.5)$	0.398	0.407	0.397	0.050	0.051	0.050	0.052	0.945	0.950
$PPV_{2.0}(2.0)$	0.398	0.428	0.400	0.066	0.065	0.067	0.066	0.935	0.951
$NPV_{1.0}(1.5)$	0.969	0.970	0.957	0.017	0.073	0.015	0.062	0.856	0.933
$NPV_{1.0}(2.0)$	0.931	0.933	0.931	0.023	0.088	0.022	0.080	0.905	0.996
$NPV_{2.0}(1.5)$	0.754	0.755	0.829	0.044	0.123	0.044	0.113	0.938	0.829
$NPV_{2.0}(2.0)$	0.724	0.728	0.793	0.042	0.120	0.042	0.112	0.939	0.867
$AUC_{1.0}$	0.821	0.824	0.832	0.046	0.107	0.043	0.103	0.917	0.987
$AUC_{2.0}$	0.628	0.637	0.687	0.056	0.111	0.055	0.105	0.925	0.888
$IAUC$	0.748	0.741	0.753	0.041	0.096	0.039	0.095	0.934	0.961

Table 3.2: Simulation results for the estimates of $\texttt{se}_t^{(k)}(y)$, $\texttt{sp}_t^{(k)}(y)$, $\texttt{PPV}_t^{(k)}(y)$, $\texttt{NPV}_t^{(k)}(y)$, $A^{(k)}(t)$ and $S^{(k)}$, when censoring rates are 50% on markers and 40% on outcomes. Ave is the average of the estimates. ESE is the empirical standard error of the estimator. BSE is the average of the bootstrap standard errors. Cov is the 95% coverage rate based on BSE.

Cause 1	True	Ave		ESE		BSE		Cov	
		CFK	IPW	CFK	IPW	CFK	IPW	CFK	IPW
$SE_{1.0}(1.5)$	0.900	0.905	0.872	0.051	0.083	0.047	0.075	0.865	0.905
$SE_{1.0}(2.0)$	0.720	0.731	0.708	0.091	0.120	0.089	0.110	0.903	0.898
$SE_{2.0}(1.5)$	0.575	0.593	0.574	0.080	0.075	0.076	0.073	0.909	0.936
$SE_{2.0}(2.0)$	0.404	0.437	0.401	0.088	0.078	0.085	0.074	0.901	0.931
$SP_{1.0}(1.5)$	0.697	0.696	0.768	0.042	0.180	0.042	0.177	0.953	0.970
$SP_{1.0}(2.0)$	0.818	0.819	0.829	0.047	0.158	0.046	0.153	0.940	0.993
$SP_{2.0}(1.5)$	0.655	0.672	0.741	0.076	0.213	0.074	0.191	0.927	0.875
$SP_{2.0}(2.0)$	0.760	0.788	0.818	0.086	0.207	0.081	0.173	0.897	0.895
$PPV_{1.0}(1.5)$	0.500	0.502	0.498	0.054	0.066	0.053	0.064	0.939	0.945
$PPV_{1.0}(2.0)$	0.571	0.591	0.572	0.084	0.096	0.082	0.092	0.942	0.928
$PPV_{2.0}(1.5)$	0.599	0.635	0.601	0.075	0.070	0.073	0.070	0.915	0.944
$PPV_{2.0}(2.0)$	0.602	0.687	0.603	0.112	0.097	0.104	0.094	0.860	0.931
$NPV_{1.0}(1.5)$	0.954	0.957	0.918	0.023	0.121	0.022	0.112	0.863	0.947
$NPV_{1.0}(2.0)$	0.897	0.901	0.871	0.035	0.142	0.034	0.137	0.898	0.992
$NPV_{2.0}(1.5)$	0.632	0.646	0.707	0.076	0.218	0.071	0.200	0.915	0.884
$NPV_{2.0}(2.0)$	0.587	0.608	0.658	0.070	0.211	0.066	0.195	0.899	0.869
$AUC_{1.0}$	0.858	0.856	0.816	0.055	0.146	0.051	0.141	0.902	0.989
$AUC_{2.0}$	0.671	0.697	0.718	0.085	0.194	0.078	0.171	0.870	0.908
$IAUC$	0.764	0.789	0.755	0.055	0.129	0.051	0.129	0.839	0.986
Cause 2		CFK	IPW	CFK	IPW	CFK	IPW	CFK	IPW
$SE_{1.0}(1.5)$	0.899	0.902	0.855	0.064	0.099	0.057	0.094	0.868	0.925
$SE_{1.0}(2.0)$	0.718	0.728	0.689	0.114	0.149	0.108	0.135	0.894	0.889
$SE_{2.0}(1.5)$	0.574	0.590	0.569	0.095	0.093	0.091	0.093	0.917	0.941
$SE_{2.0}(2.0)$	0.402	0.434	0.394	0.107	0.097	0.103	0.094	0.906	0.920
$SP_{1.0}(1.5)$	0.636	0.635	0.721	0.039	0.191	0.039	0.189	0.943	0.897
$SP_{1.0}(2.0)$	0.763	0.763	0.794	0.043	0.168	0.043	0.164	0.942	0.992
$SP_{2.0}(1.5)$	0.601	0.608	0.684	0.055	0.208	0.055	0.190	0.948	0.870
$SP_{2.0}(2.0)$	0.721	0.734	0.781	0.062	0.202	0.062	0.172	0.944	0.880
$PPV_{1.0}(1.5)$	0.332	0.335	0.333	0.047	0.059	0.046	0.060	0.944	0.945
$PPV_{1.0}(2.0)$	0.378	0.398	0.380	0.070	0.091	0.069	0.089	0.950	0.942
$PPV_{2.0}(1.5)$	0.398	0.425	0.401	0.062	0.066	0.062	0.067	0.942	0.947
$PPV_{2.0}(2.0)$	0.398	0.465	0.402	0.090	0.094	0.091	0.092	0.921	0.940
$NPV_{1.0}(1.5)$	0.969	0.970	0.921	0.019	0.119	0.018	0.109	0.872	0.917
$NPV_{1.0}(2.0)$	0.931	0.933	0.883	0.029	0.138	0.028	0.131	0.900	0.970
$NPV_{2.0}(1.5)$	0.754	0.762	0.796	0.061	0.215	0.060	0.185	0.915	0.923
$NPV_{2.0}(2.0)$	0.724	0.737	0.767	0.058	0.209	0.056	0.183	0.912	0.907
$AUC_{1.0}$	0.821	0.813	0.781	0.060	0.154	0.056	0.151	0.911	0.988
$AUC_{2.0}$	0.628	0.647	0.670	0.080	0.188	0.076	0.170	0.901	0.908
$IAUC$	0.748	0.745	0.716	0.055	0.132	0.052	0.134	0.914	0.985

In the second simulation, we aim to compare the discriminatory ability between two markers by conducting the tests given in the previous section. The natural logs of the two markers Y_1 and Y_2 are generated from a bivariate normal distribution with zero mean, marginal variance of 1 and correlation of ρ. Three values of ρ are considered $\rho = -0.5, 0, 0.5$. The event time T, which is subject to competing-risk censoring, is generated from a proportional cause-specific hazard model $\alpha_k(t) = \alpha_{0k} \exp(\theta_{k1} Y_1 + \theta_{k2} Y_2)$, based on the values of the markers, where the cause-specific hazard function $\alpha_k(t)$ for T is defined as

$$\alpha_k(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t, \epsilon = k | T \geq t),$$

where $k = 1, 2$. We assume a common constant baseline cause-specific hazard function for the two causes, $\alpha_{01}(t) = \alpha_{02}(t) = 0.2$. Since the sum of the cause-specific hazards gives the overall hazard of T, we have $P(T > t) = \exp(-\int_0^t \alpha_1(x) + \alpha_2(x) dx)$, resulting an exponential distribution for time T with a constant rate parameter $\alpha_1(t) + \alpha_2(t)$. Furthermore, it can be shown that the cause indicator ϵ is from a Bernoulli distribution with the probability $\alpha_k(t)/(\alpha_1(t) + \alpha_2(t))$ for cause k. In order to build various relationship between the markers and the event time, we consider three cases for the coefficients in the proportional cause-specific hazard model: $\theta_{11} = \theta_{12} = 0.5$, $\theta_{21} = \theta_{22} = 0.2$; $\theta_{11} = 0.5$, $\theta_{12} = 0.3$, $\theta_{21} = 0.1$, $\theta_{22} = 0.2$; $\theta_{11} = 0.5$, $\theta_{12} = 0$, $\theta_{21} = 0$, $\theta_{22} = 0.2$. Together with ρ, there are nine scenarios in total. The independent censoring variables C_Y and C_T are generated from $Uniform(0, 4)$ and $Uniform(0, 3)$, respectively, both yielding about 35% censoring. For each scenario, we simulate 2000 datasets and we consider sample sizes of $n = 300$ and $n = 150$. The test based on the AUC is conducted at the median of T. In the test for the overall prognostic capacity, IAUC is calculated between the 10% and 90% quantiles of T. The standard errors that are involved in the tests are computed based on 250 bootstrap samples.

The behaviors of the tests at 5% level are presented in Table 3.3. The rejection rates for the total nine scenarios are reported. There are three cases for each value of ρ. In case 1 where $\theta_{11} = \theta_{12} = 0.5$ and $\theta_{21} = \theta_{22} = 0.2$ in the model, the two markers are associated with the event times in exactly the same way for both causes. Therefore, as can be seen, the null is rejected about 5% of the time, which agree with the significance level of the test. When the alternative is true, the tests for cause 1 event have more power than their cause

Table 3.3: Simulation results in nine scenarios for the rejection rates of the proposed tests based on the AUC or the IAUC, with sample size 300 or 150.

$n = 300$			$\rho = -0.5$			$\rho = 0$			$\rho = 0.5$		
Case			1	2	3	1	2	3	1	2	3
AUC	CFK	cause 1	0.048	0.444	0.993	0.056	0.285	0.936	0.064	0.152	0.606
		cause 2	0.065	0.210	0.748	0.045	0.148	0.511	0.053	0.089	0.214
	IPW	cause 1	0.062	0.114	0.505	0.058	0.090	0.235	0.050	0.060	0.090
		cause 2	0.073	0.098	0.174	0.074	0.082	0.102	0.076	0.064	0.070
IAUC	CFK	cause 1	0.049	0.441	0.998	0.063	0.287	0.960	0.058	0.162	0.648
		cause 2	0.062	0.301	0.890	0.048	0.196	0.660	0.053	0.117	0.294
	IPW	cause 1	0.063	0.145	0.708	0.067	0.092	0.367	0.060	0.064	0.101
		cause 2	0.062	0.103	0.320	0.068	0.080	0.127	0.064	0.060	0.070
$n = 150$			$\rho = -0.5$			$\rho = 0$			$\rho = 0.5$		
Case			1	2	3	1	2	3	1	2	3
AUC	CFK	cause 1	0.066	0.248	0.886	0.055	0.165	0.706	0.056	0.090	0.358
		cause 2	0.069	0.144	0.457	0.068	0.114	0.314	0.061	0.081	0.136
	IPW	cause 1	0.070	0.098	0.430	0.056	0.086	0.215	0.044	0.048	0.078
		cause 2	0.077	0.078	0.165	0.072	0.076	0.102	0.067	0.060	0.072
IAUC	CFK	cause 1	0.064	0.282	0.918	0.057	0.176	0.759	0.062	0.098	0.385
		cause 2	0.073	0.193	0.621	0.058	0.134	0.423	0.069	0.089	0.169
	IPW	cause 1	0.064	0.134	0.620	0.070	0.096	0.327	0.056	0.054	0.113
		cause 2	0.068	0.089	0.260	0.062	0.071	0.142	0.064	0.060	0.080

2 counterparts because of the larger number of observed events in cause 1. In case 3, the first marker Y_1 is only associated with the cause 1 event, while the second marker Y_2 is only associated with the cause 2 event. Thus, compared with case 2, there are more power for each cause to detect the difference in the prognostic abilities between Y_1 and Y_2. Case 2 and case 3 assume that for each cause, the event times largely depend on one of the two markers and as a result, a negative association ($\rho = -0.5$ in log scale) between the markers will make it easier to reject the null. When $\rho = -0.5$ in case 3, the power of cause 1 test is very close to 1. The results also clearly suggest that the tests on the basis of IAUC have higher rejection proportions than those that are based on AUC, which can be explained by the larger amount of information integrated by the overall accuracy measure. In addition, there is substantial power under the smaller sample size $n = 150$, where the power under the CFK method is noticeably better than that under the IPW method, which can be mostly attributed to the fact that the IPW estimates have larger variability. The simulation results suggest that the CFK method generally performs better, especially in terms of providing smaller standard errors and more powerful tests.

3.4 BDCP ANALYSIS

We now apply the proposed methods to the data from the BDCP Study. We consider a subset of a total of 164 patients who entered the study in an euthymic state. Euthymia refers to the state of neutral mood, absence of depressive or manic symptoms. We focus on the future development of new episodes during the follow-up of about 2 years. The baseline scores, as markers, are of clinical interest in predicting the subject's vulnerability to recurrence, either in the short term or long term. Among the available baseline markers, we have traditional self-report scores from questionnaires, along with two novel instrument markers that are subject to right censoring, the numbers of previous episodes of depression and mania/hypomania. We evaluate the time-dependent diagnostic abilities of these censored markers to discriminate the patients that were likely or unlikely to develop new episodes. To this end, their ROC curves are presented. We also compare the new instrument with the traditional complete markers through the ROC curves, AUC and IAUC.

Of the 164 participants, 52 developed new episodes by the end of the follow-up, in which 47 were depression, 4 were mania/hypomania and 1 was mixed episode. The median time to a new depressive episode is 32 weeks. Due to the extremely small number of the manic/hypomanic and mixed episodes, we only consider time to a new depressive episode as the outcome of interest, whereas the other types of episodes are treated as competing events.

In Figure 3.1, we show the estimated ROC curves for the outcome of depression. We notice that there are not many jumps in the curves for the markers of previous depression and previous mania/hypomania, which is due to the limited number of observations for the outcome, as well as the limited marker information. We have 63 (73) patients whose information of previous depression (mania/hypomania) are available, where we assume the missingness is random. The results of both the CFK and the IPW method are presented, and their performances are almost indistinguishable. The ROC curves are just above the diagnal, indicating weak diagnostic abilities of the two markers. At both half a year and one year, the number of previous depressive episodes is slightly better than the number of previous manic/hypomanic episodes in discriminating the subjects who did and did not develop new depressive episode.

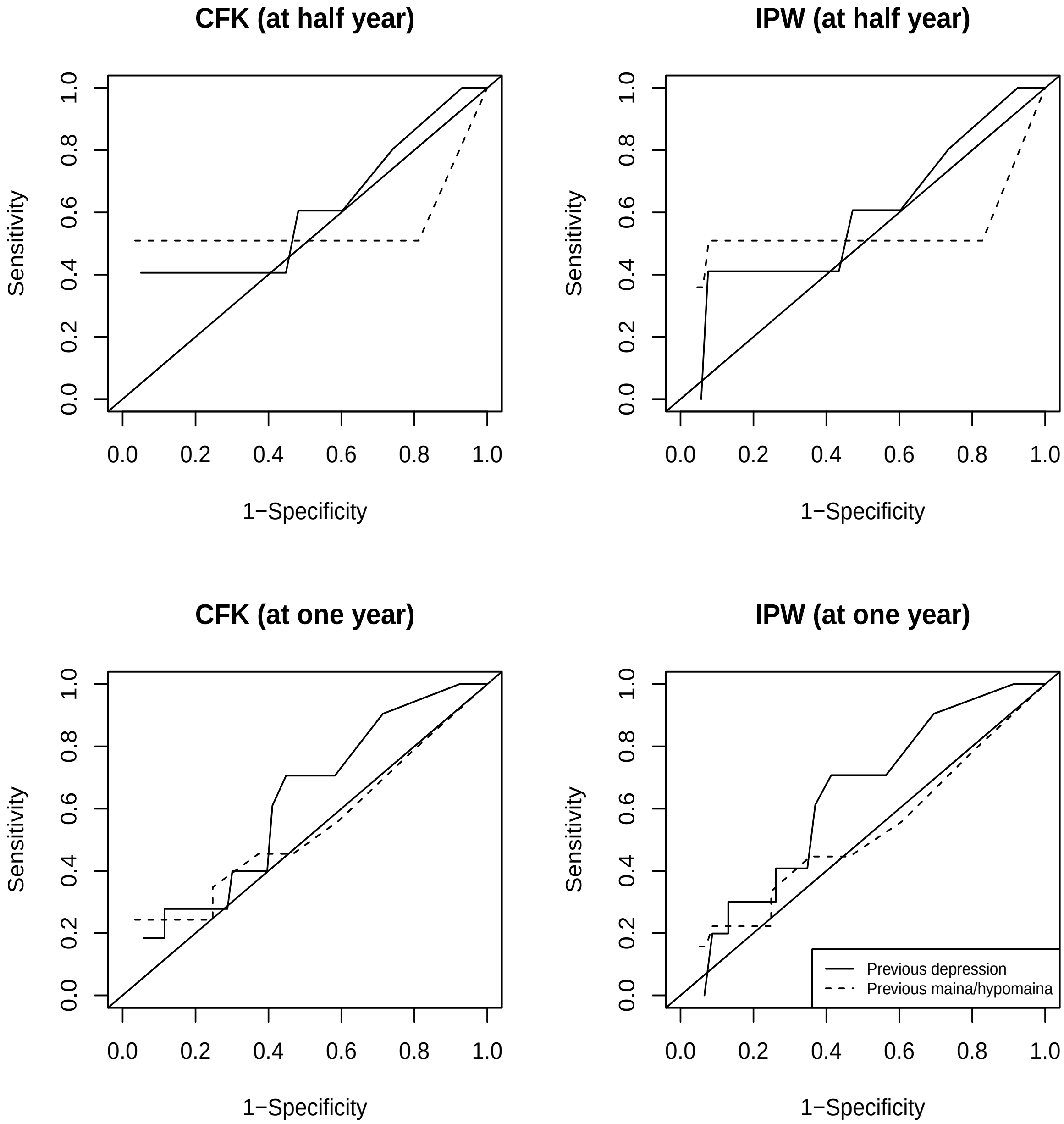

Figure 3.1: ROC curves for depressive outcome evaluated at half year and one year, based on the CFK and IPW methods.

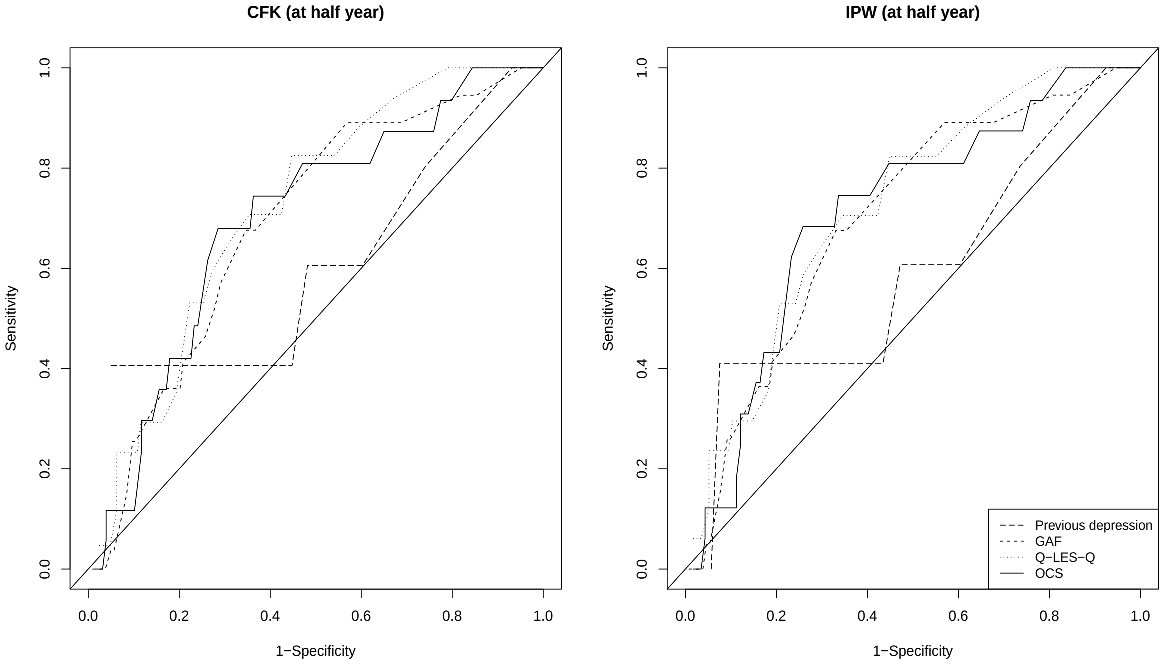

Figure 3.2: ROC curves for depressive outcome evaluated at half year, based on the CFK and IPW methods.

Among the other traditional baseline markers that are completely observed without being censored, we find three of them that are very informative in predicting depression: the past month Global Assessment of Functioning score (GAF), the Quality of Life Enjoyment and Satisfaction Questionnaire score at intake (Q-LES-Q) and the Obsessive-Compulsive Spectrum (OCS). In order to follow the definitions of sensitivity and specificity in Section 3.2, the negative of the marker values of the GAF and the Q-LES-Q is used instead. We compare their diagnostic capabilities with the censored marker of previous number of depressive episodes in discriminating between subjects who developed a new depressive episode by half a year and those who did not develop any new depressive episode by half a year. Figure 3.2 describes the estimated ROC curves for these markers. The results from both estimating methods demonstrate that the diagnostic performance of the three traditional markers are very similar, and are all better than that of the number of previous depressive episodes. This naturally leaded us to the next question on whether this difference is statistically significant.

In order to address this question, we now look at their AUCs and IAUCs for the depressive outcome. We compare each of the three traditional markers with the novel marker of the number of previous depressive episodes based on the test that we have proposed in Section 3.2. The GAF, Q-LES-Q and OCS all have larger AUC at half a year than at one year. The AUCs at half a year are 0.69 for GAF, 0.72 for Q-LES-Q, and 0.69 for OCS. Compared with the novel marker, the differences at either one year or half a year are not significant. However, we find some significant results for their discriminatory abilities for early development of depression. Table 3.4 summarizes these differences (traditional marker - novel marker) in the AUC at 2 or 3 months, along with the IAUC. Here only the results based on the CFK method are presented since the IPW method provides comparable estimates. For each pair of comparison, we compute the point estimates of the difference and the 95% confidence intervals based on the bootstrap standard errors. It turns out that none of the intervals at 2 months cover zero. This indicates that all of the three traditional markers have substantially higher capacity than the novel marker in discriminating between the subjects who experienced a new depressive episode by 2 months and the subjects who did not. While for the 3-month classification, only the past month Global Assessment of Functioning score (GAF) perform significantly better than the novel marker. Thus, it can be observed that the novel instrument marker is less accurate than traditional instruments for short-term development of new depressive episodes. As the time period turns larger, the advantages of the traditional instruments become less significant. We also show the computed IAUC based on the AUC from 8 weeks to 1.5 years using uniform weight, since there are very few observations less than 8 weeks or beyond 1.5 years. Note that all the estimated differences are less than 0.10 and the confidence intervals cover zero. The analysis shows that the number of previous depressive episodes, as a novel marker, does not necessarily provide more insights in the discrimination for future depression, compared with traditional instruments.

3.5 DISCUSSION

We have proposed two non-parametric methods for estimating time-dependent discrimination measures with censored marker and competing-risk censored outcome. The CFK

Table 3.4: Estimated difference in AUC and IAUC between traditional instrument marker and the novel marker (traditional - novel)

AUC		**GAF**	**Q-LES-Q**	**OCS**
At 2 month	Point estimate	0.54	0.44	0.42
	95% CI	(0.31, 0.77)	(0.18, 0.69)	(0.06, 0.79)
At 3 month	Point estimate	0.33	0.23	0.28
	95% CI	(0, 0.65)	(-0.10, 0.57)	(-0.04, 0.60)
IAUC		**GAF**	**Q-LES-Q**	**OCS**
	Point estimate	0.08	0.08	0.09
	95% CI	(-0.12, 0.29)	(-0.12, 0.27)	(-0.10, 0.27)

method utilizes the existing estimators for the bivariate CIF and the univariate quantities and performs generally better than the IPW method which entirely relies on the idea of inverse probability weighting. The nonparametric methods that we proposed do not rely on specific assumptions for the distribution of the marker or the outcome and hence tend to be more robust than some parametric or semiparametric counterparts. For example, using Cox model in the diagnostic accuracy analysis needs to verify the model assumption and violation of the regularity assumptions would result in bias in medical decision making (Schmid and Potapov, 2012).

In our application, the proposed methods serve as useful tools in analyzing the bipolar disorder study that is considered in this chapter. They have provided valid quantitative evidence for the comparison of diagnostic ability between the novel marker and the traditional instruments. The analysis reveals that the novel marker fails to provide more information than the exiting instruments in guiding medical diagnostic decisions. We have mainly focused on the competing-risk censored outcome and right-censored marker. In other applications, it is possible to extend the present work to other types of censoring schemes. In addition, when it is desirable in multiple endpoint studies to explore the potential of a marker in diagnosis for different diseases, testing for the diagnostic abilities across different causes or end points would be worthwhile.

APPENDIX

PROOFS

A.1 ASYMPTOTIC PROPERTIES OF THE ESTIMATORS PROPOSED IN CHAPTER 2

It follows from the uniform consistency of the Kaplan-Meier estimator $\hat{G}$ (Breslow et al., 1974) in Chapter 2 and the strong law of large numbers that the estimator for the bivariate cumulative incidence function $\hat{F}_l(x, t)$ is uniformly consistent.

Now we examine the weak convergence of $\hat{F}_l(x,t)$. Note that

$$\begin{aligned}
&\sqrt{n}\{\hat{F}_l(x,t) - P(X \leq x, T \leq t, \epsilon = l)\} \\
=&\sqrt{n}\Big[\frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{1}{\hat{G}(Y_{2i}-)}I(Y_{1i} \leq x, \eta_{1i} = 1, Y_{2i} - Y_{1i} \leq t, \eta_{2i} = l)\Big\} \\
&- \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{1}{G(Y_{2i}-)}P(X_i \leq x, T_i \leq t, \epsilon = l, C_i \geq X_i + T_i)\Big\}\Big] \\
=&\sqrt{n}\Big[\frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{1}{\hat{G}(Y_{2i}-)}I(Y_{1i} \leq x, \eta_{1i} = 1, Y_{2i} - Y_{1i} \leq t, \eta_{2i} = l)\Big\} \\
&- \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{1}{G(Y_{2i}-)}P(Y_{1i} \leq x, \eta_{1i} = 1, Y_{2i} - Y_{1i} \leq t, \eta_{2i} = l)\Big\} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{1}{G(Y_{2i}-)}I(Y_{1i} \leq x, \eta_{1i} = 1, Y_{2i} - Y_{1i} \leq t, \eta_{2i} = l)\Big\} \\
&- \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{1}{G(Y_{2i}-)}I(Y_{1i} \leq x, \eta_{1i} = 1, Y_{2i} - Y_{1i} \leq t, \eta_{2i} = l)\Big\}\Big] \\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\Big\{\frac{1}{\hat{G}(Y_{2i}-)} - \frac{1}{G(Y_{2i}-)}\Big\}I(Y_{1i} \leq x, \eta_{1i} = 1, Y_{2i} - Y_{1i} \leq t, \eta_{2i} = l)\Big] \\
&+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\frac{1}{G(Y_{2i}-)}\Big\{I(Y_{1i} \leq x, \eta_{1i} = 1, Y_{2i} - Y_{1i} \leq t, \eta_{2i} = l) \\
&- P(Y_{1i} \leq x, \eta_{1i} = 1, Y_{2i} - Y_{1i} \leq t, \eta_{2i} = l)\Big\}\Big].
\end{aligned}$$

The second term clearly converges to a Gaussian process with mean zero. Now we consider the first term, which can be written as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\frac{G(Y_{2i}-) - \hat{G}(Y_{2i}-)}{G(Y_{2i}-)\hat{G}(Y_{2i}-)}I(Y_{1i} \leq x, \eta_{1i} = 1, Y_{2i} - Y_{1i} \leq t, \eta_{2i} = l)\Big]. \tag{A.1.1}$$

Based on the martingale representation for the Kaplan-Meier estimator (Kalbfleisch and Prentice, 2002; Fleming and Harrington, 1991), we have

$$\frac{G(t) - \hat{G}(t)}{G(t)} = \int_0^t \frac{\hat{G}(s-)}{G(s)}\frac{\sum_{i=1}^{n} dM_i(s)}{\sum_{i=1}^{n} I(Y_{2i} \geq s)}, \tag{A.1.2}$$

where $t \leq \max_i Y_{2i}$ and the martingale for the censoring time is defined as

$$M_i(t) = I(C_i \leq (X_i + T_i) \wedge t) - \int_0^t I(Y_{2i} \geq s) d\Lambda_C(s),$$

where $\Lambda_C(s)$ is the cumulative hazard function of the censoring time.

Then following the arguments in Lin et al. (1999) and plugging (A.1.2) into (A.1.1), (A.1.1) now becomes

$$\frac{1}{\sqrt{n}} \int_0^{R_C} \{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\hat{G}(Y_{2i}-)} I(Y_{1i} \leq x, \eta_{1i} = 1, Y_{2i} - Y_{1i} \leq t, \eta_{2i} = l, Y_{2i} \geq s) \}$$

$$\times \frac{\hat{G}(s-)}{G(s)} \frac{\sum_{i=1}^{n} dM_i(s)}{\frac{1}{n} \sum_{i=1}^{n} I(Y_{2i} \geq s)} + o_P(1)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^{R_C} [\frac{G(s-)\{P(X \leq x, T \leq t, \epsilon = l) - P(X \leq x, T \leq t, \epsilon = l, X + T < s)\}}{P(Y_2 \geq s) G(s)}] dM_i(s)$$

$$+ o_P(1),$$

resulting from the consistency of $\hat{G}$ and $\hat{F}_l(x,t)$, where R_C is the upper bound of censoring. Note that the summands of the above are independent random variables with mean zero, which means (A.1.1) converges to a zero-mean Gaussian process. Therefore, the weak convergence of $\hat{F}_l(x,t)$ naturally follows. The uniform consistency and weak convergence for the cause-specific odds ratio $\phi_{(l)}(x,t|M)$ and the integrated odds ratio $\phi^{\star}_{(l)}(M)$ can be demonstrated by following the arguments in Cheng et al. (2007).

A.2 ASYMPTOTIC RESULTS OF THE PROPOSED METHODS IN CHAPTER 3

In this section, we justify the asymptotic properties for the two sets of estimators that have been proposed in Chapter 3. First note that the uniform consistency and weak convergence hold for the Kaplan-Meier estimator $\hat{S}_Y(y)$, the univariate CIF estimator $\hat{F}_T^{(k)}(t)$ (Kalbfleisch and Prentice, 2002), and the estimator for the bivariate CIF $\hat{F}_{Y,T}^{(k)}(y,t)$ (Cheng et al., 2007).

We can then establish the uniform consistency, as well as the weak convergence by the functional δ-method, for the CFK estimators $\hat{\mathtt{se}}_t^{(k)}(y)$, $\hat{\mathtt{sp}}_t^{(k)}(y)$, $\mathtt{P\hat{P}V}_t^{(k)}(y)$, $\mathtt{N\hat{P}V}_t^{(k)}(y)$, $\hat{R}_t^{(k)}(p)$ and $\hat{S}^{(k)}$ that are proposed in section 3.2.1.1.

The uniform consistency of the IPW estimators $\widetilde{\mathtt{se}}_t^{(k)}(y)$, $\widetilde{\mathtt{sp}}_t^{(k)}(y)$, $\widetilde{\mathtt{PPV}}_t^{(k)}(y)$, $\widetilde{\mathtt{NPV}}_t^{(k)}(y)$, $\widetilde{R}_t^{(k)}(p)$ and $\widetilde{S}^{(k)}$ naturally follows from the uniform consistency of the Kaplan-Meier estimators $\hat{S}_{C_T}$, $\hat{S}_{C_Y}$, and the Dabrowska estimator $\hat{G}$ for the bivariate censoring times (Gill et al., 1995). For weak convergence, we will first focus on the sensitivity $\widetilde{\mathtt{se}}_t^{(k)}(y)$. For the numerator of $\widetilde{\mathtt{se}}_t^{(k)}(y)$, we have

$$
\begin{aligned}
&\sqrt{n}\{\frac{1}{n}\sum_{i=1}^{n}[\hat{G}(y-,X_i-)]^{-1}I(W_i \geq y, X_i \leq t, \delta_i = k) - P(Y \geq y, T \leq t, \epsilon = k)\}\\
=&\sqrt{n}\{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\hat{G}(y-,X_i-)}I(W_i \geq y, X_i \leq t, \delta_i = k)\\
&-\frac{1}{n}\sum_{i=1}^{n}\frac{1}{G(y-,X_i-)}P(Y_i \geq y, C_{Y_i} \geq y, T_i \leq t, C_{T_i} \geq T_i, \epsilon = k)\}\\
=&\sqrt{n}\{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\hat{G}(y-,X_i-)}I(W_i \geq y, X_i \leq t, \delta_i = k)\\
&-\frac{1}{n}\sum_{i=1}^{n}\frac{1}{G(y-,X_i-)}I(W_i \geq y, X_i \leq t, \delta_i = k)\\
&+\frac{1}{n}\sum_{i=1}^{n}\frac{1}{G(y-,X_i-)}I(W_i \geq y, X_i \leq t, \delta_i = k)\\
&-\frac{1}{n}\sum_{i=1}^{n}\frac{1}{G(y-,X_i-)}P(Y_i \geq y, C_{Y_i} \geq y, T_i \leq t, C_{T_i} \geq T_i, \epsilon = k)\}\\
=&\sqrt{n}\{\frac{1}{n}\sum_{i=1}^{n}[\frac{1}{\hat{G}(y-,X_i-)} - \frac{1}{G(y-,X_i-)}]I(W_i \geq y, X_i \leq t, \delta_i = k)\}\\
&+\sqrt{n}\{\frac{1}{n}\sum_{i=1}^{n}[\frac{1}{G(y-,X_i-)}I(W_i \geq y, X_i \leq t, \delta_i = k)-\\
&\frac{1}{G(y-,X_i-)}P(W_i \geq y, X_i \leq t, \delta_i = k)]\}
\end{aligned}
$$

where $k = 1, 2$. Both of the two terms above are sums of iid random variables with mean 0 and hence converge weakly to a Gaussian process with mean 0. The asymptotic normality for the denominator of $\widetilde{\mathtt{se}}_t^{(k)}(y)$ can be established in a similar fashion. Therefore, $\sqrt{n}(\widetilde{\mathtt{se}}_t^{(k)}(y) - \mathtt{se}_t^{(k)}(y))$ converges weakly to a zero-mean Gaussian process.

The asymptotic properties for the IPW estimators of the other discrimination measures can be similarly obtained. The tests based on the CFK and IPW methods for the comparison of two markers are then justified using the above asymptotic results.

BIBLIOGRAPHY

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

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

Bandeen-Roche, K. and J. Ning (2008). Nonparametric estimation of bivariate failure time associations in the presence of a competing risk. *Biometrika 95*, 221–232.

Beyersmann, J., A. Allignol, and M. Schumacher (2011). *Competing risks and multistate models with R.* Springer Science & Business Media.

Blanche, P., J.-F. Dartigues, and H. Jacqmin-Gadda (2013a). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in medicine 32*(30), 5381–5397.

Blanche, P., J.-F. Dartigues, and H. Jacqmin-Gadda (2013b). Review and comparison of roc curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal 55*(5), 687–704.

Breslow, N., J. Crowley, et al. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics 2*(3), 437–453.

Cai, T. and S. Cheng (2008). Robust combination of multiple diagnostic tests for classifying censored event times. *Biostatistics 9*(2), 216–233.

Cai, T., T. A. Gerds, Y. Zheng, and J. Chen (2011). Robust prediction of t-year survival with data from multiple studies. *Biometrics 67*, 436–444.

Cai, T., M. Pepe, Y. Zheng, T. Lumley, and N. Jenny (2006). The sensitivity and specificity of markers for event times. *Biostatistics 7*, 182–197.

Cai, T. and Y. Zheng (2011). Nonparametric evaluation of biomarker accuracy under nested case-control studies. *Journal of the American Statistical Association 106*, 569–580.

Chambless, L. E. and G. Diao (2006). Estimation of time-dependent area under the roc curve for long-term risk prediction. *Statistics in medicine 25*(20), 3474–3486.

Chang, S.-H. and S.-J. Tzeng (2006). Nonparametric estimation of sojourn time distributions for truncated serial event dataa weight-adjusted approach. *Lifetime Data Analysis 12*(1), 53–67.

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

Cheng, Y. and J. P. Fine (2012). Cumulative incidence association models for bivariate competing risks data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(2), 183–202.

Cheng, Y., J. P. Fine, and K. Bandeen-Roche (2010). Association analyses of clustered competing risks data via cross hazard ratio. *Biostatistics 11*(1), 82–92.

Cheng, Y., J. P. Fine, and M. R. Kosorok (2007). Nonparametric analysis of bivariate competing risks data. *Journal of the American Statistical Association 102*, 1407–1416.

Cheng, Y. and J. Li (2015). Time-dependent diagnostic accuracy analysis with censored outcome and censored predictor. *Journal of Statistical Planning and Inference 156*, 90–102.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika 65*, 141–152.

Dabrowska, D. M. (1988). Kaplan meier estimate on the plane. *The Annals of Statistics 16*, 1475–1489.

Datta, S. and G. A. Satten (2001). Validity of the aalen–johansen estimators of stage occupation probabilities and nelson–aalen estimators of integrated transition hazards for non-markov models. *Statistics & probability letters 55*(4), 403–411.

de Uña Álvarez, J. and L. F. Meira-Machado (2008). A simple estimator of the bivariate distribution function for censored gap times. *Statistics & Probability Letters 78*(15), 2440–2445.

Fagiolini, A., E. Frank, D. A. Axelson, B. Birmaher, Y. Cheng, D. E. Curet, E. S. Friedman, A. G. Gildengers, T. Goldstein, V. J. Grochocinski, et al. (2009). Enhancing outcomes in patients with bipolar disorder: results from the bipolar disorder center for pennsylvanians study. *Bipolar disorders 11*(4), 382–390.

Fleming, T. R. and D. P. Harrington (1991). *Counting Processes and Survival Analysis.* John Wiley & Sons.

Foucher, Y., M. Giral, J. Soulillou, and J. Daures (2010). Time-dependent roc analysis for a three-class prognostic with application to kidney transplantation. *Statistics in medicine 29*(30), 3079–3087.

Gill, R. D. (1992). Multistate life-tables and regression models. *Mathematical Population Studies 3*(4), 259–276.

Gill, R. D., M. J. van der Laan, and J. A. Wellner (1995). Inefficient estimators of the bivariate survival function for three models. *Annales de L'Institut Henri Poincaré Probabilités et Statistiques 31*, 545–597.

Heagerty, P., T. Lumley, and M. S. Pepe (2000). Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics 56*, 337–344.

Heagerty, P. and Y. Zheng (2005). Survival model predictive accuracy and roc curves. *Biometrics 61*, 92–105.

Huan, C.-Y. and M.-C. Wang (2005). Nonparametric estimation of the bivariate recurrence time distribution. *Biometrics 61*(2), 392–402.

Huang, Y. and T. A. Louis (1998). Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika 85*(4), 785–798.

Hung, H. and C.-T. Chiang (2010a). Estimation methods for time-dependent auc models with survival data. *Canadian Journal of Statistics 38*(1), 8–26.

Hung, H. and C.-T. Chiang (2010b). Optimal composite markers for time-dependent receiver operating characteristic curves with censored survival data. *Scandinavian Journal of Statistics 37*(4), 664–679.

Jafarzadeh, S. R., W. O. Johnson, J. M. Utts, and I. A. Gardner (2010). Bayesian estimation of the receiver operating characteristic curve for a diagnostic test with a limit of detection in the absence of a gold standard. *Statistics in medicine 29*(20), 2090–2106.

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

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 81–93.

Klein, J. P. and M. L. Moeschberger (2003). *Survival Analysis: Techniques for Censored and Truncated Data.* New York: Springer-Verlag Inc.

Li, J. and S. Ma (2011). Time-dependent roc analysis under diverse censoring patterns. *Statistics in medicine 30*(11), 1266–1277.

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

Lin, D. Y. and Z. Ying (2001). Nonparametric tests for the gap time distributions of serial events based on censored data. *Biometrics 57*(2), 369–375.

Mumford, S. L., E. F. Schisterman, A. Vexler, and A. Liu (2006). Pooling biospecimens and limits of detection: effects on roc curve analysis. *Biostatistics 7*(4), 585–598.

Pepe, M. S. (1991). Inference for events with dependent risks in multiple endpoint studies. *Journal of the American Statistical Association 86*(415), 770–778.

Pepe, M. S. and M. Mori (1993). Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in Medicine 12*, 737–751.

Pepe, M. S., Y. Zheng, Y. Jin, Y. Huang, C. R. Parikh, and W. C. Levy (2008). Evaluating the roc performance of markers for future events. *Lifetime data analysis 14*(1), 86–113.

Perkins, N. J., E. F. Schisterman, and A. Vexler (2007). Receiver operating characteristic curve inference from a sample with a limit of detection. *American journal of epidemiology 165*(3), 325–333.

Perkins, N. J., E. F. Schisterman, and A. Vexler (2009). Generalized roc curve inference for a biomarker subject to a limit of detection and measurement error. *Statistics in medicine 28*(13), 1841–1860.

Perkins, N. J., E. F. Schisterman, and A. Vexler (2011). Roc curve inference for best linear combination of two biomarkers subject to limits of detection. *Biometrical Journal 53*(3), 464–476.

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

Sankaran, P. G., J. F. Lawless, B. Abraham, and A. A. Antony (2006). Estimation of distribution function in bivariate competing risk models. *Biometrical Journal 48*(3), 399–410.

Schaubel, D. E. and J. Cai (2004). Non-parametric estimation of gap time survival functions for ordered multivariate failure time data. *Statistics in Medicine 23*(12), 1885–1900.

Scheike, T. H., Y. Sun, M.-J. Zhang, and T. K. Jensen (2010). A semiparametric random effects model for multivariate competing risks data. *Biometrika 97*(1), 133–145.

Schmid, M. and S. Potapov (2012). A comparison of estimators to evaluate the discriminatory power of time-to-event models. *Statistics in medicine 31*(23), 2588–2609.

Shen, P. (2010). Interval estimation of the joint survival function for successive duration times under left truncation and right censoring. *Journal of Statistical Computation and Simulation 80*(11), 1265–1277.

Shi, H., Y. Cheng, and J. Li (2014). Assessing diagnostic accuracy improvement for survival or competing-risk censored outcomes. *Canadian Journal of Statistics 42*(1), 109–125.

Shih, J. H. and P. S. Albert (2010). Modeling familial association of ages at onset of disease in the presence of competing risk. *Biometrics 66*(4), 1012–1923.

Song, X. and X.-H. Zhou (2008). A semiparametric approach for the covariate specific roc curve with survival outcome. *Statistica Sinica 18*(3), 947–965.

Uno, H., T. Cai, L. Tian, and L. Wei (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association 102*(478).

van der Laan, M. J., A. E. Hubbard, and J. M. Robins (2002). Locally efficient estimation of a multivariate survival function in longitudinal Studies. *Journal of the American Statistical Association 97*(458), 494–507.

Visser, M. (1996). Nonparametric estimation of the bivariate survival function with an application to vertically transmitted AIDS. *Biometrika 83*, 507–518.

Wang, M.-C. (1999). Gap time bias in incident and prevalent cohorts. *Statistica Sinica 9*, 999–1010.

Wang, W. and M. T. Wells (1998). Nonparametric estimation of successive duration times under dependent censoring. *Biometrika 85*, 561–572.

Wolf, P., G. Schmidt, and K. Ulm (2011). The use of roc for defining the validity of the prognostic index in censored data. *Statistics & Probability Letters 81*(7), 783–791.

Zheng, Y., T. Cai, Y. Jin, and Z. Feng (2012). Evaluating prognostic accuracy of biomarkers under competing risk. *Biometrics 68*(2), 388–396.

Zheng, Y. and P. J. Heagerty (2004). Semiparametric estimation of time-dependent roc curves for longitudinal marker data. *Biostatistics 5*(4), 615–632.

Zheng, Y. and P. J. Heagerty (2007). Prospective accuracy for longitudinal markers. *Biometrics 63*(2), 332–341.

Zhu, H. and M.-C. Wang (2012). Analysing bivariate survival data with interval sampling and application to cancer epidemiology. *Biometrika*, ass009.