A Unified Parametric Approach to the Estimation of Dependence and Marginal Distributions in Bivariate Competing Risks Survival Data

Hyun-Soo Zhang

The Graduate School
Yonsei University
Department of Biostatistics and Computing

A Unified Parametric Approach to the Estimation of Dependence and Marginal Distributions in Bivariate Competing Risks Survival Data

A Dissertation Submitted to the
Department of Biostatistics and Computing
and the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Biostatistics and Computing

Hyun-Soo Zhang

June 2023

This certifies that the doctoral thesis of ***Hyun-Soo Zhang*** is approved.

Chung Mo Nam: Thesis Supervisor

Inkyung Jung: Thesis Committee Member #1

Sohee Park: Thesis Committee Member #2

Min Jin Ha: Thesis Committee Member #3

Boyoung Park: Thesis Committee Member #4

The Graduate School

Yonsei University

June 2023

Contents

List of Tables

Appendix

List of Figures

Abstract

A Unified Parametric Approach to the Estimation of Dependence and Marginal Distributions in Bivariate Competing Risks Survival Data

In bivariate competing risks survival data where only the minimum of the time-to-events is observed and never both, dependence between the survival endpoints is known to be non-identifiable. If dependence or correlation exists between the time-to-events, cause-specific hazards analysis under independent censoring or inference under incorrectly assumed correlations become biased. Arguably, the most important parameter for estimation when dependence exists is the correlation between the time-to-events. However, maximum likelihood estimation (MLE) is known to be biased with large variance, and no practical methods to estimate the correlation exist.

Using the fact that bivariate normally (BVN) distributed competing risks data is identifiable, we propose a unified parametric approach where the bivariate central limit theorem provides a connection between a given bivariate competing risks data and the identifiable BVN distribution. We demonstrate that the correlation in the given data is estimable by finding a BVN distribution that produces the same sample mean information

as that of the given data. Estimating the correlation subsequently enables an unbiased estimation of the marginal survival or hazard functions of the event of interest. Simulations showed that the proposed method works well over various marginal distributions, copulas, and sizes of the correlation.

Our study provides a potential contribution to the existing literature in that the proposed method is applicable to any parametric bivariate competing risks data, requires no covariate information to estimate the correlation, and shows accurate and precise results where the conventional MLE fails to do so. We expect the current study to have further applications in biomedical time-to-event analyses where dependence between the survival endpoints exist such as disease etiology research or RCTs of drug efficacy.

Keywords: Competing risks survival analysis, Correlation, Dependence, Identifiability, Bivariate central limit theorem.

Chapter 1

Introduction

1.1 Basic Quantities in Survival Analysis

In the biomedical setting, a cohort of patients with some disease are often followed longitudinally until the occurrence of some disease-related event. Survival analysis is the statistical inference of the time to such an event, which differs from other longitudinal studies in that subjects may drop out or become lost to follow-up (f/u) during the study period. This is called '(right-)censoring', and many methods unique to survival analysis are related to dealing with such censoring.

In this regard, we first start with some basic quantities in survival analysis, for which the following notations will be used:

X: the time to some biomedical event of interest, which is a random variable

$F_X(x)$: the cumulative distribution function (cdf) at time X=x. A related quantity is the probability density function (pdf) of X, $f_X(x)$

S(x): the survival function at X=x, which is equal to $1 - F_X(x)$

h(x): the hazard rate (or function) at X=x. A related quantity is the cumulative hazard function H(x) at X=x

Since censoring occurs during or at the end of the study (some patients become right-censored if they do not experience the event of interest until the end of study), the time-to-event random variable X is not always fully observed, i.e. we observe the minimum of two times T: time to the actual event of interest, and C: time to either (right-) censoring or some other competing event(s). Assuming we know each patient's event status of whether he/she has experienced the event of interest during the study, the time-to-event $X = \min(T, C)$, and a status indicator variable $\delta = I(T \leq C)$ is defined. Similarly, we define the other basic quantities above as follows:

The cdf $F_X(x) = \Pr(X \leq x)$

The pdf $f_X(x) = \frac{d}{dx} F_X(x)$

The survival function $S(x) = \Pr(X > x) = 1 - \Pr(X \leq x) = 1 - F_X(x)$

The hazard function $h(x) = \lim_{\Delta x \to 0} \frac{\Pr(x \leq X \leq x + \Delta x \mid X \geq x)}{\Delta x}$

$= \lim_{\Delta x \to 0} \frac{F_X(x+\Delta x) - F_X(x)}{\Delta x} \times \frac{1}{(1 - F_X(x))} = \frac{f_X(x)}{S(x)}$, where any $h(x) \geq 0$

The cumulative hazard function $H(x) = \int_0^x h(u)du$

These basic quantities characterize the (probability) distribution of the time-to-event X (Klein & Moeschberger, 2003), and thus have inter-relationships such as $h(x) = f(x)/S(x)$ (foregoing the random variable subscript X notation in the pdf $f_X(x)$), $S(x) = \exp[-H(x)]$ and vice versa $H(x) = -\log[S(x)]$, where 'log' is the natural log with base e.

Among these basic quantities, we can non-parametrically (or empirically) estimate the survival function S(x) and cumulative hazard function H(x), which are the well-known Kaplan-Meier (K-M) product limit estimator of survival and the Nelson-Aalen (N-A) estimator of cumulative hazard. For notation, let the events of interest occur at D distinct times x_i (i = 1, 2, …, D), $x_1 < x_2 < \ldots < x_D$, and d_i be the number of such events occurring among Y_i number of subjects at risk (the risk-set), at each time x_i. Then, $\frac{d_i}{Y_i} = \Pr(X = x_i \mid X \geq x_i)$, a crude estimate of the hazard rate at an event time x_i, is the basic quantity from which the K-M survival estimator $\widehat{S(x)}$ and the N-A cumulative hazard estimator $\widetilde{H(t)}$ are constructed.

The K-M survival estimator $\widehat{S(x)} = 1$, if $x < x_1$,

$$= \prod_{x_i \leq x}(1 - \frac{d_i}{Y_i}), \text{ if } x \geq x_1.$$

The N-A cumulative hazard estimator $\widetilde{H(x)} = 0$, if $x < x_1$,

$$= \sum_{x_i \leq x} \frac{d_i}{Y_i}, \text{ if } x \geq x_1.$$

From $\widehat{S(x)}$ we can calculate $\widehat{H(x)}$ as $-\log[\widehat{S(x)}]$, and similarly, from $\widetilde{H(x)}$ calculate $\widehat{S(x)}$ as $\exp[-\widetilde{H(x)}]$. A point to be stressed is that any (right-)censoring is considered to be non-informative, or independent, of the main event of interest when estimating survival or cumulative hazard. Thus, any censored subjects can be conveniently subtracted from the denominator Y_i with no contribution to the numerator d_i in the $\widehat{S(x)}$

and $\widetilde{H(x)}$ estimations above.

Using these basic quantities (cdf, pdf, survival, hazard, cumulative hazard) and non-parametric estimators of survival and cumulative hazard, we can expand our case into two or more survival outcomes (time-to-events) that compete with each other, i.e. only the first occurring event's event type and its time-to-event are observed.

1.2 Competing Risks in Survival Analysis

Expanding our perspective to allow more than one event type when observing the time to occurrence of an event, two or more types of events may "compete" with each other to be the first-occurring, and hence, the term "competing risks". Such situations occur frequently in the biomedical setting, where deaths from other causes (heart disease etc.) are competing risks to the time to death from cancer, or death without relapse being a competing risk to the time to relapse in leukemia patients.

Here, we would like to distinguish between the two competing risk situations described above. While theoretically, one can die of only one cause (either from heart disease or from cancer in the above example) such that the two types of events are mutually exclusive, time-to-death and time-to-relapse among leukemia patients are not so. A patient can experience disease relapse and subsequently die, such that both event times may be observed. This situation is termed "semi-" competing risks, while we shall call the situation

of only one type of event and its time-to-event being observable as “classical” competing risks. In the current study, we will focus on this “classical” competing risks situation.

Analogous to the basic quantities described in 1.1, we now define such quantities under competing risks. Notation-wise, the events occur at D distinct times x_i (i = 1, 2, …, D), $x_1 < x_2 < \dots < x_D$, d_i is the number of events occurring among Y_i number of subjects at risk (the risk-set), at each time x_i, but now there are k = 1, 2, …, K different types of “competing” events such that d_{ik} is the number of events of type k occurring at time x_i. The quantities defined below thus pertain to a specific event of type or cause k among K total types or causes possible (such as death due to 1. old age, 2. an accident, 3. heart disease, or 4. cancer, corresponding to four types or causes of death).

The cumulative incidence function (CIF; also called a “sub-” distribution function) of the kth event type $F_k(x) = \Pr(X \le x, \text{type}=k)$

The derivative of the CIF (also called a “sub-” density function) $f_k(x) = \frac{d}{dx} F_k(x)$

The survival function $S(x) = \Pr(X > x) = 1 - \Pr(X \le x) = 1 - \sum_{k=1}^{K} F_k(x)$

The “cause-specific” hazard function $h_k(x) = \lim_{\Delta x \to 0} \frac{\Pr(x \le X \le x+\Delta x, type=k \mid X \ge x)}{\Delta x}$

$= \lim_{\Delta x \to 0} \frac{F_k(x+\Delta x) - F_k(x)}{\Delta x} \times \frac{1}{(1-\sum_{k=1}^{K} F_k(x))} = \frac{f_k(x)}{S(x)}$,

while the “overall” hazard $h(x) = \sum_{k=1}^{K} h_k(x) = \frac{\sum_{k=1}^{K} f_k(x)}{S(x)} = \frac{f(x)}{S(x)}$

The cumulative "cause-specific" hazard function $H_k(x) = \int_0^x h_k(u)du$

Using the sub-density function $f_k(x)$ and the cause-specific hazard $h_k(x)$, the CIF $F_k(x)$ can be expressed as

$F_k(x) = \int_0^x f_k(u)du = \int_0^x S(u) \times h_k(u)du,$

and since $\int_0^\infty S(u) \times h_k(u)du$ is always < 1 due to the cause-specific hazard $h_k(x)$ pertaining only to the kth event type, the CIF is thus also known as a (cumulative) "sub-" distribution function.

From the CIF (or sub-distribution function), Gray (1988) proposed a sub-distribution hazard, from which a semi-parametric regression model of the CIF was developed later on by Fine & Gray (1999). To distinguish between the cause-specific hazard, sub-distribution hazard, and marginal hazard (to be introduced shortly), we will use superscript notation as $h_k^{c-s}(x)$ for the cause-specific hazard, $h_k^{s-d}(x)$ for the sub-distribution hazard, and $h_k^{marg}(x)$ for the marginal hazard of event type (or cause) k at time X=x. Recalling that $h_k^{c-s}(x) = \lim_{\Delta x \to 0} \frac{\Pr(x \le X \le x+\Delta x, type=k \mid X \ge x)}{\Delta x}$, the "sub-distribution" hazard is defined as

$$h_k^{s-d}(x) = \lim_{\Delta x \to 0} \frac{\Pr(x \le X \le x+\Delta x, type=k \mid X \ge x \cup (X<x \cap type \ne k))}{\Delta x} = \frac{f_k(x)}{(1-F_k(x))}$$

$$= -\frac{\mathrm{d}}{\mathrm{dx}} \log\left[1 - \mathrm{F}_k(x)\right] \text{ (Gray, 1988),}$$

where the relationship between the hazard, cumulative hazard, and survival functions are

invoked in the last expression of $h_k^{s-d}(x)$. Intuitively, the sub-distribution hazard adds back to the denominator (or the risk-set) at time x those subjects who had experienced an event, but not of type k, before time x. As the denominator becomes larger due to subjects added back, the calculated hazard becomes smaller, and thus $h_k^{s-d}(x) < h_k^{c-s}(x)$. From this sub-distribution hazard function, the cumulative sub-distribution hazard function follows as $H_k^{s-d}(x) = \int_0^x h_k^{s-d}(u)du$.

We now examine a non-parametric (or empirical) estimator of the CIF, analogous to the K-M estimator of the survival function in 1.1. First, we note that the N-A estimator of the cumulative hazard function naturally extends to the estimators of cumulative "cause-specific" hazard and cumulative "sub-distribution" hazard as $\widetilde{H}_k^{c-s}(x) = \sum_{x_i \le x} \frac{d_{ik}}{Y_i}$ and $\widetilde{H}_k^{s-d}(x) = \sum_{x_i \le x} \frac{d_{ik}}{Y_i^*}$, respectively, where d_{ik} is the number of events of type k occurring at time x_i, and the denominator (or risk-set) Y_i^* is expanded to include subjects whose event time $X < x \cap type \neq k$, as in the sub-distribution hazard definition above. From the expression of $F_k(x) = \int_0^x S(u) \times h_k(u)du = \int_0^x S(u)dH_k u$,

The Aalen-Johansen (A-J) CIF estimator

$\widehat{F_k(x)} = 0$, if $x < x_1$,

$= \sum_{x_i \le x} \widehat{S(x)} \times \widetilde{H}_k^{c-s}(x) = \sum_{x_i \le x} \left[\prod_{j:x_j<x_i}(1 - \frac{d_j}{Y_j}) \times \frac{d_{ik}}{Y_i}\right]$, if $x \ge x_1$.

The A-J CIF estimate is the non-parametric maximum likelihood estimate (NPMLE)

of the true CIF (Pintilie, 2005), and estimation of the sub-density function follows as $\widehat{f_k(x)}$ = $\frac{\mathrm{d}}{\mathrm{dx}}\widehat{F_k(x)}$.

An important point of note is that the usual K-M survival estimator always over-estimates the CIF, because it treats all event types other than the type (or cause) k of interest as right-censored. This can be verified from how $\widehat{S(x)}$ is calculated by the K-M method. Treating all event types ≠ k as independently right-censored, those events would not be included in the numerator of $\frac{d_j}{Y_j}$ (for $j: x_j < x_i$), and $\widehat{S(x)}$ = $\prod_{j:x_j<x_i}(1-\frac{d_{jk}}{Y_j})$, where the number of events occurring at time x_j uses d_{jk} (< d_j) of the kth event type only, rather than d_j of all event types. The K-M method thus over-estimates $\widehat{S(x)}$, and $\widehat{F_k(x)}$ = $\sum_{x_i \le x}\widehat{S(x)} \times \tilde{H}_k^{c-s}(x)$ becomes over-estimated as well. We provide an extreme yet intuitive example of the K-M method over-estimating the CIF: Consider 10 patients waiting for an organ transplant. If 9 die while waiting and the single surviving patient receives a transplant, the K-M method would simply right-censor (and subtract only from the denominator of $\frac{d}{Y}$) those 9 deceased such that the estimated survival from receiving a transplant = 1 - $\frac{d}{Y}$ = $1-\frac{1}{1}$ = 0, resulting in a CIF = 1 – survival = 1 – 0 = 1 (or 100%). In contrast, the A-J CIF estimate of receiving a transplant = $\left(1-\frac{9}{10}\right)\times\frac{1}{1}=0.1$ (or 10%), which is intuitively the right estimate.

1.3 Cause-specific, Sub-distribution, and Marginal hazards

In this section, we restate the definitions of the cause-specific hazard and sub-distribution hazard, and define the "marginal" hazard. Recall that

$$h_k^{c-s}(x) = \lim_{\Delta x \to 0} \frac{\Pr(x \le X \le x+\Delta x, type=k \mid X \ge x)}{\Delta x},$$

$$h_k^{s-d}(x) = \lim_{\Delta x \to 0} \frac{\Pr(x \le X \le x+\Delta x, type=k \mid X \ge x \cup (X < x \cap type \ne k))}{\Delta x}.$$

Now, think of a "marginal" hazard for the event type or cause k, where all other types or causes (1, 2, …, k-1, k+1, …, K) are removed or eliminated. When K competing events are present, the observed survival time $X = \min(T_1, T_2, \ldots, T_K)$, but when types or causes other than k have been eliminated, the observed survival time $X = T_k$. Under this situation where only the single event type k is available to occur, we define the marginal hazard as

$$h_k^{marg}(x) = \lim_{\Delta x \to 0} \frac{\Pr(x \le T_k \le x+\Delta x \mid T_k \ge x)}{\Delta x}.$$

Thus, the marginal hazard is simply a hazard function of some "marginally" distributed T_k among the jointly distributed vector of random variables $(T_1, T_2, \ldots, T_K)$.

Emura et al. (2020) distinguished between these three types of hazard functions in the competing risks setting and characterized the interrelationships between the sub-distribution, cause-specific and marginal hazards, which can be summarized as follows.

$$h_k^{s-d}(x) = \frac{S(x)}{1-\int_0^x S(u) \times h_k^{c-s}(u) du} \times h_k^{c-s}(x),$$

$$h_k^{s-d}(x) = \frac{C_\theta(S_1(x), \ldots, S_K(x))}{1-\int_0^x C_\theta(S_1(u), \ldots, S_K(u)) \times h_k^{marg}(u) du} \times h_k^{marg}(x).$$

Notice that we used a yet undefined notation in $C_\theta(S_1(x), \dots, S_K(x))$ for the relationship between the sub-distribution and marginal hazards, which stands for a "copula" function with copula parameter(s) θ that defines the possible dependence or correlation structure among the K competing risk events. $C_\theta(S_1(x), \dots, S_K(x))$ is thus the multivariate "joint" survival probability, and the definition and further use of copulas will be detailed in Chapter 3 & beyond. Up to this point, we have implicitly assumed that all competing risk events are mutually independent of each other, and in fact, this mutual independence assumption is the cornerstone of cause-specific hazards estimation and modeling (Klein & Moeschberger, 2003). In other words, it has been implicitly assumed that the overall survival function $S(x) = Pr(T_1 > x, T_2 > x, \dots, T_K > x) = Pr(T_1 > x) \times Pr(T_2 > x) \times \dots \times Pr(T_K > x)$, which is clearly not the case in many real-life situations (Tsiatis, 1975; Klein, 2010). While this unrealistic assumption in cause-specific hazards may be somewhat justified by the classical non-informative censoring argument, the possible dependence among the K competing events has to be dealt with in marginal hazards estimation, since finding a "marginal" distribution is naturally connected to separating a single event time from its dependency within the "joint" distribution of K event times. Therefore, the above relationship between the sub-distribution and marginal hazards includes a copula dependence structure, and we can see that their relationship is essentially the same as that between the sub-distribution and cause-specific hazards, other than the copula notation for "joint" overall survival. In fact, cause-specific hazards estimation is no more than using an independence copula, among abundant families of other copulas, to estimate the marginal

hazards. In contrast, sub-distribution hazard estimation and modeling doesn't suffer from this potential dependency problem, since it does not assume that the competing risks are non-informative or independent of each other.

To further contrast the marginal hazard against the cause-specific hazard, we briefly mention regression modeling with competing risks survival data. For convenience purposes, let's assume the time to an event of interest is T, and the time to all other competing events (including random right-censoring) is C, which then retains the previous notation of the observed time-to-event X = min(T, C) and the status indicator variable $\delta = I(T \leq C)$. Including covariates Z for regression modeling, we obtain the so-called triplet survival data of (X, δ, Z). In the univariate case (one time-to-event of interest with independent right censoring), the monumental Cox proportional hazards (PH) regression (Cox, 1972) utilizes the partial likelihood

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp\left[\beta^T \cdot Z_{(i)}\right]}{\sum_{j \in R(t_i)} \exp\left[\beta^T \cdot Z_j\right]}$$

i = 1, 2, …, D for distinct event times t_i, j = 1, 2, …, n subjects, $R(t_i)$ = the number of subjects at risk (the risk-set) at event times t_i,

to estimate the regression coefficients vector β of a semi-parametric regression model

$$h(x|Z) = h_0(x) \cdot \exp\left[\beta^T \cdot Z\right]$$

For regression modeling of the "cause-specific" hazard $h_k^{c-s}(x)$, the above Cox regression

model is identically applied as

$$h_k^{c-s}(x|Z) = h_0^{c-s}(x) \cdot \exp\,[\beta^T \cdot Z]$$

where all other competing events are simply assumed as independent right-censoring. In the similar spirit of semi-parametric modeling, but with additional weights in the denominator of the likelihood function to add back the subjects who experienced any competing event before experiencing the event of interest (as in the previous definition of the sub-distribution hazard $h_k^{s-d}(x)$), Fine & Gray (1999) developed a proportional "sub-distribution" hazard regression model as

$$h_k^{s-d}(x|Z) = h_0^{s-d}(x) \cdot \exp\,[\beta^T \cdot Z]$$

which essentially models the CIF as

$$F_k(x|Z) = 1 - \exp\,[-\int_0^x h_0^{s-d}(u) \cdot \exp\,\{\beta^T \cdot Z\} du]$$

The modified partial likelihood in the Fine & Gray regression model is

$$\mathrm{L}(\beta) = \prod_{i=1}^{D} \frac{\exp\,[\beta^T \cdot Z_{(i)}]}{\sum_{j \in R(t_i)} w_{ij} \cdot \exp\,[\beta^T \cdot Z_j]}$$

i = 1, 2, …, D for distinct event of interest times t_i, j = 1, 2, …, n subjects, $R(t_i)$ = the number of subjects at risk (the risk-set) at event of interest times t_i, weight $w_{ij} = \frac{\hat{G}(t_i)}{\hat{G}(\min\,(s_j, t_i))}$, where s_j = jth subject's all other competing event times, $\hat{G}(\cdot)$ = K-M survival estimate of the right-censoring distribution.

The weight w_{ij} can be intuitively understood as the conditional probability of survival from censoring (due to all competing events or random right-censoring itself) at event of interest times t_i, given the survival from censoring at times $\min(s_j, t_i)$.

Consensus of the current literature on competing risks regression is to use either the cause-specific or sub-distribution hazard (Lau et al., 2009; Andersen et al., 2012; Wolbers et al., 2014; Austin et al., 2016; Hsu et al., 2017). Since the sub-distribution hazard is not an actual "hazard" in the usual sense (recall why it is called a "sub-" hazard, and also the term "sub-" distribution), interpretation of the model results in terms of covariates' effects on the sub-hazard is not straightforward. Also, since devising a regression model for the CIF was the original intention of Fine & Gray (1999), the sub-hazard model is usually recommended for prognosis or prediction model development purposes. Regarding the cause-specific hazard model, it treats any other competing or censoring events as independent right-censoring, and thus enables one to concentrate on the single event of interest's cause-specific hazard and its hazard ratio (HR) by covariate levels. Thus, the cause-specific hazard model is deemed appropriate for disease etiology research purposes, where the HR and its statistical significance by an exposure or treatment on some biomedical time-to-event is of primary interest. However, what if the assumption of mutual independence among competing or censoring events isn't met? A straightforward example of dependent censoring is when a patient drops out of a clinical trial (right-censored; corresponds to time C) due to deteriorating health in an overall survival (event of interest; corresponds to time T) study, and it is easy to see that assuming independence in this case

of positive dependence over-estimates the “marginal” overall survival function. Hence, when T and C are dependent upon each other, using the “cause-specific” hazard model and its independent censoring assumption results in a biased analysis.

We restate that the cause-specific hazard is also a marginal hazard, with the independence copula assumed to be its dependence structure. The cause-specific hazard is merely a single possibility among many other potential marginal hazards, depending on the actual correlation between T and C. Thus, if strong evidence exists against mutual independence among competing event times, then specifying their dependence structure is required for unbiased estimation modeling of the marginal hazard(s). This leads us to the infamous “non-identifiability of competing risks” dilemma (Tsiatis, 1975; Prentice, 1978; Crowder 1994; Klein, 2010). This dilemma basically states that given the observable data of $X = \min(T, C)$ and $\delta = I(T \leq C)$, there is no way to distinguish independence from dependence of T and C (Klein, 2010). In other words, independent T and C with some marginal hazards produce the same observables of X and δ as dependent T and C with some other marginal hazards would produce. Thus, additional information regarding the dependence structure is needed, and various efforts to overcome this problem (additional covariates information, parametric distribution assumptions, etc.) continue today. A literature review of these approaches toward analyzing dependent competing risks data, with an emphasis on the use of copulas for dependence modeling, is the main topic of Chapter 2. First, we take a closer look at the “non-identifiability of competing risks” dilemma.

1.4 The Non-identifiability Dilemma

A well-known framework in the analysis of competing risks survival data, is that of "latent failure times" (Prentice, 1978). The term "failure" time, as opposed to "event" time, comes from the historical context of viewing survival analysis as "time-to-failure" analysis. Latent failure times assumes a potential event time for each type (or cause) of competing event, regardless of this being realistic or not. For example, if the competing risks pertain to possible causes of death, a patient obviously cannot die from a competing cause if already deceased due to some other cause. However, this framework provides an easy way to conceptualize a joint multivariate distribution of the individual competing risks which are usually mutually exclusive in reality.

Retaining our notation of bivariate competing risks, $X = \min(T, C)$, $\delta = I(T \leq C)$, the "joint" survival function $S(t, c) = Pr(T > t, C > c)$, and the "marginal" survival functions $S(t) = Pr(T > t)$ and $S(c) = Pr(C > c)$ can be expressed with the joint survival function as $Pr(T > t, C > 0) = S(t, 0)$ and $Pr(T > 0, C > c) = S(0, c)$, respectively. Given the actually observable information of (X, δ), our knowledge of the survival functions is limited to $Pr(X > x) = Pr(\min(T, C) > x) = Pr(T > x \mid \delta=1) + Pr(C > x \mid \delta=0)$, a sum of conditional probabilities that involve both the joint and marginal survivals. Thus, given (X, δ), many different possibilities exist for the joint and marginal survival functions, i.e. the survival functions are not "identifiable" from the observable information alone.

Identifiability is defined as follows: For the pdf f of the observables (X, δ) with parameters vector θ, θ is identifiable if any given θ uniquely determines the density f of (X, δ), i.e. if $f_{X,\delta,\theta_1} \equiv f_{X,\delta,\theta_2}$, then $\theta_1 = \theta_2$ (Czado & Van Keilegom, 2022). This can also be thought of as the pdf f of (X, δ) being an injective function, such that $f_{X,\delta,\theta_1} \equiv f_{X,\delta,\theta_2}$ guarantees $\theta_1 = \theta_2$.

Given the observables (X, δ) in a "classical" competing risks situation, we cannot uniquely identify the joint and/or marginal survival functions non-parametrically, i.e., they are "non-identifiable" from (X, δ) alone. The bivariate case was noted by Cox (1959), and Tsiatis (1975) extended the result to the k-dimensional case. Here, we provide an example of such non-identifiability using the bivariate exponential distribution of Gumbel (1960). The joint survival function S(t, c) = Pr(T >t, C >c) is defined as $\exp[-\lambda_T t - \lambda_C c - \rho tc]$ with parameters λ_T, λ_C, and ρ, where ρ denotes the dependency between T ~ Exp(λ_T) and C ~ Exp(λ_C), $0 \leq \rho \leq \lambda_T \lambda_C$. From S(x) = exp[-H(x)], the joint cumulative hazard function H(t, c) = $\lambda_T t + \lambda_C c + \rho tc$, and under the assumption of independence, the cause-specific hazards are $\lambda_T + \rho t$ and $\lambda_C + \rho c$, respectively for T and C. Using these cause-specific hazards to compute the cause-specific survival functions,

$$S_T(t) = \exp[-H_T(t)] = \exp\left[-\int_0^t (\lambda_T + \rho u)\, du\right] = \exp\left[-\lambda_T t - \frac{1}{2}\rho t^2\right],$$

$$S_C(c) = \exp[-H_C(c)] = \exp\left[-\int_0^c (\lambda_C + \rho u)\, du\right] = \exp\left[-\lambda_C c - \frac{1}{2}\rho c^2\right].$$

These results (under "independence"), however, are unequal to the marginal survival

functions for the exponential distribution, $\exp[-\lambda_T t]$ and $\exp[-\lambda_C c]$ (under some "dependence" ρ). The problem is that it is not possible to distinguish between these two different models from the observation of (X, δ) alone, since X = min(T, C) allows only one of either T or C to be observed, and there is seemingly no way to estimate the correlation between T and C. In fact, Tsiatis (1975) states and proves that for any true model with dependent competing risks, there exists an independent competing risks model that yields identical (X, δ) information, which is precisely the non-identifiability dilemma.

1.5 Study Outline and Objectives

The current study aims to investigate methods that enable us to robustly identify the yet known as "unidentifiable" dependence structure in competing risks or dependently censored survival times for which only the minimum (or first-occurring), event time and its event type (or cause) are known. Without loss of generality, we will assume bivariate survival data where the observed survival time X = min(T, C), status indicator $\delta = I(T \leq C)$, and dependence between T and C is allowed, T: time to event of interest, C: time to all other competing or dependent censoring events. This bivariate case can be expanded to the multivariate case of three or more mutually dependent time-to-events.

Further chapters will be comprised of: Chapter 2. Previous Literature in Modeling Dependent Competing Risks Survival Data; Chapter 3. Proposed Method; Chapter 4. Simulation Study; Chapter 5. Real Data Analysis, and Chapter 6. Discussion and

Conclusion.

In summary, the main objectives of this study are:

(i) identify whether dependent censoring is present by explicitly estimating the correlation between T (time to the event of interest) and C (time to any other competing event or dependent censoring),

(ii) use the estimated correlation to estimate the marginal distributions (survival curves, hazard functions) via copula-based methods, and

(iii) unbiasedly estimate a main exposure or treatment's effect on the identified marginal survival or hazard (of the event of interest), which we expect will have further applications in disease etiology research or randomized clinical trials of drug efficacy.

Chapter 2

Previous Literature in Modeling Dependent Competing Risks Survival Data

2.1 Definition and Basic Properties of Copulas

The copula function $C_\theta(S_1(x), \dots, S_K(x))$ with copula parameter(s) θ has been mentioned in Section 1.3 as the multivariate joint survival probability of 1, 2, …, K competing events. The definition, properties, and use of copulas in bivariate (or multivariate) survival analysis will be detailed here.

First, the term "copula" refers to a multivariate cdf for which the marginal distribution of each random variable is the Uniform distribution on the interval [0, 1]. In the latent failure times context, for the random vector (X_1, X_2, …, X_K) of competing event times, the copula of (X_1, X_2, …, X_K) is defined as the joint cdf $\Pr(X_1 \le x_1, X_2 \le x_2, \dots, X_K \le x_K)$. The probability integral transform is applied to each marginal distribution to write the joint cdf in copulae notation as below.

(Lemma) Probability integral transform: Let F be a continuous cdf and let X have the cdf F. Then, $F_X(X)$ follows the standard Uniform distribution, i.e. $F_X(X) \sim \text{Unif}(0,1)$.

$\therefore \Pr(X_1 \le x_1, X_2 \le x_2, \dots, X_K \le x_K) = \Pr(U_1 \le u_1, U_2 \le u_2, \dots, U_K \le u_K) = C_\theta(u_1, u_2, \dots, u_K)$.

The book by Nelsen (2006) includes more precise definitions and properties of copulas, among which some are noted here. Let the unit square $\mathbf{I}^2$ be the product $\mathbf{I}$x$\mathbf{I}$, where $\mathbf{I} = [0,1]$. Then, a two-dimensional copula is a function C from $\mathbf{I}^2$ to $\mathbf{I}$ with the following properties:

(Definition) Two-dimensional copula C:

1) For every u, v in $\mathbf{I}$, $C(u,0) = C(0,v) = 0$, $C(u,1) = u$ and $C(1,v) = v$.

2) For every u_1, u_2, v_1, v_2 in $\mathbf{I}$ such that $u_1 \leq u_2$, $v_1 \leq v_2$,

$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$.

A strict definition of a K-dimensional copula requires it to be "grounded" and "K-increasing", for which the details are referred to Nelsen (2006).

An important cornerstone of copula theory is its bounds from above and below, namely the Frechet-Hoeffding bounds.

(Theorem) Frechet-Hoeffding bounds: For every u, v in $\mathbf{I}^2$,

$\max(u+v-1,0) \leq C(u,v) \leq \min(u,v)$.

The stated bounds can be easily verified in the two-dimensional case using (U, 1-U), which is perfect negative dependence or the counter-monotone copula, and (U, U), which is perfect positive dependence or the co-monotone copula, for $U \sim \text{Unif}(0,1)$.

The central theorem of copulas for most of its statistical applications is the theorem by Sklar (1959), explaining why a copula determines the dependence among the components

of a random vector (Hofert et al., 2018).

(Theorem) Sklar's theorem: Let H be a joint distribution (or cdf) of two marginal cdfs F and G.

1) There exists a copula C such that for all x, y in the real line R, H(x,y) = C(F(x), G(y)), where F(x) can be obtained as $F^{-1}(U=u)$, U ~ Unif(0,1), and G(y) as $G^{-1}(V=v)$, V ~ Unif(0,1), from the copula C(u,v), and $F^{-1}(\cdot)$ and $G^{-1}(\cdot)$ are generalized inverses of F(·) and G(·). If F and G are continuous, then C is unique.

2) Conversely, for H(x,y) with marginals F(x) and G(y), a copula C(u,v) can be constructed as $H(F^{-1}(U=u), G^{-1}(V=v))$.

Part 2) of Sklar's theorem, which provides a method of constructing copulas from joint distributions, is exemplified below from the book by Nelsen (2006). Consider a bivariate joint cdf H with marginal cdfs F and G as

$H(x,y) = \frac{(x+1)(e^y-1)}{(x+2e^y-1)}, for\ (x,y) \in [-1,1] \times [0,\infty],$

$= 1 - e^{-y}, for\ (x,y) \in (1,\infty] \times [0,\infty],$

= 0, elsewhere.

F(x) = 0, for x < -1, G(y) = 0, for y < 0,

= (x+1)/2, x ∈ [-1,1], $= 1 - e^{-y}$, y ≥ 0.

= 1, x > 1.

Then, the quasi-inverses of F and G are $F^{-1}(u) = 2u-1$, $G^{-1}(v) = -\ln(1-v)$, for u, v $\in$ **I**, and the copula $C(u,v) = H(F^{-1}(u), G^{-1}(v)) = \frac{2u(\frac{1}{(1-v)}-1)}{\{2(u-1)+\frac{2}{(1-v)}\}} = \frac{uv}{(u+v-uv)}$. While C(u,v) = Pr(U ≤u, V ≤v) is a cdf, we are more interested in the "survival" function Pr(U >u, V >v) in bivariate (or multivariate) survival analysis. Inserting marginal survival functions Pr(T >t) and Pr(C >c) for time-to-events T and C in a copula thus provides a useful measure such that C(Pr(T >t)=some s, Pr(C >c)=some t) = Pr(U ≤s, V ≤t) = C(s,t) = Pr(T >t, C >c), the joint survival function of interest.

A major application of copulas in multivariate survival analysis is the construction of dependent (or correlated) survival times for simulation purposes. For simplicity, we consider the case of bivariate survival times that are correlated through a copula, and introduce a general method of generating such dependent survival times, which is known as the "conditional distribution (cdf)" method (Nelsen, 2006). First, we give a theorem that guarantees the obtention of the "conditional" cdf for any given copula.

(Theorem) For a copula C and $\forall\, v \in \boldsymbol{I}, \frac{\partial}{\partial u} C(u,v)$ exists for almost all u, and satisfies $0 \leq \frac{\partial}{\partial u} C(u,v) \leq 1$. Here, the copula partial derivative $\frac{\partial}{\partial u} C(u,v)$ is a conditional cdf, since $\frac{\partial}{\partial u} C(u,v) = \lim_{\Delta u \to 0} \frac{1}{\Delta u}\{C(u+\Delta u, v) - C(u,v)\}$ = Pr(V ≤v | U=u).

Writing $\frac{\partial}{\partial u} C(u,v)$ as $c_u(v)$, generate a pair of data (u,v) correlated via C(u,v) as:

i) Generate instances of two independent Unif(0,1) random variables u and t.

ii) Calculate v using $c_u(v) = t$, and $v = c_u^{-1}(t)$.

We illustrate the conditional distribution method using the example of finding the copula $C(u,v) = H(F^{-1}(u), G^{-1}(v)) = \frac{uv}{(u+v-uv)}$ from the joint cdf H and marginals F and G of page 22 above. From the copula function, the partial derivative $c_u(v) = \frac{1}{(u+v-uv)^2} \times \{v(u+v-v) - uv(1-v)\} = (\frac{v}{u+v-uv})^2$, from which $c_u^{-1}(t) = v = \frac{u\sqrt{t}}{\{1-(1-u)\sqrt{t}\}}$. Now, generate the correlated random variates (x,y) whose marginal distributions are originally Unif(-1,1) and Exp(1), as:

i) Sample two independent Unif(0,1) random variates u and t.

ii) Set $v = \frac{u\sqrt{t}}{\{1-(1-u)\sqrt{t}\}}$.

For $Y \sim Exp(1)$, the cdf $G(y) = 1-e^{-y}$, and thus $G^{-1}(v) = y = -\ln(1-v)$.

For $X \sim Unif(-1,1)$, the cdf $F(x) = (x+1)/2$, and thus $F^{-1}(u) = x = 2u-1$.

iii) Setting $x = 2u-1$, $y = -\ln(1-v)$, where $v = \frac{u\sqrt{t}}{\{1-(1-u)\sqrt{t}\}}$, (x,y) is the desired pair, correlated via $C(u,v) = \frac{uv}{(u+v-uv)}$ (Nelsen, 2006).

We will use this conditional distribution method throughout our study's simulations of dependent censoring in Chapter 4.

Among the many different copula functions that currently exist, two major families of copulas are the elliptical and Archimedean copulas. First, elliptical copulas are the copulas

of elliptical distributions, where an elliptical distribution is defined as

(Definition) Elliptical distribution: A k-dimensional random vector X has an elliptical distribution with location vector $\mu \in \mathbb{R}^k$, scale matrix $\Sigma = AA^t$ with rank(Σ) = r ≤k for a matrix A $\in \mathbb{R}^{k\times r}$ and radial part R ≥0 if $X \equiv \mu + AY, for\ Y \equiv RS$, where S is uniformly distributed on the unit sphere in $\mathbb{R}^k$, R and S independent (Hofert et al., 2018).

The most commonly used elliptical copula is the Normal or Gaussian copula, which is a copula with its cdf being the cdf of a multivariate normal distribution. In the bivariate case, a Normal copula $C(u,v) = Pr(U \le u, V \le v) = Pr(\Phi_T(T) \le u, \Phi_C(C) \le v) = Pr(T \le \Phi_T^{-1}(u), C \le \Phi_C^{-1}(v))$, for the standard normal distribution cdf Φ, where the bivariate Normal pdf

$$f_{T,C}(t,c) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\{t^2 - 2\rho tc + c^2\}\right], \text{ for } (t,c) \in \mathbb{R}^2,\ \rho \in [-1,1],$$

such that the bivariate Normal cdf, or the Normal copula, is

$$Pr(T \le \Phi_T^{-1}(u), C \le \Phi_C^{-1}(v)) = \int_{-\infty}^{\Phi_T^{-1}(u)} \int_{-\infty}^{\Phi_C^{-1}(v)} f_{T,C}(t,c)\, dcdt =$$

$$\frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\Phi_T^{-1}(u)} \int_{-\infty}^{\Phi_C^{-1}(v)} \exp\left[-\frac{1}{2(1-\rho^2)}\{t^2 - 2\rho tc + c^2\}\right] dcdt.$$

Second, Archimedean copulas are also widely applied due to the ease with which they can be constructed, and the great variety of copulas that belong to this class (Nelsen, 2006). To define the Archimedean copula, let X and Y be continuous random variables with joint cdf H and marginal cdfs F and G. In the case where X, Y were independent, then H(x,y) = F(x)·G(Y). Now, for the general case of dependent X and Y, suppose some function λ(·), its

range positive in [0,1], enables $\lambda(H) = \lambda(F)\cdot\lambda(G)$. Then setting $\varphi(\cdot) = -\ln\{\lambda(\cdot)\}$ or $\varphi(t) = -\ln\{\lambda(t)\}$, $\varphi(H) = -\ln\{\lambda(H)\} = -\ln\{\lambda(F)\cdot\lambda(G)\} = -\ln\{\lambda(F)\} - \ln\{\lambda(G)\} = \varphi(F) + \varphi(G)$. That is, a function of H (the joint cdf) can be expressed as a function of F and G (the marginal cdfs), and expressing H as $\varphi^{-1}\{\varphi(F) + \varphi(G)\}$ derives a copula function, named an Archimedean copula.

$\therefore H = \varphi^{-1}\{\varphi(F) + \varphi(G)\} \iff C(F(X),G(Y)) = C(u,v) = \varphi^{-1}\{\varphi(F(X)) + \varphi(G(Y))\}$

$= \varphi^{-1}\{\varphi(u) + \varphi(v)\}$.

The function $\varphi(\cdot)$ is called the generator function of an Archimedean copula, and the three most popular Archimedean copula functions derived from their respective generator functions are now introduced: the Clayton, Gumbel, and Frank copulas.

(Definition) Clayton copula: Consider the generator function $\varphi(t) = \frac{1}{\theta}\left(t^{-\theta} - 1\right)$. Then, $\varphi^{-1}(t) = (\theta\varphi + 1)^{\frac{-1}{\theta}}$, and the corresponding Archimedean copula generation of $\varphi^{-1}\{\varphi(u) + \varphi(v)\}$ results in $[\theta\left\{\frac{1}{\theta}\left(u^{-\theta} - 1\right) + \frac{1}{\theta}\left(v^{-\theta} - 1\right)\right\} + 1]^{\frac{-1}{\theta}} = \left(u^{-\theta} + v^{-\theta} - 1\right)^{\frac{-1}{\theta}}$, which is the Clayton copula with copula parameter $\theta \in [-1, \infty)\backslash\{0\}$.

(Definition) Gumbel copula: Consider the generator function $\varphi(t) = (-lnt)^{\theta}$.

Then, $\varphi^{-1}(t) = \exp(-t^{\frac{1}{\theta}})$, and $\varphi^{-1}\{\varphi(u) + \varphi(v)\} = \exp[-\left\{(-lnu)^{\theta} + (-lnv)^{\theta}\right\}^{\frac{1}{\theta}}]$, which is the Gumbel copula with copula parameter $\theta \in [1, \infty)$.

(Definition) Frank copula: Consider the generator function $\varphi(t) = -\ln\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right)$.

Then, $\varphi^{-1}(t) = -\frac{1}{\theta}\ln\{e^{-t}(e^{-\theta}-1)+1\}$, and $\varphi^{-1}\{\varphi(u) + \varphi(v)\} = -\frac{1}{\theta}\ln[1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}]$, which is the Frank copula with copula parameter $\theta \in (-\infty, \infty)\backslash\{0\}$.

In the current study, the Normal, Clayton, Gumbel, and Frank copulas introduced above will be used in generating dependent (or correlated) time-to-events data and the subsequent analysis of such data to unbiasedly estimate the marginal hazard of the event of interest.

The direction and degree of dependence (or correlation) between the time-to-events T and C, corresponding to the event of interest and all other competing or dependent censoring events, respectively, is another very important measure that must be defined. The usual Pearson correlation coefficient is rarely used in practice due to its many fallacies and its measuring of only a 'linear' trend (Hofert et al., 2018). Instead, non-parametric rank-correlation coefficients such as Kendall's tau or Spearman's rho are preferred, among which Kendall's tau will be described and used here.

Kendall's tau is defined in terms of concordance, where two pairs of observed data (t_i,c_i) and (t_j,c_j) are 'concordant' if $t_i < t_j$ and $c_i < c_j$ or $t_i > t_j$ and $c_i > c_j$. In contrast, the pairs are 'discordant' if $t_i < t_j$ and $c_i > c_j$ or $t_i > t_j$ and $c_i < c_j$. Among a random sample of n observations from a vector (T, C) of continuous survival times, there exist a total of ${}_nC_2$ distinct pairs of observations in the sample, and letting c denote the number of 'concordant'

pairs and d denote the number of 'discordant' pairs, Kendall's tau for this sample is defined as: Kendall's tau $\tau = \frac{c-d}{c+d} =$ (c-d) / ${}_nC_2$. From $\tau = \frac{c-d}{c+d}$, we see that it is the difference between the probability of concordance and discordance in a random sample of bivariate survival times, and thus can be stated more formally as below.

(Theorem) Equivalent expressions for Kendall's tau:

Let (X_1, Y_1) and (X_2, Y_2) be independent vectors of continuous random variables with joint cdfs H_1 and H_2, respectively, with common margins F (of X_1 and X_2) and G (of Y_1 and Y_2). Let C_1 and C_2 denote the copulas of (X_1, Y_1) and (X_2, Y_2), respectively, such that $H_1(x,y) = C_1(F(x), G(y))$ and $H_2(x,y) = C_2(F(x), G(y))$. Let τ denote the difference between the probabilities of concordance and discordance of (X_1, Y_1) and (X_2, Y_2), i.e.,

$\tau = \Pr[(X_1 - X_2)(Y_1 - Y_2) > 0] - \Pr[(X_1 - X_2)(Y_1 - Y_2) < 0]$. This is equivalently expressed using the copulas C_1 and C_2 as $\tau = 4 \iint_{I^2} C_2(u,v) dC_1(u,v) - 1$.

Since $\iint_{I^2} C_2(u,v) dC_1(u,v) = \int_0^1 \int_0^1 C_2(u,v) \frac{\partial^2}{\partial u \partial v} C_1(u,v) du dv$, this is again equivalent to $\tau = 1 - 4 \iint_{I^2} \frac{\partial}{\partial u} C_2(u,v) \frac{\partial}{\partial v} C_1(u,v) du dv$.

For an Archimedean copula generated by $\varphi(\cdot)$ such that $C(u,v) = \varphi^{-1}\{\varphi(u) + \varphi(v)\}$, Kendall's tau can be found as $\tau = 1 + 4 \int_0^1 \frac{\varphi(t)}{\varphi'(t)} dt$ (Nelsen, 2006).

Hence, Kendall's tau can be found via its relationship with an Archimedean copula's dependence parameter θ, shown in Table 3.1 below for the Clayton, Gumbel, and Frank

copulas, and also for the Normal (or Gaussian) copula with dependence parameter $\rho \in [-1,1]$.

Table 2.1 Relationship between Kendall's tau and the copula dependence parameter in several parametric copulas

Copula	Copula dependence parameter	Kendall's tau
Normal (Gaussian)	$\rho \in [-1,1]$	$\frac{2}{\pi}\arcsin(\rho)$
Clayton	$\theta \in [-1,\infty)\backslash\{0\}$	$\frac{\theta}{2+\theta}$
Gumbel	$\theta \in [1,\infty)$	$\frac{\theta-1}{\theta}$
Frank	$\theta \in (-\infty,\infty)\backslash\{0\}$	$1-\frac{4}{\theta}\{1-D_1(\theta)\}$, Debye function $D_k(\theta) = \frac{k}{\theta^k}\int_0^\theta \frac{t^\theta}{e^t-1}dt$

2.2 The Assumed Copula Approach to Estimate the Marginal Hazards

As copulas are a natural tool in modeling multivariate dependence structures, they have been widely adapted in modeling the possible dependence in competing risks survival data. A seminal work in this area has been done by Zheng and Klein (1994, 1995), where they proposed an 'assumed' copula to first determine the dependence structure of the given

competing risks data, which then leads to unbiased identifiability of the marginal hazards of the competing risk events. More specifically, assuming that the copula is known, the authors used the notion of self-consistency to construct an estimator of the marginal survival functions based on dependent competing risks data (Zheng & Klein, 1994). This concept of self-consistency has been used by Efron (1976) to derive the K-M estimator in the case of 'independent' competing risks, or bivariate survival, data.

For a bivariate survival time sample (T_i, C_i), i = 1, 2, …, n subjects, nonparametric estimators of the 'marginal' survival functions of T and C (with marginal cdfs F and G, respectively) are denoted as $\widehat{Pr}(T \geq t) = \hat{S}(t) = \sum_{i=1}^{n} \mathrm{I}(T_i \geq t)$ and $\widehat{Pr}(C \geq c) = \hat{R}(c) = \sum_{i=1}^{n} \mathrm{I}(C_i \geq c)$. When the status indicator variable $\delta = \mathrm{I}(T \leq C) = 0$, one does not observe T_i but only knows that $T_i > X_i = \min(T_i, C_i)$. If $x \leq X_i < T_i$, then one is certain that $I(T_i \geq x)$=1 to count towards the survival estimator, but if $X_i < x$, one doesn't know whether the unobservable T_i will be $\geq x$ or not. Thus, the probability that needs to be estimated is $\Pr(T \geq t \mid X_i = x_i < t, \delta_i = 0) = \Pr(T \geq t \mid T > x_i, C = x_i)$, and likewise, $\Pr(C \geq c \mid C > x_i, T = x_i)$, from the viewpoint of the dependent censoring or competing event time C. The self-consistent marginal survival estimators at some time x are

$$\hat{S}_{new}(x) = \frac{1}{n}\sum_{i=1}^{n}\{I(X_i \geq x) + I(X_i < x) \cdot (1 - \delta_i) \cdot \widehat{\Pr}_{old}(T \geq x \mid T > x_i, C = x_i)\}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\{I(X_i \geq x) + I(X_i < x) \cdot (1 - \delta_i) \cdot \frac{1 - C_v(1 - \hat{S}_{old}(x), 1 - \hat{R}_{old}(x_i))}{1 - C_v(1 - \hat{S}_{old}(x_i), 1 - \hat{R}_{old}(x_i))}\}$$

where X_i (= x_i) is the event time of subject i, $C_v(u,v) = \frac{\partial}{\partial v}C(u,v) = \Pr(U \le u \mid V = v)$.

This can be shown by $\widehat{Pr}(T \ge x \mid T > x_i, C = x_i) = \frac{\widehat{Pr}(T>x, C=x_i)}{\widehat{Pr}(T>x_i, C=x_i)}$, where

$$\widehat{Pr}(T > x, C = x_i) = \int I[(u,v)|F(x) < u \le 1, G(x_i \le v < G(x_i + \theta)] \cdot c(u,v)dudv$$

$$= \Pr\big(u > F(x), v > G(x_i)\big) - \Pr\big(u > F(x), v > G(x_i + \theta)\big)$$

$$= C\big(F(x), G(x_i)\big) - C\big(F(x), G(x_i + \theta)\big) + G(x_i + \theta) - G(x_i),$$

and similarly, $\widehat{Pr}(T > x_i, C = x_i) = C\big(F(x_i), G(x_i)\big) - C\big(F(x_i), G(x_i + \theta)\big) + G(x_i + \theta) - G(x_i)$. Then,

$$\lim_{\theta \to 0} \frac{\widehat{Pr}(T > x, C = x_i)}{\widehat{Pr}(T > x_i, C = x_i)} = \lim_{\theta \to 0} \frac{-\left[\frac{\{C\big(F(x), G(x_i)\big) - C\big(F(x), G(x_i + \theta)\big)\}}{\{G(x_i + \theta) - G(x_i)\}}\right] + 1}{-\left[\frac{\{C\big(F(x_i), G(x_i)\big) - C\big(F(x_i), G(x_i + \theta)\big)\}}{\{G(x_i + \theta) - G(x_i)\}}\right] + 1}$$

$$= \frac{1 - \frac{\partial}{\partial v}C(F(x), G(x_i))}{1 - \frac{\partial}{\partial v}C(F(x_i), G(x_i))} = \frac{1 - C_v(1 - S(x), 1 - R(x_i))}{1 - C_v(1 - S(x_i), 1 - R(x_i))}.$$

Following a similar procedure as above,

$$\hat{R}_{new}(x) = \frac{1}{n}\sum_{i=1}^{n}\{I(X_i \ge x) + I(X_i < x) \cdot \delta_i \cdot \widehat{\Pr}_{old}(C \ge x \mid C > x_i, T = x_i)\}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\{I(X_i \ge x) + I(X_i < x) \cdot \delta_i \cdot \frac{1 - C_v(1 - \hat{S}_{old}(x_i), 1 - \hat{R}_{old}(x))}{1 - C_v(1 - \hat{S}_{old}(x_i), 1 - \hat{R}_{old}(x_i))}\}.$$

Thus, first beginning with an initial guess (usually by simply assuming independence between T and C) of $\hat{S}_{old}(x)$ and $\hat{R}_{old}(x)$, iteratively update to $\hat{S}_{new}(x)$ and $\hat{R}_{new}(x)$ until convergence. The point to stress is that the joint survival function of T and C is a function of the copula and the unknown 'marginal' survival functions S(x) and R(y), and that these marginal survival functions can be estimated by the above self-consistent equations when the copula is assumed to be known.

In Zheng and Klein (1995), the estimation of marginal survival functions is provided with a graphical tool named the "copula-graphic estimator" for dependent competing risks data, which reduce to the usual K-M estimator when the bivariate survival times are independent. Again, it is shown that the marginal survival functions are identifiable if the copula (dependence structure) is assumed to be known, and an iterative bisection root-finding algorithm is presented with a graphical tool that depicts the relationship between F(x) of the survival time T and G(x) of the dependent competing risk or censoring time C on a unit square. The resulting estimators $\hat{F}_n$ and $\hat{G}_n$ are strongly consistent for F and G such that as $n \to \infty, \hat{F}_n \to F$ and $\hat{G}_n \to G$ for all $x \in [0, \infty)$. The authors noted in the results of their simulation study that the important requirement for a good estimate of the marginal survival function is a reasonable guess of the strength of association between T and C (hence, an 'assumed' copula), and not the functional form of the copula.

2.3 Likelihood-based Semi-parametric Modeling Under an Assumed Copula

Following the works of Zheng and Klein, Huang and Zhang (2008) applied the assumed copula & self-consistent estimator of marginal survival to a Cox PH regression setting. More systematically, Chen (2010) extended the assumed copula approach to the regression setting by generalizing the marginal regressions via semiparametric transformation models, which include both proportional hazards and proportional odds. Under the premise that the marginal and/or joint distributions of competing event times cannot be identified without additional information (the non-identifiability problem), both the functional form of the copula and the copula parameter(s) that control the level of association between event times were assumed to be known. Similar to the comments of earlier works (Zheng and Klein, 1994, 1995; Huang and Zhang, 2008), Chen found that the proposed model was robust to the functional form of the copula (e.g. whether the underlying copula was Normal, Clayton, Gumbel etc.) but sensitive to the assumed level of association (e.g. the copula parameter value that directly corresponds to Kendall's tau; Table 3.1).

Using the author's notation, T_1^* and T_2^* are two dependent event times for two competing events, and C(u,v), u, v $\in [0,1]$, is a chosen copula function where the joint survival function of T_1^* and T_2^* is $\Pr(T_1^* > t_1, T_2^* > t_2) = C\{S_1(t_1), S_2(t_2)\}$, $t_1 \geq 0$, $t_2 \geq 0$. The functional form and the association parameter for C are assumed to be known. For two sets

of covariates Z_1 and Z_2, Chen conducted marginal regression analyses to estimate the effect of Z_k on T_k^* (k=1, 2). The two marginal regressions are specified as the following semiparametric transformation models for k=1, 2:

$$\Lambda_k(t;\beta_k,R_k) = G_k[\int_0^t I(T_k^* \geq s)\cdot \exp\{\beta_k^T Z_k(s)\}dR_k(s)],$$

where $\Lambda_k(t;\beta_k,R_k)$ is the cumulative intensity (or hazard) function for the counting process $N_k^*(t) = I(T_k^* \leq t)$, G_k is some specified strictly increasing and differentiable transformation function, and R_k is an unspecified increasing function to be estimated from the data (or some baseline hazard function).

Also, for subject i, i = 1, 2, …, n, the observed data triplet are $\{T_i, \delta_{ki}, Z_{ki}(t)$; k=1, 2, $0 \leq t \leq$ maximum follow-up time $t_{end}\}$, the counting process of an observed event k (=1 or 2) for subject i is $N_{ki}(t) = \delta_{ki}\cdot I(T_i \leq t) = I(T_i = T_{ki}^*)\cdot I(T_i \leq t)$ such that δ_{ki} is subject i's event 1 or 2 status indicator, and $I(T_i \leq t)$ is a counter for subject i's survival time being $\leq t$. The corresponding at-risk process is $Y_i(t) = I(T_i \geq t)$, which indicates whether subject i is still at risk at time t or beyond.

After specifying the marginal regression model's functional form, Chen made an important comment that the focus on "marginal" or "net" event time regression analysis implies the hypothetical setting of artificially removing the other competing risk, i.e. the assessment of covariate effects on either one of event k=1 or k=2, but having the other correlated event removed. This is precisely what our current study aims to achieve, e.g. the

effect of a new drug in a randomized clinical trial under the ideal scenario of no patient being dependently censored due to one's deteriorating or improving health condition. Chen also mentioned that this counterfactual scenario of removing the competing risk event(s) may have controversial interpretation issues (Prentice, 1978), which was also noted in our study's outline of the 'latent failure times' framework (section 1.4).

To describe the above semiparametric transformation model in more detail, consider the following cause-specific (or marginal) intensity (or hazard) for the counting process $N_{ki}(t)$, k=1, 2:

$$Y_i(t) \cdot \exp\{\beta_k^T Z_{ki}(t)\} \cdot \eta_{ki}(t-;\beta,R) \cdot r_k(t),$$

where $r_k(t) = R_k'(t)$, $\eta_{ki}(t-;\beta,R) \equiv \eta_k^*[\int_0^t Y_i(u) \cdot \exp\{\beta_1^T Z_{1i}(u)\} \cdot dR_1(u), \int_0^t Y_i(u) \cdot \exp\{\beta_2^T Z_{2i}(u)\} \cdot dR_2(u)]$,

for $\eta_k^*(t_1,t_2) = \frac{\partial}{\partial t_k}\Phi[\exp\{-G_1(t)\}, \exp\{-G_2(t)\}]$

$= g_k(t_k) \cdot \exp\{-G_k(t_k)\} \cdot D_k[\exp\{-G_1(t_1)\}, \exp\{-G_2(t_2)\}]$,

$\Phi = -\log(\mathrm{C}), g_k = G_k', D_k(u_1,u_2) = -\frac{\partial}{\partial u_k}\Phi(u_1,u_2)$.

Thus, $\exp\{-G_k(t)\}$ being a survival function, $\Phi[\exp\{-G_1(t)\}, \exp\{-G_2(t)\}]$ being a joint cumulative hazard, $\frac{\partial}{\partial t_k}\Phi[\exp\{-G_1(t)\}, \exp\{-G_2(t)\}]$ is some form of marginal hazard that considers the kth event's correlation with the (k$\pm$1)th event. Hence, the whole expression $Y_i(t) \cdot \exp\{\beta_k^T Z_{ki}(t)\} \cdot \eta_{ki}(t-;\beta,R) \cdot r_k(t)$ is a marginal semiparametric

transformation model, with $r_k(t)$ and $\exp\{\beta_k^T Z_{ki}(t)\}$ corresponding to a Cox model's baseline hazard and proportional hazards of covariates formulation.

Using this marginal semiparametric transformation model, Chen provides the log-likelihood function for parameters $\{dR\}$ and β, for which we provide the original likelihood function using the general likelihood formulation of

$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} \cdot S(t_i)^{1-\delta_i} = \prod_{i=1}^{n} h(t_i)^{\delta_i} \cdot S(t_i)$, where

$f(\cdot): pdf, S(\cdot): survival\ function, h(\cdot): hazard\ function, \delta: event\ status\ indicator$

in time-to-event regression models.

$$L = \prod_{i=1}^{n} [\prod_{k=1}^{2} \int_0^{t_{end}} \exp\{\beta_k^T Z_{ki}(t)\} \cdot \eta_{ki}(t-; \beta, R) \cdot dR_k(t) \cdot dN_{ki}(t)$$
$$\cdot \exp\{-\Phi[\exp\{-\Lambda_{1i}(t_{end}; \beta_1, R_1)\}, \exp\{-\Lambda_{2i}(t_{end}; \beta_2, R_2)\}]\}],$$

such that the first integral portion of the likelihood expresses the joint hazard function, and the second exp{-Φ(·)} portion expresses the joint survival function.

Utilizing this likelihood function, one can obtain the score functions (first-order partial derivative vectors of the parameters of interest) and information matrices (second-order partial derivative matrices of the parameters) for estimation and statistical inference (point estimation, interval estimation, testing, and P-values). Chen provides explicit expressions for the score functions and information matrices and shows that the subsequent maximum likelihood estimations result in unbiased, consistent, and asymptotically normal (weak

convergence to a zero-mean Gaussian process) parameter estimates.

Conclusively, Chen (2010) presented a systematic way to conduct marginal regression analysis for dependent censoring or competing risks, applicable to the broad class of semiparametric transformation models, when the dependence structure is completely specified through an assumed copula. Chen also made an important note that in the presence of regression covariates, although the copula association parameter may be estimated from the data along with the regression coefficients (Heckman & Honore, 1989), the variability of the resulting estimates is quite large, i.e. the estimates are unstable, and is not recommended. Thus, one is still left with the question of how accurate the 'assumed' copula would be and how to possibly estimate the copula association parameter if a reasonable guess of it is challenging.

Subsequent contributions to copula-based dependent competing risks analysis were made by Emura and Chen (2016, 2018). Emura and Chen (2016) applied Chen's copula-based framework to gene selection for survival data with dependent censoring, where the authors noted that no practical methods exist to date that simultaneously estimate the copula association parameter and the marginal regression models. However, a novel approach to estimating the copula parameter through cross-validation of a survival time prediction model was proposed, where the copula parameter that maximizes a k-fold cross-validated Harrell's c-index is chosen to be the corresponding parameter of the given data. An important comment here is that due to the non-identifiability of competing risks data, the usual likelihood equation may provide little information about the true association

parameter (or corresponding Kendall's tau correlation), rescinding maximum likelihood-based approaches. An implementation of the authors' work was also provided as an R package "compound.Cox", which we utilize in our simulations and analyses as well. The more recent Emura and Chen (2018) is a textbook on copula-based approaches to survival data with dependent censoring, and to the best of our knowledge, seems to be the first textbook to comprehensively cover the topic.

2.4 Parametric Identifiability of Copulas and Marginal Distributions

Van Keilegom et al. (2013, 2019, 2020, 2021, 2022, 2023) have extensively studied the identifiability of dependent competing risks data using parametrically specified copulas and marginal distributions. Schwarz, Jongbloed, and Van Keilegom (2013) mainly studied the mathematical conditions under which some parametric copulas may be used to identify the joint distribution of (T, C), $X = \min(T, C)$, $\delta = I(T \leq C)$. The more recent Deresa and Van Keilegom (2019) substantially formulated the parametric identifiability approach, utilizing monotone increasing transformations to obtain bivariate normally distributed linear regression error terms, for which the bivariate normal distribution is known to be identifiable under bivariate competing risks (Nadas, 1971; Basu and Ghosh, 1978).

Following the authors' notation, T and C are log-transformed event of interest and

dependent censoring times, for which the proposed joint regression model is

$$\Lambda_\theta(T) = X^t\beta + \varepsilon_T$$

$$\Lambda_\theta(C) = W^t\eta + \varepsilon_C,$$

where X and W are covariates of dimension p and q that are associated with T and C, respectively, and $\Lambda_\theta(\cdot)$ is a parametric class of monotone transformations that preserves the rank of its domain (the Yeo-Johnson family of power transformations). The vector of error terms (ε_T, ε_C) has a bivariate normal (BVN) distribution such that

$$\binom{\varepsilon_T}{\varepsilon_C} \sim N_2\left(\binom{0}{0}, \Sigma = \begin{pmatrix} \sigma_T^2 & \rho\sigma_T\sigma_C \\ \rho\sigma_T\sigma_C & \sigma_C^2 \end{pmatrix}\right),\ \Sigma\text{: positive definite.}$$

The observed event time Z = min(T, C), event status indicator δ = I(T ≤ C), and the dataset has n independent and identically distributed (i.i.d.) observations of $(\mathrm{Z}_i, \delta_i, \mathrm{X}_i, \mathrm{W}_i), i = 1,2,\dots,n$. The parameter vector for estimation $\alpha = (\theta, \beta, \eta, \sigma_T^2, \sigma_C^2, \rho) \in \mathbb{R}^{p+q+4}$, for which the authors showed is identifiable from the observed data, and noted that this is non-trivial, since for a given subject we observe either T or C but never both.

For the BVN distributed error terms linear model above, the conditional cdfs and pdfs (given the covariates X and W) are expressed as

$$\begin{cases} F_{T|X}(t|x) = F_{\epsilon_T}(\Lambda_\theta(t) - x^t\beta) = \Phi\left(\dfrac{\Lambda_\theta(t) - x^t\beta}{\sigma_T}\right) \\ F_{C|W}(c|w) = F_{\epsilon_C}(\Lambda_\theta(c) - w^t\eta) = \Phi\left(\dfrac{\Lambda_\theta(c) - w^t\eta}{\sigma_C}\right) \end{cases},$$

$$\begin{cases} f_{T|X}(t|x) = \sigma_T^{-1} \cdot \phi(\frac{\Lambda_\theta(t) - x^t\beta}{\sigma_T}) \cdot \Lambda_\theta(t)' \\ f_{C|W}(c|w) = \sigma_C^{-1} \cdot \phi(\frac{\Lambda_\theta(w) - w^t\eta}{\sigma_C}) \cdot \Lambda_\theta(w)' \end{cases}.$$

Recall that the sub-distribution function $F_{Z,\delta}(z,1) = \Pr(Z \le z, \delta = 1) = \Pr(\mathrm{T} \le \mathrm{t}, \mathrm{T} \le \mathrm{C}) = \Pr(C \ge T \mid T \le z) \cdot \Pr(T \le z) = \int_0^z \Pr(C \ge u \mid T = u) \cdot f_T(u)du$, and for BVN distributed random variables X_1 and X_2, the conditional distribution of $X_2 \mid X_1 = x_1 \sim N(\mu_2 + \frac{\rho(x_1-\mu_1)\sigma_2}{\sigma_1}, {\sigma_2}^2(1-\rho^2))$, for BVN parameters $(\mu_1,\ \mu_2,\ {\sigma_1}^2,\ {\sigma_2}^2,\ \rho)$ (Hogg, McKean, & Craig, 2013).

Then, the sub-distribution and sub-density functions for the proposed model are

$$\begin{aligned} F_{Z,\delta|\mathrm{X},\mathrm{W}}(z,1|x,w;\alpha) &= \Pr(Z \le z, \delta = 1|\mathrm{X} = \mathrm{x}, \mathrm{W} = \mathrm{w}) \\ &= \Pr(\Lambda_\theta(T) \le \Lambda_\theta(z), \Lambda_\theta(T) \le \Lambda_\theta(C)|X = x, W = w) \\ &= \int_{-\infty}^{\Lambda_\theta(z)-x^t\beta} \Pr(\varepsilon_C \ge e + x^t\beta - w^t\eta \mid \varepsilon_T = e) \cdot f_{\varepsilon_T}(e)de \\ &= \frac{\lambda}{\sigma_T}\int_{-\infty}^{\Lambda_\theta(z)-x^t\beta} [1 - \Phi(\frac{e + x^t\beta - w^t\eta - \rho\frac{\sigma_C}{\sigma_T}e}{\sigma_C\sqrt{1-\rho^2}})]\phi(\frac{e}{\sigma_T})de, \end{aligned}$$

$$f_{Z,\delta|\mathrm{X},\mathrm{W}}(z,1|x,w;\alpha) = \frac{1}{\sigma_T}\left[1 - \Phi\left(\frac{\Lambda_\theta(z) - w^t\eta - \rho\frac{\sigma_C}{\sigma_T}(\Lambda_\theta(z) - x^t\beta)}{\sigma_C\sqrt{1-\rho^2}}\right)\right]$$

$$\times\ \phi\left(\frac{\Lambda_\theta(z) - x^t\beta}{\sigma_T}\right)\Lambda_\theta(z)',$$

and similarly,

$$f_{Z,\delta|X,W}(z,0|x,w;\alpha)=\frac{1}{\sigma_C}\left[1-\Phi\left(\frac{\Lambda_\theta(z)-x^t\beta-\rho\frac{\sigma_T}{\sigma_C}(\Lambda_\theta(z)-w^t\eta)}{\sigma_T\sqrt{1-\rho^2}}\right)\right]$$

$$\times\phi\left(\frac{\Lambda_\theta(z)-w^t\eta}{\sigma_C}\right)\Lambda_\theta(z)'.$$

The authors then provided a theorem and proof of the identifiability of the parameter vector $\alpha=(\theta,\beta,\eta,\sigma_T^2,\sigma_C^2,\rho)$ such that if $f_{Z_1,\delta_1|X,W}(\cdot,\ell|x,w;\alpha_1)=f_{Z_2,\delta_2|X,W}(\cdot,\ell|x,w;\alpha_2)$, then $\alpha_1=\alpha_2$. Naturally, estimation of the parameter vector α now follows, with the likelihood function expressed as

$$L(\alpha)=\prod_{i=1}^{n}f_{Z,\delta|X,W}(Z_i,\delta_i|X_i,W_i;\alpha)$$

$$=\prod_{i=1}^{n}\{\frac{1}{\sigma_T}\left[1-\Phi\left(\frac{\Lambda_\theta(Z_i)-W_i{}^t\eta-\rho\frac{\sigma_C}{\sigma_T}\left(\Lambda_\theta(Z_i)-X_i{}^t\beta\right)}{\sigma_C\sqrt{1-\rho^2}}\right)\right]\phi\left(\frac{\Lambda_\theta(Z_i)-X_i{}^t\beta}{\sigma_T}\right)\}^{\delta_i}$$

$$\times\{\frac{1}{\sigma_C}\left[1-\Phi\left(\frac{\Lambda_\theta(Z_i)-X_i{}^t\beta-\rho\frac{\sigma_T}{\sigma_C}\left(\Lambda_\theta(Z_i)-W_i{}^t\eta\right)}{\sigma_T\sqrt{1-\rho^2}}\right)\right]\phi\left(\frac{\Lambda_\theta(Z_i)-W_i{}^t\eta}{\sigma_C}\right)\}^{1-\delta_i}$$

$$\times\Lambda_\theta(Z_i)'.$$

Standard errors (SEs) and confidence intervals (CIs) were subsequently found by utilizing the asymptotic normality of maximum likelihood estimators (MLEs).

Deresa and Van Keilegom (2020) extended the above paper (Deresa and Van Keilegom, 2019) by additionally including administrative right-censoring times, and considering multivariate (more than two) competing risks. Afterwards, Deresa and Van Keilegom (2021) further generalized the "BVN transformation for identifiability" approach by leaving the transformation function unspecified and simultaneously estimating a non-parametric transformation function from the given survival data subject to dependent censoring.

The most recent work by Czado and Van Keilegom (2023) considerably expands the parametric approach to copulas and marginal distributions widely used in practice. The authors consider a survival time T where T is stochastically dependent on a censoring time C, for which the researcher's interest is in the marginal distribution of T. They provide sufficient conditions where a parametric copula and parametric marginal distributions of T and C completely identifiable without assuming that the copula parameter is known (the previous "assumed copula" approach).

Similar to our notation, consider an event (of interest) time T and dependent censoring time C, where the observables are Y = min(T, C) and $\delta = I(T \leq C)$. Assume parametric marginal distributions and copulas $F_T \in \{F_{T,\theta_T}: \theta_T \in \Theta_T\}$, $F_C \in \{F_{C,\theta_C}: \theta_C \in \Theta_C\}$, $C \in \{C_\theta: \theta \in \Theta\}$. Then, from Sklar's theorem (1959), for continuous marginals F_T and F_C, there exist s a unique copula C such that $F_{T,C}(t,c) = C\{F_T(t), F_C(c)\} = C(u,v), for\ t,c \geq 0$. Also, express the conditional cdfs via copula partial derivatives as

$$\begin{cases} h_{C|T,\theta}(v|u) = \dfrac{\partial}{\partial u} C_\theta(u,v) = \Pr\,(V \le v | U = u) \\ h_{T|C,\theta}(u|v) = \dfrac{\partial}{\partial v} C_\theta(u,v) = \Pr\,(U \le u | V = v) \end{cases},$$

such that the sub-distribution functions (CIFs) are expressed as

$$\begin{cases} F_{Y,\delta}(y,1) = \int_0^y \{1 - h_{C|T}\big(F_C(t)\big|F_T(t)\big)\} f_T(t) dt \\ F_{Y,\delta}(y,0) = \int_0^y \{1 - h_{T|C}\big(F_T(c)\big|F_C(c)\big)\} f_C(c) dc \end{cases},$$

and the corresponding sub-density functions are then

$$\begin{cases} f_{Y,\delta}(y,1) = \{1 - h_{C|T}\big(F_C(y)\big|F_T(y)\big)\} f_T(y) \\ f_{Y,\delta}(y,0) = \{1 - h_{T|C}\big(F_T(y)\big|F_C(y)\big)\} f_C(y) \end{cases}.$$

Again, "identifiability" is defined as the parameters $(\theta, \theta_T, \theta_C) \in \Theta \times \Theta_T \times \Theta_C$ uniquely determining the density function of the observable random variables (Y, δ) such that if $f_{Y,\delta,\alpha_1} \equiv f_{Y,\delta,\alpha_2}$ then $\alpha_1 = \alpha_2,\ \alpha_j = (\theta_j, \theta_{Tj}, \theta_{Cj})^t, j = 1,2.$

The following theorems (*Thm.1 – Thm.4*) regarding the identifiability of specific parametric distributions and copulas are stated without proof, and additional details are deferred to the original paper (Czado & Van Keilegom, 2023).

(Thm.1) Suppose that the following two conditions hold.

(i) For $\theta_{T1}, \theta_{T2} \in \Theta_T$ and $\theta_{C1}, \theta_{C2} \in \Theta_C$, we have the four equivalences

$$\lim_{t\to 0} \frac{f_{T,\theta_{T1}}(t)}{f_{T,\theta_{T2}}(t)} = 1 \Leftrightarrow \theta_{T1} = \theta_{T2},\ \lim_{t\to \infty} \frac{f_{T,\theta_{T1}}(t)}{f_{T,\theta_{T2}}(t)} = 1 \Leftrightarrow \theta_{T1} = \theta_{T2},$$

$\lim_{t\to 0}\frac{f_{C,\theta_{C1}}(t)}{f_{C,\theta_{C2}}(t)}=1\Leftrightarrow\theta_{C1}=\theta_{C2},\ \lim_{t\to\infty}\frac{f_{C,\theta_{C1}}(t)}{f_{C,\theta_{C2}}(t)}=1\Leftrightarrow\theta_{C1}=\theta_{C2}.$

(ii) The parameter space $\Theta\times\Theta_T\times\Theta_C$ is such that

$\lim_{t\to 0}h_{T|C,\theta}(u_t|v_t)=0,\forall(\theta,\theta_T,\theta_C)\in\Theta\times\Theta_T\times\Theta_C$ or

$\lim_{t\to\infty}h_{T|C,\theta}(u_t|v_t)=0,\forall(\theta,\theta_T,\theta_C)\in\Theta\times\Theta_T\times\Theta_C$, and similarly for $h_{C|T,\theta}(v_t|u_t)$,

where $u_t=F_{T,\theta_T}(t)$, and $v_t=F_{C,\theta_C}(t)$.

Then the model $F_{T,C}(t,c)=C\{F_T(t),F_C(c)\}$ is identified.

(Thm.2) Condition (i) of *Thm.1* is satisfied for the families of log-Normal, log-Student-t, Weibull, and log-Logistic densities.

(Thm.3) [Frank, Gumbel, and Gaussian copulas] Condition (ii) of *Thm.1* is satisfied by the following:

(i) the Frank copula, independently of the marginal distributions and the parameter space,

(ii) the Gumbel copula if $\lim_{t\to 0}\frac{\log\{f_{T,\theta_T(t)}\}}{\log\{f_{C,\theta_C(t)}\}}\in(0,\infty),\forall(\theta,\theta_T)\in(\Theta_T\times\Theta_C)$,

(iii) the Gaussian copula if $\lim_{t\to 0}A_{\theta,F_{T,\theta_T},F_{C,\theta_C}}(t)=-\infty,\forall(\theta,\theta_T,\theta_C)\in\Theta\times\Theta_T\times\Theta_C$

or $\lim_{t\to\infty}A_{\theta,F_{T,\theta_T},F_{C,\theta_C}}(t)=-\infty,\forall(\theta,\theta_T,\theta_C)\in\Theta\times\Theta_T\times\Theta_C$,

and similarly for $A_{\theta,F_{C,\theta_C},F_{T,\theta_T}}$, where $A_{\theta,F_1,F_2}(t)=\Phi^{-1}\{F_1(t)\}-\theta\Phi^{-1}\{F_2(t)\}$.

(Thm.4) [Clayton copula] Suppose that condition (i) of *(Thm.1)* holds, $\Theta_T \times \Theta_C$ is such that $\lim_{t\to 0} \frac{f_{T,\theta_T}(t)}{f_{C,\theta_C}(t)}$ is either 0 or $+\infty$ for all $\theta_T \in \Theta_T$ and $\theta_C \in \Theta_C$, and the copula C_θ is a Clayton copula ($\theta > 0$). Then the model $F_{T,C}(t,c) = C_\theta\{F_T(t), F_C(c)\}$ is identified.

For marginal distributions and copulas that are identifiable via *(Thm.1)* to *(Thm.4)*, the parameters are estimated using maximum likelihood, for which the likelihood function is composed of the two probabilities $\Pr(T_i = y_i, C_i > y_i)$ and $\Pr(C_i = y_i, T_i > y_i)$, as the following:

$$L(\alpha) = \prod_{i=1}^{n} [\{1 - h_{C|T}(F_C(y_i)|F_T(y_i))\}f_T(y_i)]^{\delta_i} [\{1 - h_{T|C}(F_T(y_i)|F_C(y_i))\}f_C(y_i)]^{1-\delta_i},$$

The usual log-likelihood function, score vector, and information matrix are utilized for point & interval estimation & inference, and consistency and asymptotic normality of the estimators are obtained from the properties of the MLE.

The main utility of this paper is its claimed novelty in proving the identifiability of a copula model for dependent censoring, where the copula association parameter is not assumed to be known. Although the study covers several survival time distributions and copulas that are often used in practice, Deresa and Van Keilegom (2022) note in another study that some survival time distributions, such as the Gompertz distribution, do not satisfy the above identifiability condition (*Thm.1*), which leaves room for improvement.

2.5 Other Approaches to Modeling Dependent Competing Risks in Survival Analysis

Here, we briefly mention some approaches other than that of copula-based modeling for dependent competing risks or dependently censored survival data. Neither the approaches themselves nor the list of studies within each approach are exhaustive or complete, as they are outside the immediate scope of the current study and were not extensively reviewed.

Heckman and Honore (1989) proposed identifiability conditions for models with regressors, or covariates, showing the possibility of identifying the joint distribution of competing risks survival times without parametric functional form assumptions. This approach essentially relies on a sufficient variation in the marginal survival times by different values of a highly predictive covariate, where the covariate is associated with both or all of the marginal survival times. Abbring and van den Berg (2003) showed that the conditions of Heckman and Honore (1989) can be considerably relaxed in the mixed PH case, such that the marginal survival times require much less variation by the different values of a covariate.

Inverse probability of censoring weights (IPCWs) proposed by Robins and Finkelstein (2000) is another approach of utilizing available covariates that are associated with both the time to event of interest and the time to other competing or dependent censoring events. The main idea is to compensate for dependently censored subjects by giving extra weight

to subjects who are not yet censored and have similar characteristics to those censored in terms of the available covariates. By estimating the probability of censoring at time t = $P_c(t)$ through a Cox model including the available covariates, the IPCW is calculated as = $1/(1 - P_c(t))$ (Willems et al., 2018). For the IPCWs to fully adjust for the dependence, all covariates that might be associated with the event of interest and the dependent censoring event, thus inducing a correlation (dependence) between the two, must be measured. This is also known as the "no unmeasured confounders" assumption.

A third approach to modeling competing risks survival data, especially for semi-competing risks such as the association between progression-free survival and overall survival in oncology trials, is the multi-state model or illness-death model (Meller et al., 2019; Weber & Titman, 2018; Putter et al., 2007). The multi-state model does not assume latent failure times, but rather uses transition intensity matrices to model the probability of transitioning between possible states or arriving at a particular state. This allows for clinical prediction modeling of disease incidence or patient survival, especially in semi-competing risks survival data where both the time-to event of interest and the time to some other intermediate or non-terminal event are observed in at least some cases. However, this approach is inherently unable to address the potential correlation (dependence) in the "classical" competing risks situation of mutually exclusive time-to-events, as the latent failure times framework is not utilized and the counterfactual transition or association between the mutually exclusive states is undefined.

Chapter 3

Proposed Method

3.1 Previous studies on the parametric identifiability of bivariate Normally distributed competing risks data

Among the previous works on identifiability within parametric families of bivariate distributions, we focus on the ubiquitous bivariate Normal distribution (the BVN distribution), where the distribution of the observable X = min(T, C) and δ = I(T ≤ C) uniquely determines the distribution of the (unobservable) underlying BVN distribution (Nadas, 1971; Basu & Ghosh, 1978).

Using the notations of Nadas (1971), consider a BVN-distributed random vector (X_0, X_1) with mean (μ_0, μ_1) and covariance matrix (σ_{ij}) where $\sigma_{ii} = \sigma_i^2$ and $\sigma_{ij} = \rho\sigma_i\sigma_j$ (i, j = 0, 1 and i≠j). Here, Z = min(X_0, X_1), and I is defined by Z = X_I. Also, let $n(\cdot \mid a, b^2)$ be the univariate Normal density with mean a and standard deviation b, and let $N(\cdot \mid a, b^2)$ be the corresponding cumulative distribution. Then, the conditional density $f_i(z)$ of Z for I = i is given by the following lemma (proof deferred to the original paper).

(Lemma) $f_i(z) = \begin{cases} p_i^{-1} \cdot n(z \mid \mu_i, {\sigma_i}^2) \cdot \left(1 - N\left(\frac{z-\mu_i^*}{\sigma_i^*}\middle|0,1\right)\right), if\ \rho\sigma_{1-i} \neq \sigma_i \\ n(z \mid \mu_i, {\sigma_i}^2),\ otherwise \end{cases}$, where

$\mu_i^* = \alpha_i\mu_{1-i} + (1-\alpha_i)\mu_i,\ \sigma_i^* = \alpha_i\sigma_{1-i}(1-\rho^2)^{1/2},\ \alpha_i = (1-\rho\frac{\sigma_{1-i}}{\sigma_i})^{-1},$

$and\ p_i = \Pr(I=i) = N(0 \mid \mu_i - \mu_{1-i}, \sigma_0^2 + \sigma_1^2 - 2\rho\sigma_0\sigma_1).$

From the lemma, $\frac{p_i \cdot f_i(z)}{n(z|\,\mu_i,{\sigma_i}^2)} = 1 - N\left(\frac{z-\mu_i^*}{\sigma_i^*}\middle|0,1\right)$, and now, let μ_0', μ_1', and σ_{ij}' be the parameters defining any other BVN density for which its (Z, I) has the same conditional density $f_i(z)$ of Z for I = i.

Since $\lim_{z\to-\infty}\frac{p_i \cdot f_i(z)}{n(z|\,\mu_i,{\sigma_i}^2)} = \lim_{z\to-\infty}\left\{1 - N\left(\frac{z-\mu_i^*}{\sigma_i^*}\middle|0,1\right)\right\} = 1$ such that $\lim_{z\to-\infty}\frac{p_i \cdot f_i(z)}{n(z|\,\mu_i,{\sigma_i}^2)} = 1,$

$\lim_{z\to-\infty}\frac{p_i \cdot f_i(z)}{n(z|\,\mu_i,{\sigma_i}^2)} = \lim_{z\to-\infty}\frac{p_i \cdot f_i(z)}{n(z|\,{\mu'}_i,{\sigma'}_i^2)} = 1$. Therefore, $\lim_{z\to-\infty}\frac{n(z|\,\mu_i,{\sigma_i}^2)}{n(z|\,{\mu'}_i,{\sigma'}_i^2)} = 1$, from which $\mu_i = \mu_i', \sigma_i = \sigma_i'$ is deduced, and $\rho_i = \rho_i'$ also, from $p_i = p_i'$.

Basu and Ghosh (1978) additionally showed in a lemma that if $\sigma_i < \sigma_i'$ or $\sigma_i = \sigma_i'$ but $\mu_i > \mu_i'$ then the above $\lim_{z\to-\infty}\frac{n(z|\,\mu_i,{\sigma_i}^2)}{n(z|\,{\mu'}_i,{\sigma'}_i^2)} = 0$, and if $\sigma_i > \sigma_i'$ or $\sigma_i = \sigma_i'$ but $\mu_i < \mu_i'$ then $\lim_{z\to-\infty}\frac{n(z|\,\mu_i,{\sigma_i}^2)}{n(z|\,{\mu'}_i,{\sigma'}_i^2)} = \infty$, and that these results equally apply when $z \to +\infty$. They also dealt with the case where $\alpha_i = (1-\rho\frac{\sigma_{1-i}}{\sigma_i})^{-1}$ may not be positive (which Nadas implicitly assumed), and proved BVN's identifiability in this situation as well.

3.2 Proposed method of estimating the correlation in various parametric bivariate competing risks data

We now present our proposed method of a unified parametric approach to estimating the correlation (or the strength of dependence) in bivariate competing risks survival data, where the minimum of either T (time-to-event of interest) or C (time-to competing event or dependent censoring) is observed, but never both. The motivation for such an approach to correlation (dependence) estimation is that no practical methodology exists for simultaneously estimating the correlation and marginal hazards (Emura & Chen, 2016), and the usual ML-based estimation of the correlation parameter together with the marginal hazards parameters results in a large variability of the estimates due to the likelihood function having little information for the correlation (Chen, 2010; Michimae & Emura, 2022). Therefore, we propose a method to first estimate the correlation in some given bivariate competing risks data, and afterwards, estimate the marginal survival and/or hazard function of the time-to-event of interest, taking the estimated correlation into account.

Our proposal starts with a possible connection between some given bivariate competing risks data and the previously proven to be identifiable (Section 4.1) bivariate normal (BVN) or bivariate Weibull distributions, via the bivariate central limit theorem (CLT). First, the multivariate CLT is stated as below.

(Theorem) Multivariate central limit theorem (CLT)

For some k-dimensional random vector $\begin{bmatrix} X_{(1)} \\ X_{(2)} \\ \cdots \\ X_{(k)} \end{bmatrix}$, consider n i.i.d. samples $\begin{bmatrix} X_{1(1)} \\ X_{1(2)} \\ \cdots \\ X_{1(k)} \end{bmatrix}$, …,

$\begin{bmatrix} X_{n(1)} \\ X_{n(2)} \\ \cdots \\ X_{n(k)} \end{bmatrix}$ with mean vector $\mu_{k\times 1}$ and covariance matrix $\Sigma_{k\times k}$, with k finite (<∞) variances on the diagonal. The sample mean $\bar{X}_n = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n} X_{i(1)} \\ \sum_{i=1}^{n} X_{i(2)} \\ \cdots \\ \sum_{i=1}^{n} X_{i(k)} \end{bmatrix}$ then converges in distribution to a k-variate Normal distribution, i.e. $\sqrt{n}\cdot(\bar{X}_n - \mu_{k\times 1}) \xrightarrow{d} N_k(0_{k\times 1}, \Sigma_{k\times k})$.

In the bivariate case of $\begin{bmatrix} Y_{(1)} \\ Y_{(2)} \end{bmatrix}$ with mean $\begin{bmatrix} \mu_{(1)} \\ \mu_{(2)} \end{bmatrix}$ and covariance $\begin{bmatrix} \sigma_{(1)}^2 & \rho\sigma_{(1)}\sigma_{(2)} \\ \rho\sigma_{(1)}\sigma_{(2)} & \sigma_{(2)}^2 \end{bmatrix}$,

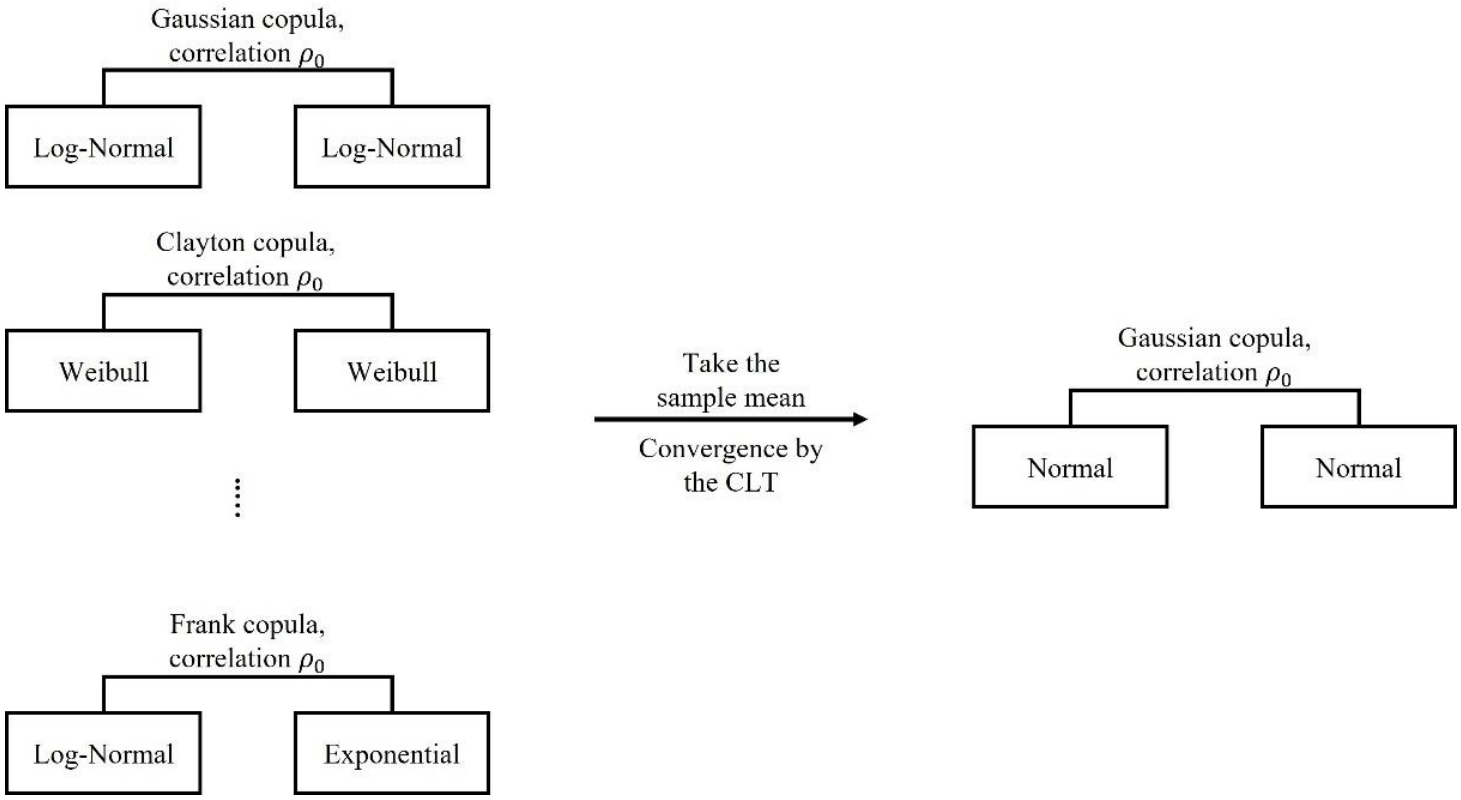
$$\sqrt{n}\cdot\left(\begin{bmatrix} \bar{Y}_{n(1)} \\ \bar{Y}_{n(2)} \end{bmatrix} - \begin{bmatrix} \mu_{(1)} \\ \mu_{(2)} \end{bmatrix}\right) \xrightarrow{d} N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{(1)}^2 & \rho\sigma_{(1)}\sigma_{(2)} \\ \rho\sigma_{(1)}\sigma_{(2)} & \sigma_{(2)}^2 \end{bmatrix}\right),$$

$$\sqrt{n}\cdot(\bar{Y}_{n(1)} - \mu_{(1)}) \xrightarrow{d} N(0, \sigma_{(1)}^2) \quad , \quad \sqrt{n}\cdot(\bar{Y}_{n(2)} - \mu_{(2)}) \xrightarrow{d} N(0, \sigma_{(2)}^2) \quad , \quad -1 \le \rho \le 1$$

(Hogg, McKean, & Craig, 2013).

Now, consider taking the sample mean of some parametric bivariate dependent censoring (or competing risks) data correlated through some parametric copula with correlation ρ_0. Regardless of what the marginal distributions are or what kind of functional form the copula has, taking its sample mean would result in a convergence to a BVN distribution with the same correlation ρ_0 (or two univariate Normal distributions linked by a Gaussian copula with correlation ρ_0), by the bivariate CLT. This situation is depicted in the following diagram (Figure 3.1).

Figure 3.1 Convergence of the sample mean to a bivariate normal distribution in various parametric copula models with parametric marginal distributions



Thinking in terms of this unique Gaussian (or Normal) copula with Normally distributed marginals for the sample means of a given bivariate competing risks data, the correlation ρ_0 is preserved, regardless of the original data's copula functional form or marginal distributions. Also, we described in Section 4.1 of how a BVN-distributed (Y_1, Y_2) is free from the non-identifiability issue such that the sub-density function of its observables of min(Y_1, Y_2) and I($Y_1 \leq Y_2$) uniquely identify the underlying BVN distribution parameters. Using this fact to find ρ_0, the following questions need to addressed:

(i) Does a BVN distribution that produces the same sample mean information as those of (X = min(T, C), δ = $I(T \leq C)$), and thus with its correlation parameter being equal to ρ_0 of (T, C), uniquely exist?

(ii) Is the desired BVN distribution in (i) estimable from the sample mean information of (X = min(T, C), δ = $I(T \leq C)$)?

These questions are addressed via a couple of conjectures below.

First, the setting and notations of our conjectures are:

(a) Some given bivariate competing risks data (T, C) of sample size n with parametric marginal distributions for a time-to-event of interest T and time-to competing event or dependent censoring C, which have a correlation ρ_0 through a parametric copula,

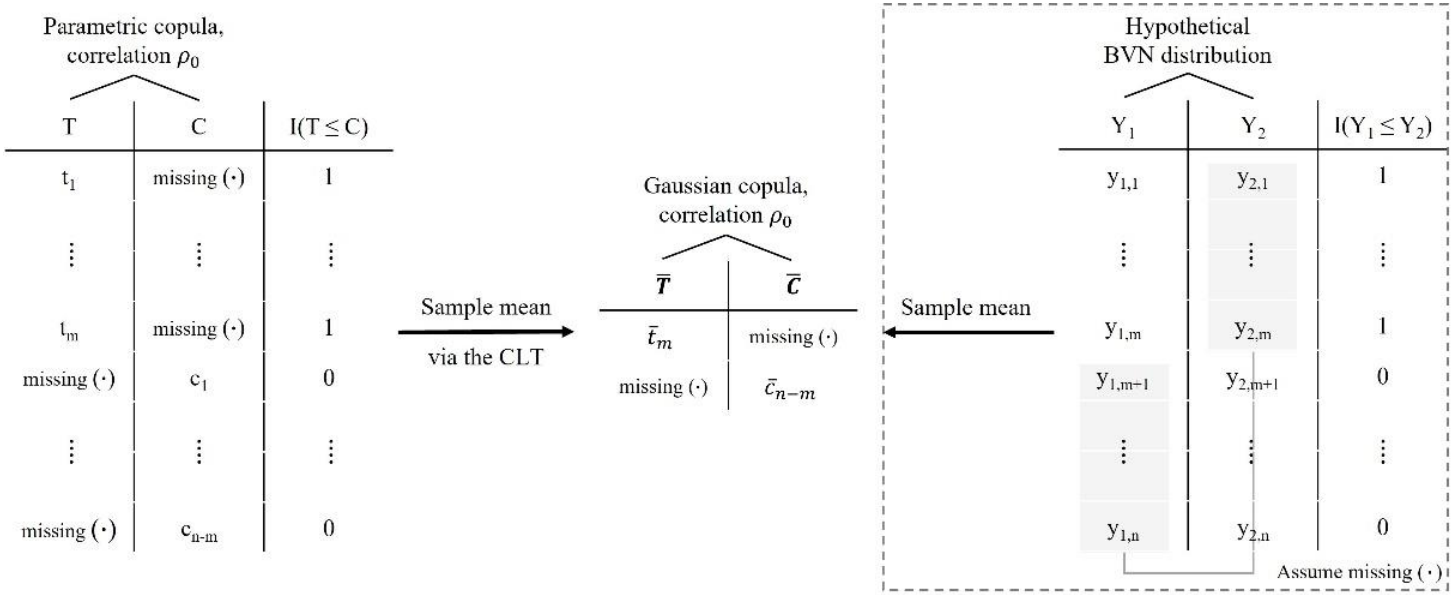
(b) Sample mean data of ($\bar{T}$, $\bar{C}$), with $\bar{T} = \bar{t}_m$ (with mean $\mu_{\bar{t}_m}$, variance $\sigma^2{}_{\bar{t}_m}$) or

missing (·), $\bar{C} = \bar{c}_{n-m}$ (with mean $\mu_{\bar{c}_{n-m}}$, variance $\sigma^2{}_{\bar{c}_{n-m}}$) or missing (·), where $m = \sum_{i=1}^{n} I(T \le C)$, $\bar{T} = \frac{1}{m}\sum_{i=1}^{n} T_i \cdot I(T \le C)$, $\bar{C} = \frac{1}{(n-m)}\sum_{i=1}^{n} C_i \cdot \{1 - I(T \le C)\}$, and $(\bar{T}, \bar{C})$ have the same correlation ρ_0 as in (T, C) through a Gaussian (or Normal) copula, via the bivariate CLT,

(c) A hypothetical BVN-distributed (Y_1, Y_2) of sample size n with parameters $\theta = (\mu_{Y_1}, \mu_{Y_2}, \sigma^2_{Y_1}, \sigma^2_{Y_2}, \rho)$, for which min($Y_1$, Y_2) and I($Y_1 \le Y_2$) produce the same sample mean information as in (ii), i.e. $\bar{Y}_1 = \bar{y}_{1,m} = \bar{t}_m$ (with mean $\mu_{\bar{t}_m}$, variance $\sigma^2{}_{\bar{t}_m}$) or missing (·), $\bar{Y}_2 = \bar{y}_{2,n-m} = \bar{c}_{n-m}$ (with mean $\mu_{\bar{c}_{n-m}}$, variance $\sigma^2{}_{\bar{c}_{n-m}}$) or missing (·), where $m = \sum_{i=1}^{n} I(Y_1 \le Y_2)$, $\bar{Y}_1 = \frac{1}{m}\sum_{i=1}^{n} Y_{1,i} \cdot I(Y_1 \le Y_2)$, $\bar{Y}_2 = \frac{1}{(n-m)}\sum_{i=1}^{n} Y_{2,i} \cdot \{1 - I(Y_1 \le Y_2)\}$.

This setting is also depicted by the following diagram (Figure 3.2).

Figure 3.2 A hypothetical bivariate normal distribution that has the same sample mean information as that of a given bivariate competing risks data with a parametric copula linking the parametric marginal distributions



Conjectures 1~2 now follow as:

(Conjecture 1) For the sample mean information from the given bivariate competing risks data (T, C), a BVN distribution that produces the same sample mean information, and thus has the same correlation ρ_0 as that of the given (T, C) as its correlation parameter, uniquely exists.

(Justification 1) (i) Correlation parameter of the hypothesized BVN distribution equals to ρ_0 of the given bivariate competing risks data (T, C):

For now, suppose that such a BVN distribution uniquely exists such that its correlation parameter to be estimated also uniquely exists. Justification of existence and uniqueness follow below. Then for large enough sample size n, the bivariate CLT guarantees that the sample mean's distribution of a given bivariate competing risks data has the same covariance matrix Σ, and hence, the same correlation ρ_0 as that of (T, C).

∴ By the bivariate CLT in reverse direction, if a BVN distribution's min(Y_1, Y_2) and I($Y_1 \leq Y_2$) data produces the same sample mean information as that of X = min(T, C) and δ = $I(T \leq C)$, then the BVN distribution's correlation parameter ρ is also equal to that of the sample mean information, which is ρ_0.

(ii) Existence: The sample mean information from bivariate competing risks data are such that $\mu_{\bar{t}_m}$, $\mu_{\bar{c}_{n-m}}$ are both > 0, ${\sigma^2}_{\bar{t}_m}$, ${\sigma^2}_{\bar{c}_{n-m}}$ are both finite (< ∞), 0 < m < n, and correlation -1 ≤ ρ_0 ≤ 1. For the BVN-distributed (Y_1, Y_2) parameters $\theta = (\mu_{Y_1}, \mu_{Y_2}, \sigma^2_{Y_1}, \sigma^2_{Y_2}, \rho)$, the distribution's support has a range of (-∞, ∞) for each of (Y_1, Y_2),

each of $\sigma_{Y_1}^2, \sigma_{Y_2}^2$ can take any finite value, and the correlation ρ covers the whole interval [-1, 1]. Especially, if the missing n – m observations of T and m observations of C were counterfactually known, then the sample mean random vector ($\bar{T}$, $\bar{C}$) would be distributed as $\sim N_2(\begin{bmatrix}\bar{t}_n \\ \bar{c}_n\end{bmatrix}, \begin{bmatrix}\sigma_{\bar{t}_n}^2 & \rho_0\sigma_{\bar{t}_n}\sigma_{\bar{c}_n} \\ \rho_0\sigma_{\bar{t}_n}\sigma_{\bar{c}_n} & \sigma_{\bar{c}_n}^2\end{bmatrix})$ by the bivariate CLT, for the total number of samples n without missing or unknown values. Apparently then, a BVN distribution (Y_1, Y_2) for which its sample mean ($\bar{Y}_1$, $\bar{Y}_2$) is identical to ($\bar{T}$, $\bar{C}$) would be distributed as $\sim N_2(\begin{bmatrix}\bar{t}_n \\ \bar{c}_n\end{bmatrix}, n \cdot \begin{bmatrix}\sigma_{\bar{t}_n}^2 & \rho_0\sigma_{\bar{t}_n}\sigma_{\bar{c}_n} \\ \rho_0\sigma_{\bar{t}_n}\sigma_{\bar{c}_n} & \sigma_{\bar{c}_n}^2\end{bmatrix})$ by the bivariate CLT in reverse direction, i.e. such a BVN distribution exists.

∴ There exists some BVN distribution (Y_1, Y_2) that produces the same sample mean information as that of the given bivariate competing risks data (T, C).

(iii) Uniqueness: Suppose that two different BVN distributions, using only their respective min(Y_1, Y_2) and I($Y_1 \leq Y_2$) data, produce the same sample mean information as that of the given bivariate competing risks data. Since we know that a given bivariate dataset's sample mean is unique (a given dataset cannot produce two different sample means), the above assumption of two different BVN distributions having the same sample mean information implies that the two different distributions produced the same min(Y_1, Y_2) and I($Y_1 \leq Y_2$) data, i.e. the same competing risks format data, which would then result in the same sample mean information. However, by Nadas (1971) and Basu and Ghosh (1978), the identified minimum of a BVN-distributed pair uniquely determines the distribution of the pair (their

underlying BVN distribution), so the assumption that two different BVN distributions produced the same competing risks format data cannot be true.

∴ A given sample mean information of ($\bar{T}$, $\bar{C}$) is produced from a unique BVN distribution (Y_1, Y_2) and its competing risks format data of min(Y_1, Y_2) and I($Y_1 \leq Y_2$).

(Conjecture 2) Such a BVN distribution described in Conjecture 1, especially its correlation parameter ρ, can be numerically estimated with the given bivariate competing risks data's sample mean information of $\frac{m}{n}$, $\mu_{\bar{t}_m}$, $\mu_{\bar{c}_{n-m}}$, $\sigma^2{}_{\bar{t}_m}$, and $\sigma^2{}_{\bar{c}_{n-m}}$.

(Justification 2) From the given bivariate competing risks data's sample mean, we know the proportions $\frac{m}{n}$ or $\frac{(n-m)}{n}$ of T or C that are observed as the minimum of the two, $\mu_{\bar{t}_m}$ and $\sigma^2{}_{\bar{t}_m}$ of the observed T = min(T, C), and $\mu_{\bar{c}_{n-m}}$ and $\sigma^2{}_{\bar{c}_{n-m}}$ of the observed C = min(T, C). Using this sample mean information, the parameters $\theta = (\mu_{Y_1}, \mu_{Y_2}, \sigma^2_{Y_1}, \sigma^2_{Y_2}, \rho)$ of the BVN distribution (Y_1, Y_2) can be estimated via a method-of-moments type estimation as below.

(i) The proportion of T observed, $\frac{m}{n}$: For BVN random variables S and T, use the fact that Pr(S < T) = Pr(S - T < 0), where S – T ~ $N(\mu_S - \mu_T, \sigma_S^2 + \sigma_T^2 - 2\rho\sigma_S\sigma_T)$, Pr(S - T < 0) = Pr(Z < $-\frac{\mu_S - \mu_T}{\sqrt{\sigma_S^2+\sigma_T^2-2\rho\sigma_S\sigma_T}}$), Z ~ N(0, 1). Then, compare $\frac{m}{n}$ of ($\bar{T}$, $\bar{C}$) with Pr(Z < $-\frac{\mu_{Y_1} - \mu_{Y_2}}{\sqrt{\sigma^2_{Y_1}+\sigma^2_{Y_2}-2\rho\sigma_{Y_1}\sigma_{Y_2}}}$) of ($Y_1$, Y_2).

(ii) Means of $\bar{T}$ and $\bar{C}$, $\mu_{\bar{t}_m}$ and $\mu_{\bar{c}_{n-m}}$: For the observed $Y_1 = \min(Y_1, Y_2)$, obtain its sub-density function as

$$\frac{1}{\sigma_{Y_1}}\left[1-\Phi\left(\frac{y_1-\mu_{Y_2}-\rho\frac{\sigma_{Y_2}}{\sigma_{Y_1}}(y_1-\mu_{Y_1})}{\sigma_{Y_2}\sqrt{1-\rho^2}}\right)\right]$$

$$\times\phi\left(\frac{y_1-\mu_{Y_1}}{\sigma_{Y_1}}\right),$$

where Φ is the standard Normal cdf, ϕ the standard Normal pdf, and for the jth such Y_1 (= y_1), calculate its contributing weight as $wt_j = sub-density_j / \sum_{j=1}^{m} sub-density_j$ to estimate $\hat{E}[Y_1] = \sum_{j=1}^{m}(y_{1,j}\cdot wt_j)$.

Similarly, For the observed $Y_2 = \min(Y_1, Y_2)$, obtain its sub-density function, calculate its contributing weight as wt_k, and estimate $\hat{E}[Y_2] = \sum_{k=1}^{n-m}(y_{2,k}\cdot wt_k)$. Then, compare $\mu_{\bar{t}_m}$ of $(\bar{T}, \bar{C})$ with $\hat{E}[Y_1] = \sum_{j=1}^{m}(y_{1,j}\cdot wt_j)$ of (Y_1, Y_2), and compare $\mu_{\bar{c}_{n-m}}$ of $(\bar{T}, \bar{C})$ with $\hat{E}[Y_2] = \sum_{k=1}^{n-m}(y_{2,k}\cdot wt_k)$ of (Y_1, Y_2).

(iii) Variances of $\bar{T}$ and $\bar{C}$, $\sigma^2{}_{\bar{t}_m}$ and $\sigma^2{}_{\bar{c}_{n-m}}$: For the observed $Y_1 = \min(Y_1, Y_2)$, use the sub-density and corresponding weight in (ii) to estimate $\hat{E}[Y_1^2] = \sum_{j=1}^{m}(y_{1,j}^2\cdot wt_j)$, and use the estimated $\hat{E}[Y_1]$ in (ii) to estimate $\widehat{Var}[Y_1] = \hat{E}[Y_1^2] - \{\hat{E}[Y_1]\}^2$. Similarly, For the observed $Y_2 = \min(Y_1, Y_2)$, use the sub-density, corresponding weight, and $\hat{E}[Y_2]$ in (ii) to estimate $\widehat{Var}[Y_2] = \hat{E}[Y_2^2] - \{\hat{E}[Y_2]\}^2$. Then, compare $\sigma^2{}_{\bar{t}_m}$ of $(\bar{T}, \bar{C})$ with $\widehat{Var}[Y_1]/m$ of (Y_1, Y_2), and compare $\sigma^2{}_{\bar{c}_{n-m}}$ of $(\bar{T}, \bar{C})$ with $\widehat{Var}[Y_2]/(n-m)$ of $(Y_1,$

Y_2).

∴ A numerical estimation algorithm may be conducted as follows:

(a) Starting from some initial parameters $\theta^{(0)} = (\mu_{Y_1}{}^{(0)}, \mu_{Y_2}{}^{(0)}, \sigma_{Y_1}^{2\,(0)}, \sigma_{Y_2}^{2\,(0)}, \rho^{(0)})$, generate a random BVN-distributed sample of size n from the initial parameters.

(b) Perform the comparisons (i)~(iii) above for the error calculation of an objective function, where the objective is to minimize the aggregated error. Update the BVN distribution parameters to those currently evaluated if the objective function value becomes smaller.

(c) Repeat (b) above for a given number of iterations, or until some convergence criteria is met. The resulting BVN distribution parameters $\theta^{(k)} = (\mu_{Y_1}{}^{(k)}, \mu_{Y_2}{}^{(k)}, \sigma_{Y_1}^{2\,(k)}, \sigma_{Y_2}^{2\,(k)}, \rho^{(k)})$, after k iterations, are the BVN parameters that produce the same sample mean information as that of the given bivariate competing risks data, and specifically, $\rho^{(k)} = \rho_0$, the correlation parameter of interest.

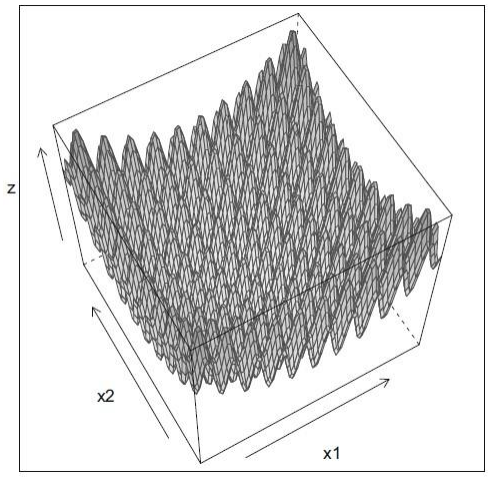
Conclusively, the correlation (dependence) in bivariate competing risks data for any parametric marginal distributions linked through parametric copulas can be estimated by our unified approach of estimating the identifiable BVN distribution's correlation parameter, as stated in Conjectures 1~2. Since the bivariate Weibull distribution, which is often used in survival and reliability analyses, was also shown to be identifiable from its observed minimum (Moeschberger & Klein, 1995; Moeschberger, 1974), one can use the bivariate Weibull in place of the ubiquitous BVN distribution to proceed in a similar manner.

3.3 Optimization procedures in estimating the correlation in bivariate competing risks data

We now focus on the numerical estimation algorithm in Conjecture 2 of Section 3.2. More specifically, we present the objective function of minimization and its components, such as the minimization criteria, weights assigned within the objective function, and sample size considerations. In addition, some optimization algorithms for global and local searches, as well as the R packages used for their implementation, are introduced. A parallel computing procedure to simultaneously process multiple runs of newly generated data and their bootstrap samples is also briefly mentioned.

The basic structure of an optimization problem is to find the global minimum or maximum of a pre-defined objective function f(s) by evaluating f with changing s $\in S$, where S is the possible search space. Without loss of generality, we hereafter assume a minimization problem. Finding a unique solution to the optimization problem may become complicated due to a brute-force blind search being unrealistic, constraints on the search space or no constraints resulting in too broad a search space, the difficulty of pinpointing a global minimum among possibly numerous local minima, etc. An example figure of numerous local minima is shown below (courtesy to Cortez, 2021).

Figure 3.3 Plot of the 'rastrigin' objective function with many local minima



The above figure plots the 'rastrigin' function, $f_{rastrigin}(\vec{x}) = \sum_{i=1}^{D}(x_i^2 - 10\cos 2\pi x_i + 10)$, D: number of dimensions, which is a popular benchmark for real-valued optimization algorithm evaluations. Numerous local minima are depicted by the many valleys in the figure, where the overall global minimum lies on the origin. In this situation, relying only on local search methods, such as the Newton-Raphson variants based on gradient descent, may result in failure to find the desired global minimum, due to convergence to a local solution. Among the broad classes of optimization methods: 1. Blind search (grid search, Monte Carlo search etc.), 2. Single-state or 'Local' search (gradient descent, tabu search etc.), 3. Population-based or 'Global' search (simulated annealing, genetic algorithms etc.) (Cortez, 2021), we will focus on a mixed use of global and local search methods.

Recalling Conjecture 2 of Section 3.2, $\Pr(Z < -\frac{\mu_{Y_1}-\mu_{Y_2}}{\sqrt{\sigma_{Y_1}^2+\sigma_{Y_2}^2-2\rho\sigma_{Y_1}\sigma_{Y_2}}})$, $Z \sim N(0,1)$, of

the BVN distribution (Y_1, Y_2) was used to compare against $\frac{m}{n}$ of ($\bar{T}$, $\bar{C}$), which can be thought of as regressing $\frac{m}{n}$ of ($\bar{T}$, $\bar{C}$) upon a function f of the BVN parameters $\mu_{Y_1}, \mu_{Y_2}, \sigma_{Y_1}^2, \sigma_{Y_2}^2, \rho$, such as $\frac{m}{n} = f_1(\mu_{Y_1}, \mu_{Y_2}, \sigma_{Y_1}^2, \sigma_{Y_2}^2, \rho) + \varepsilon_1,\ \varepsilon_1: error\ term\ 1.$

Similarly, the other four possible comparisons of $\mu_{\bar{t}_m}$ with $\hat{E}[Y_1] = \sum_{j=1}^{m}(y_{1,j} \cdot wt_j)$, $\mu_{\bar{c}_{n-m}}$ with $\hat{E}[Y_2] = \sum_{k=1}^{n-m}(y_{2,k} \cdot wt_k)$,

$\sigma^2{}_{\bar{t}_m}$ with $\widehat{Var}[Y_1]/m = \left[\hat{E}[Y_1^2] - \{\hat{E}[Y_1]\}^2\right]/m$,

and $\sigma^2{}_{\bar{c}_{n-m}}$ with $\widehat{Var}[Y_2]/(n-m) = \left[\hat{E}[Y_2^2] - \{\hat{E}[Y_2]\}^2\right]/(n-m)$

can be thought of as

$\mu_{\bar{t}_m} = f_2(\mu_{Y_1}, \mu_{Y_2}, \sigma_{Y_1}^2, \sigma_{Y_2}^2, \rho) + \varepsilon_2,\ \varepsilon_2: error\ term\ 2,$

$\mu_{\bar{c}_{n-m}} = f_3(\mu_{Y_1}, \mu_{Y_2}, \sigma_{Y_1}^2, \sigma_{Y_2}^2, \rho) + \varepsilon_3,\ \varepsilon_3: error\ term\ 3,$

$\sigma^2{}_{\bar{t}_m} = f_4(\mu_{Y_1}, \mu_{Y_2}, \sigma_{Y_1}^2, \sigma_{Y_2}^2, \rho) + \varepsilon_4,\ \varepsilon_4: error\ term\ 4,$

and $\sigma^2{}_{\bar{c}_{n-m}} = f_5(\mu_{Y_1}, \mu_{Y_2}, \sigma_{Y_1}^2, \sigma_{Y_2}^2, \rho) + \varepsilon_5,\ \varepsilon_5: error\ term\ 5.$

Hence, our objective function $f_{obj}(\cdot)$ can be defined as some aggregate function of the five error terms to be minimized, such as

$f_{obj}(\cdot) = \min\ (\sum_{i=1}^{5} {\varepsilon_i}^2)$, or $f_{obj}(\cdot) = \min\ (\sum_{i=1}^{5} |\varepsilon_i|)$,

such that $\hat{\theta} = (\hat{\mu}_{Y_1}, \hat{\mu}_{Y_2}, \hat{\sigma}_{Y_1}^2, \hat{\sigma}_{Y_2}^2, \hat{\rho}) = argmin_{\theta \in \Theta}(\sum_{i=1}^{5} {\varepsilon_i}^2)\ or\ argmin_{\theta \in \Theta}(\sum_{i=1}^{5} |\varepsilon_i|)$,

for the BVN parameter vector $\theta = (\mu_{Y_1}, \mu_{Y_2}, \sigma^2_{Y_1}, \sigma^2_{Y_2}, \rho)$ and its parameter space Θ. Among the often-used minimization criteria (or performance metrics) of mean absolute error (MAE) = $\frac{1}{5}\sum_{i=1}^{5}|\varepsilon_i|$, mean absolute percentage error (MAPE) = $\frac{100}{5}\sum_{i=1}^{5}\frac{|\varepsilon_i|}{true_i}$, where $true_i$ is one of the five sample mean information components $\frac{m}{n}$, $\mu_{\bar{t}_m}$, $\mu_{\bar{c}_{n-m}}$, $\sigma^2{}_{\bar{t}_m}$, $\sigma^2{}_{\bar{c}_{n-m}}$, depending on which comparison is being made, or root mean squared error (RMSE) = $\sqrt{\frac{1}{5}\sum_{i=1}^{5}\varepsilon_i{}^2}$, we considered either the MAPE or RMSE criterion, and chose to use MAPE based on the rationale that MAPE more closely resembles our objective of finding BVN parameters that exactly reproduce the sample mean information of ($\bar{T}$, $\bar{C}$).

In terms of the relative weights assigned within the objective function, we experimented with different weights wt_i, $\sum_{i=1}^{5} w_i = 1$, for the error terms of ε_1 to ε_5. Weights such as being proportional to the I(T ≤ C) proportion, or putting more emphasis on reducing a certain error term such as $\varepsilon_1 = |\frac{m}{n} - f_1(\mu_{Y_1}, \mu_{Y_2}, \sigma^2_{Y_1}, \sigma^2_{Y_2}, \rho)|$ were applied and compared against. Consequently, equal weights of 1/5 each was chosen to provide the relatively best performance (the smallest objective function value).

The sample size n of the given bivariate competing risks data and the hypothesized BVN distribution is another important consideration, as it directly determines the amount of information available when searching for the BVN parameters that minimize our objective function. This is especially true in our case, since the proposed algorithm requires generating a random BVN sample with sample size n from a given set of interim parameters

for each comparison with the 'true' sample mean information of ($\bar{T}$, $\bar{C}$). Thus, it may be hypothesized that a larger sample size n would provide more information for both sides of distributions (T, C) and (Y_1, Y_2) when comparing their competing risks format sample mean information for the BVN parameter updates on (Y_1, Y_2).

We used a combination of first globally searching the parameter space for promising candidates, and then refining the possible solutions with a local search, after which the candidate providing the lowest objective function value is chosen as the final solution. For the global (or population-based) search, stochastic Monte Carlo, simulated annealing, and genetic algorithms were considered (Cortez, 2021; Givens & Hoeting, 2005). For the subsequent local (or single-state) search, a Newton-Raphson variant of gradient descent by the Broyden-Fletcher-Goldfarb-Shanno (BFGS) formula was used (Nash, 2014). We first note that the stochastic Monte Carlo search was discarded due to its pure randomness and inefficiency in searching for a solution, and the genetic algorithm as well, because the so-called crossovers and mutations frequently resulted in incalculable numbers for generating a BVN distribution, such as the covariance matrix not being positive definite.

Therefore, we adopted simulated annealing (Kirkpatrick et al., 1983) as the first stage global search, which also employs gradient descent to search for solutions but with additional stochastic randomness. The 'annealing' phenomenon is observed in metallurgy, which is the process of heating up a solid and then cooling it slowly (Cortez, 2021; Givens & Hoeting, 2005). By adopting a temperature parameter, say "temp", which governs the probability of accepting an inferior option as the updated solution, simulated annealing is

able to escape from possible local minima and move toward the desired global minimum. As this stochastic acceptance of inferior options is key to the algorithm, the temp parameter should start with a high value and then gradually decrease as the algorithm proceeds. The Pr(inferior option acceptance) is set by Boltzmann's probability $\exp\{-\Delta E/(k \cdot temp)\}$ in thermodynamics, where ΔE is the internal energy increase and k is Boltzmann's constant. Intuitively, $\Delta E = f_{obj}(\theta^{new}) - f_{obj}(\theta^{curr})$ is how much worse the new inferior option is compared to the current solution, and temp is the quantitative measure of willingness to accept such an inferior option as the updated solution. As the algorithm progresses and temp decreases to zero, the Pr(inferior option acceptance) also goes to zero, thus hopefully honing in on the global minimum. The overall process is summarized in the following algorithm by Givens and Hoeting (2005).

(Algorithm) Simulated annealing

Conduct an iterative procedure with an initial solution vector $\theta^{(0)}$ and initial temperature $temp_0$. Iterations within each stage of constant temperature are indexed by i. The stages are indexed by j = 0, 1, 2 … and each stage consists of m_j number of iterations, that is, i = 1, 2, …, m_j for stage j.

(i) Select a candidate solution θ^* within the neighborhood of $\theta^{(i)}$, say $\mathcal{N}(\theta^{(i)})$ according to a proposal density $g^{(i)}(\cdot|\theta^{(i)})$. The Gaussian or Cauchy-Lorentz visiting distributions are frequently used proposal densities (Xiang et al., 2013).

(ii) Randomly decide whether to adopt θ^* as the next candidate solution or to keep

the current solution $\theta^{(i)}$. That is, let $\theta^{(i+1)} = \theta^*$ with probability

p = min(1, exp $\{-[f_{obj}(\theta^*) - f_{obj}(\theta^{(i)})]/temp_j\}$.

Otherwise, keep the current solution as $\theta^{(i+1)} = \theta^{(i)}$.

(iii) Repeat steps (i) and (ii) m_j number of times.

(iv) Increment j. Update $temp_{j+1} = \alpha(temp_j)$ and $m_{j+1} = \beta(m_j)$ by some functions $\alpha(\cdot)$ and $\beta(\cdot)$. $\alpha(\cdot)$ should slowly decrease the temperature to zero, and $\beta(\cdot)$ should scale the number of iterations m_j within each temperature exponentially in probability p.

(v) Return to step (i).

For additional details on simulated annealing, we refer the reader to Cortez (2021), Xiang et al. (2013), or Givens and Hoeting (2005).

A possible enhancement of simulated annealing is refining its result with a local search (Givens & Hoeting, 2005), for which we used the BFGS formula of gradient descent. Since the inverse Hessian matrix is calculated within each run, gradient descent-based methods also have the potential advantage of utilizing MLE's asymptotic normality for statistical inference of the estimated parameters (Nash, 2014; Michimae & Emura, 2022).

To implement the proposed global and local searches, we used the *GenSA* and base *stats* packages in R (R foundation for statistical computing). The generalized simulated annealing method implemented by the *GenSA* package has been validated to perform well

in solving non-linear objective functions with many local minima (Xiang, Gubian, & Martin, 2017), and the optim() function in the *stats* package is routinely used in ML-based estimations. Hyper-parameters such as the maximum number of iterations within each temperature and the initial temperature in the GenSA() function, or the number of function calls for each candidate solution in optim() were tuned specific to the current study's objective function. Due to the long and greatly varying runtimes of GenSA(), the *parallel* and *doParallel* packages in R that utilize multiple cores of the computer CPU were used to simultaneously process multiple runs of newly generated data and their bootstrap samples for MAE and bootstrap CI calculations.

The nonparametric bootstrap and the percentile method (Givens & Hoeting, 2005) were used to construct bootstrap CIs as follows. B independent random resamples were taken with replacement from the originally simulated (T, C) data with same sample sizes n. The correlation (dependence) between T and C was estimated via the proposed method in each of the b = 1, 2, …, B bootstrap samples, where the b^{th} estimated Kendall's tau = $\hat{\tau}_{(b)}$. Rearranging the tau estimates in increasing order,

$$\hat{\tau}_{(1)} \leq \hat{\tau}_{(2)} \leq \cdots \leq \hat{\tau}_{(B)}$$

the $100 \cdot (1-\alpha)$ percentiles CI of τ for number of bootstraps B was calculated as

$$\left[\hat{\tau}_{\left(B\cdot\frac{\alpha}{2}\right)}, \hat{\tau}_{\left(B\cdot\left(1-\frac{\alpha}{2}\right)\right)}\right].$$

Chapter 4

Simulation Study

4.1 Part 1: Estimation of correlation (dependence) in bivariate competing risks survival data

4.1.1 Estimation of correlation with different marginal distributions for an underlying Normal (Gaussian) copula

- *Simulation settings*

The overall simulation configuration was based upon varying the three factors below.

(i) Marginal distributions of bivariate competing risks data: Exponential, Weibull, and Log-Normal marginals

(ii) Functional form of copulas: Normal (Gaussian), Clayton, Frank, and Gumbel copulas

(iii) Size of correlations: 0, 0.3, 0.5, and 0.8

We first present the results for varying factors (i) and (iii) under a Normal (Gaussian) copula linking the bivariate survival times. The results for varying the functional form of copulas, such as the Archimedean copulas of Clayton, Frank, and Gumbel, are presented in section 4.1.2.

The parameter settings for each of the marginal distributions followed those used in previous studies of copula modeling in competing risks survival analysis. Specifically, the Exponential distributions' parameters were set as

T ~ *Exponential(0.023)*, C ~ *Exponential(0.025)*, for the time to event of interest T, time to other competing events or dependent censoring C, respectively, which is from the renal transplant data example of Sorrell et al. (2021).

The Weibull distributions' (shape, scale) parameters were set as

T ~ *Weibull(0.63, 0.06)*, C ~ *Weibull(0.86, 0.04)*, which is also from Sorrell et al. (2021).

The Log-Normal distributions' (mean, standard deviation) of log(T) and log(C) were set as

T ~ *LogNormal(2.2, 1.0)*, C ~ *LogNormal(2.0, 0.25)*, from the simulation study of Czado and Van Keilegom (2022).

The Normal (Gaussian) copula's dependence parameter ρ was set as 0, 0.4539905, 0.7071068, and 0.9510565, corresponding to a Kendall's tau of 0 (independence), 0.3, 0.5, and 0.8, respectively (Table 3.1).

As demonstrated in section 2.1, the conditional distribution (conditional cdf) method was used to generate bivariate survival times that follow certain marginal distributions and are correlated by a certain size or strength and functional form. For example, after generating a random variable U = u ~ *Uniform(0,1)* and its correlated random variable V = v ~ *Uniform(0,1)* through the conditional cdf (or copula partial derivative) of a Normal

copula, consider a bivariate Weibull distribution with marginals T ~ *Weibull*(α_T, λ_T) and C ~ *Weibull*(α_C, λ_C) for shape α and scale λ. From a Weibull distribution's hazard, cumulative hazard, and survival functions

$$h(t) = \alpha\lambda t^{\alpha-1}, H(t) = \int_0^t h(u)du = \lambda t^{\alpha}, S(t) = \exp[-H(t)] = \exp[-\lambda t^{\alpha}],$$

the correlated bivariate Weibull survival times of T (= t) and C (= c) were generated as

$$u = S(t) = exp[-\lambda_T t^{\alpha_T}], t = S^{-1}(u) = \{-\frac{\log(u)}{\lambda_T}\}^{1/\alpha_T},$$

$$v = S(c) = exp[-\lambda_C t^{\alpha_C}], c = S^{-1}(v) = \{-\frac{\log(v)}{\lambda_C}\}^{1/\alpha_C}.$$

For the simulated bivariate competing risks data (T, C) of correlation ρ_0 and sample size n, for which we varied n from 500 to 4000, the observed survival time X = min(T, C) and status indicator δ = I(T ≤ C) were defined. The "sample mean" information of (X, δ): mean of observed T = $\mu_{\bar{t}_m}$, variance of observed T = $\sigma^2{}_{\bar{t}_m}$, mean of observed C = $\mu_{\bar{c}_{n-m}}$, variance of observed C = $\sigma^2{}_{\bar{c}_{n-m}}$, and the proportion of T that was observed = $\frac{m}{n}$, were set as the true values to compare those of a hypothetical BVN distribution against, where min(T, C) = T (i.e. δ = 1) m out of n times (section 3.2). An objective function for minimization was defined using the MAPE criterion with equal weights, such that the five BVN distribution parameters $\theta = (\mu_{Y_1}, \mu_{Y_2}, \sigma^2_{Y_1}, \sigma^2_{Y_2}, \rho)$ minimizing the MAPE of the five error terms defined in section 3.3 were estimated as the desired true parameter values. Especially, the resulting BVN correlation parameter ρ was estimated as the true

correlation of interest ρ_0 of the bivariate competing risks data (T, C).

For each simulated data (T, C), random resampling with replacement bootstraps of equal sample sizes n were used for the estimation of standard errors (SEs) and 95% CIs. 200 bootstrap samples were taken for the point estimate, bootstrap SE, and bootstrap 95% CI of the correlation ρ_0. In addition, multiple runs of data generation and subsequent bootstrap sampling were conducted to obtain the mean bias or mean absolute error (MAE) and the coverage probability (CP) of the bootstrap 95% CI regarding the true correlation ρ_0. Here, we conducted 50 multiple runs of 50 bootstrap samples each.

As noted in section 3.3, a global search of possible solutions followed by a local search of refining the candidate solutions was used for the estimation of ρ_0. First, the global search utilized simulated annealing via the *GenSA* package in R for the four possible correlation ranges of [-0.1, 0.2), [0.2, 0.4), [0.4, 0.6), and [0.6, 0.9] for zero, low, intermediate, and high correlations, assuming the situations of either independence or positive correlation between T and C. Negative correlations were considered to be dealt with by negating either one of the bivariate survival times. Conservative lower and upper bounds for the remaining four BVN mean and variance parameters were assigned as the possible parameter space for simulated annealing's search of candidate solutions. For example, 0.5 times the observed sample means and standard deviations of (T, C) were assigned as the lower bound, and 5 times the observed values as the upper bound. For each candidate solution (the BVN parameters vector θ), four separate GenSA searches corresponding to the four possible correlation ranges were conducted, and the correlation

range resulting in the smallest objective function value among the four possible ranges of [-0.1, 0.2), [0.2, 0.4), [0.4, 0.6), and [0.6, 0.9] received an upvote. Thus, for the 200 bootstrap samples of a single data generation, the 200 voting results of GenSA determined the most likely range of correlation among zero, low, intermediate, or high correlation of the initially generated (T, C) data.

Because the runtime of simulated annealing varied greatly by its settings, the GenSA() function hyper-parameters of maximum number of iterations within each temperature (maxit) and the initial starting temperature (temp) were tuned specific to the current study's objective function using iterated racing of the *irace* package in R (Cortez, 2021; Lopez-Ibanez et al., 2016). Specifically, they were set as maxit=7310 and temp=820, compared to the default setting of maxit=10000 and temp=1000 by Cortez (2021), resulting in relatively faster runtimes.

Next, based on the range of correlation determined by the global search, a local search using gradient descent was performed via the optim() function in R. Here, all five BVN distribution parameters were allowed to vary freely within their possible bounds. Especially, the correlation parameter ρ was allowed to vary within its pre-determined range among one of [-0.1, 0.2), [0.2, 0.4), [0.4, 0.6), or [0.6, 0.9], based on its previous global search result. Overall, the resulting local search estimate of ρ of the hypothesized BVN distribution served as the point estimate of the true ρ_0 of the simulated (T, C) data, and the estimates of ρ among the bootstrap samples of the simulated (T, C) data were used for the bootstrap SE and 95% CI calculations. The dataset & bootstrap samples generation and

global & local searches were performed multiple times (=50) for the MAE and CP calculations.

The simulated sample size n varied from 500 to 4000, for which the resulting correlation estimations were affected to some degree. Specifically, a smaller sample size sufficed for the estimation of either zero or high (= 0.8) correlations, while a larger sample size was required to adequately distinguish between low (= 0.3) or intermediate (= 0.5) correlations. Thus, sample sizes of 500 or 1000 were used for the estimation of zero or large (= 0.8) correlations, while larger sample sizes of 2000, 3000, or 4000 were used for the estimation of low (= 0.3) or intermediate (= 0.5) correlations. After obtaining the BVN correlation parameter estimate $\hat{\rho}$, it was transformed to Kendall's tau as $\hat{\tau} = 2/\pi \cdot \sin^{-1}\hat{\rho}$ for the presentation of results and subsequent interpretations. Due to the comparatively long runtimes of the GenSA() function, the number of multiple runs and bootstrap samples were set to 50. The 50 multiple runs were simultaneously conducted using 25 cores of a 64-core computer for 50/25 = 2 parallel jobs each per core.

- *Simulation results*

Simulation results of correlation (dependence) estimation in bivariate competing risks data using the proposed method of estimating the correlation parameter in a hypothesized BVN distribution are shown in Tables 4.1 and 4.2.

Table 4.1 shows the good performance of the proposed method in terms of the point estimate and its bootstrap SE of the true correlation in a simulated (T, C) dataset. The

proposed method works well under various parametric distributions often used in survival data analysis, such as the Exponential, Weibull, or Log-Normal distributions. The bootstrap 95% CIs also demonstrate the proposed method's capability of distinguishing between independence, low, intermediate, or high correlation in bivariate competing risks data, which is especially true when the survival times are Log-Normally distributed.

The results of correlation estimation in multiple runs of an initially simulated (T, C) dataset and its bootstrap samples are presented in Table 4.2. The mean estimate and mean absolute error (MAE, or mean bias) demonstrate the proposed method's accuracy in estimating the underlying correlation, and the coverage probabilities (CPs) of the bootstrap 95% CIs are also near the desired 95% under a 0.05 significance level. In addition, since zero correlation implies independence only in Normal (Gaussian) copulas (Hogg, McKean, & Craig, 2013), the use of Normal copulas in our simulations enabled us to make conclusive statements of independence between the competing time-to-events when the bootstrap 95% CI of the estimated correlation included zero.

Table 4.1: Simulation results of correlation (dependence) estimation in bootstrap samples of bivariate competing risks data (T, C) [a] with the proposed method, where the underlying copula linking the two marginal distributions is the Normal (Gaussian) copula

Marginal distribution	True Kendall's tau	Estimated Kendall's tau with proposed method			
		Point estimate	Bootstrap [b] MAE	Bootstrap SE	Bootstrap 95% CI
Exponential	0	0.088	0.007	0.024	(-0.099, 0.245)
	0.3	0.374	0.000	0.050	(0.206, 0.398)
	0.5	0.497	0.005	0.040	(0.413, 0.615)
	0.8	0.696	0.039	0.033	(0.586, 0.863)
Weibull	0	0.005	0.043	0.077	(-0.067, 0.186)
	0.3	0.249	0.042	0.085	(0.127, 0.471)
	0.5	0.450	0.031	0.085	(0.347, 0.679)
	0.8	0.720	0.015	0.033	(0.694, 0.824)
Log-Normal	0	-0.070	0.064	0.081	(-0.108, 0.207)
	0.3	0.266	0.022	0.045	(0.188, 0.381)
	0.5	0.483	0.023	0.032	(0.389, 0.552)
	0.8	0.825	0.059	0.071	(0.628, 0.912)

Abbreviations: CI, confidence interval; MAE, mean absolute error; SE, standard error
a Simulated (T, C) data's sample sizes were 500 or 1000 for Kendall's tau of zero or 0.8, and 2000, 3000, or 4000 for Kendall's tau of 0.3 or 0.5
b 200 bootstrap samples of the originally simulated (T, C) data were used for the calculations

Table 4.2: Simulation results of correlation (dependence) estimation in multiple runs of bootstrap samples of bivariate competing risks data (T, C) [a] with the proposed method, where the underlying copula linking the two marginal distributions is the Normal (Gaussian) copula

Marginal distribution	True Kendall's tau	Estimated Kendall's tau with proposed method			
		Mean estimate	MAE	Empirical SE	CP [b]
Weibull	0	0.061	0.077	0.081	96
	0.3	0.262	0.069	0.075	92
	0.5	0.425	0.091	0.068	82
	0.8	0.733	0.080	0.031	90
Log-Normal	0	0.013	0.063	0.083	98
	0.3	0.287	0.055	0.060	94
	0.5	0.492	0.054	0.049	96
	0.8	0.751	0.068	0.030	96

Abbreviations: CP, coverage probability; MAE, mean absolute error; SE, standard error
a Simulated (T, C) data's sample sizes were 500 or 1000 for Kendall's tau of zero or 0.8, and 2000, 3000, or 4000 for Kendall's tau of 0.3 or 0.5
b 50 multiple runs of 50 bootstrap samples of the originally simulated (T, C) data were used for the CP calculations

4.1.2 Estimation of correlation with different copulas for underlying Weibull marginal distributions

- *Simulation settings*

Here, the functional form of copulas was varied among the Normal (Gaussian), Clayton, Frank, and Gumbel copulas in estimating correlations of 0 (independence), 0.3, 0.5, and 0.8. The marginal distributions of T and C were set to follow a Weibull distribution.

For the copula dependence parameter values corresponding to a Kendall's tau of 0.3, 0.5, and 0.8, the Normal (Gaussian) copula's values were 0.4539905, 0.7071068, and 0.9510565, the Clayton copula's 0.8571429, 2, and 8, the Frank copula's 2.917434, 5.736283, and 18.19154, and the Gumbel copula's 1.428571, 2, and 5. (Table 3.1). The independence copula $u \cdot v$ with dependence parameter value 0 was used to generate independent (T, C) data. The marginal distribution parameters for *Weibull (shape, scale)* were T ~ *Weibull(0.63, 0.06)* and C ~ *Weibull(0.86, 0.04)*, as noted in section 4.1.1.

The process of dataset & bootstrap samples generation and global & local searches followed the settings outlined in section 4.1.1, as well as the simulated sample sizes and parallel execution via multiple cores. Due to time constraints, 50 bootstrap samples were taken for the calculation of SE and 95% CIs.

A peculiar phenomenon was observed when bootstrap sample means were taken from the initially generated (T, C) data. A shrinkage or enlargement of the correlation between T and C occurred for the sample means of (T, C) when the original dataset was generated

using the Archimedean copulas of Clayton, Frank, or Gumbel. By the bivariate CLT, the correlations were theoretically expected to be identical between the original (T, C) data and the aggregation of bootstrap sample means $(\bar{T}, \bar{C})$ data. Thus, we underwent a separate simulation of this phenomenon, where for each initially generated (T, C) data of sample sizes 500 or 1000, 1000 bootstrap samples were taken, and the calculations of $(\bar{T}, \bar{C})$ in each bootstrap sample were aggregated into a single dataset for the subsequent calculation of the correlation between $\bar{T}$ and $\bar{C}$. This process was repeated 200 times that resulted in 200 datasets of 1000 bootstrap sample means as each of their observations, for which the mean and 95% CI of the 200 correlation values were calculated (Table 4.3). The underlying marginal distributions were Weibull-distributed as described above, and the phenomenon was investigated in the Normal (Gaussian), Clayton, Frank, and Gumbel copulas for correlations of either 0.3, 0.5, or 0.8 to be the true values. The cases for Exponential or Log-Normal marginals are included in the Appendix (Tables A1~A2).

In addition, we compared the performance of our proposed method to that of conventional maximum likelihood estimation (MLE). MLE constructs the likelihood function of (X, δ), X = min(T, C), and δ = I(T ≤ C) by considering the (conditional cdf of C)·(pdf of T) and (conditional cdf of T)·(pdf of C) terms separately, depending on whether δ = 1 and 0, and creates the composite likelihood by multiplying the separate terms. Specifically, the likelihood function can be constructed as

$$L(\Theta) = \prod_{i=1}^{n} \left\{ \Pr(T_i = x_i, C_i > x_i)^{\delta_i} \cdot \Pr(C_i = x_i, T_i > x_i)^{1-\delta_i} \right\}$$

$$= \prod_{i=1}^{n} [\{\frac{\partial}{\partial S_T(x_i;\theta_T)} \mathcal{C}_\alpha(S_T(x_i;\theta_T), S_C(x_i;\theta_C)) \cdot f_T(x_i;\theta_T)\}^{\delta_i} \cdot$$

$$\{\frac{\partial}{\partial S_C(x_i;\theta_C)} \mathcal{C}_\alpha\big(S_T(x_i;\theta_T), S_C(x_i;\theta_C)\big) \cdot f_C(x_i;\theta_C)\}^{1-\delta_i}],$$

where the parameters vector $\Theta = (\alpha, \theta_T, \theta_C)^T$, α: copula dependence parameter(s),

θ_T, θ_C: marginal distribution parameters of T, C, respectively,

$X_i = \min(T_i, C_i) = x_i, \delta_i = \mathrm{I}(T_i \leq C_i)$: the realized time-to-event and event status of the i^{th} subject,

$\mathcal{C}_\alpha(u, v)$: the copula or joint cdf of U = u ~ *Uniform(0,1)*, V = v ~ *Uniform(0,1)* with copula parameter α,

$S_T(x_i;\theta_T), S_C(x_i;\theta_C)$: the marginal survival functions of T, C, respectively,

$f_T(x_i;\theta_T), f_C(x_i;\theta_C)$: the marginal pdfs of T, C, respectively.

Subsequent numerical iterations to maximize the log-likelihood, usually by gradient descent, essentially attempts to estimate the copula dependence parameter(s) α and the marginal distribution parameters θ_T, θ_C simultaneously. We used the R code provided by Sorrell et al. (2021) for MLE of the correlation (dependence) where the copulas and correlations were varied as described above, and the marginals were Weibull-distributed (Table 4.4).

- Simulation results

Table 4.3 demonstrates the phenomenon of correlation (dependence) shrinkage or enlargement in the Archimedean copulas of Clayton, Frank, and Gumbel, compared to that of the Normal (Gaussian) copula. The average correlation value of the 200 simulated (T, C) datasets was used as the 'true' correlation to compare against when calculating the percentage (%) of shrinkage or enlargement in each case. Although some correlation shrinkage was observed, the Normal copula had the least difference in correlations between those of the initially generated data and the bootstrap sample means.

An enlargement in correlation was observed when the correlated data was generated via the Clayton copula, where the enlargement was as high as 37.9%. When the marginals were correlated using the Frank or Gumbel copulas, a correlation shrinkage of around 20~30% was observed. The relative size of correlation shrinkage/enlargement seemed to differ not only by the functional form of copulas, but also by the underlying size of the true correlation and the marginal distributions of the bivariate survival times (Appendix tables A1~A2). No detailed previous literature was found regarding this phenomenon.

Table 4.4 shows the simulation results of our proposed method and those of MLE for the estimation of correlation (dependence) in bivariate competing risks data when the copulas are varied and the marginals are Weibull-distributed. The point estimates, SEs, and 95% CIs of our proposed method demonstrate its robust performance across different copulas, regardless of their functional forms. The proposed method's ability to distinguish

between different sizes of the correlation is again shown by the non-overlapping 95% CIs, especially in the case of the Clayton copula linking the marginal distributions.

A point of note is that due to the correlation shrinkage/enlargement by Archimedean copulas in Table 4.3, correlation estimates via the proposed method in Table 4.4 were compensated for by their respective amounts of shrinkage or enlargement. This is because the proposed method utilizes the bivariate CLT and inevitably, its theoretical property that the correlation of the original data is preserved in its sample mean. Thus, for example, the point estimate of 0.554 for a true tau of 0.5 in a (T, C) dataset linked by a Frank copula in Table 4.4 was calculated by compensating for the Frank copula's correlation shrinkage of 30.1% in a simulated dataset with true Kendall's tau = 0.5 in Table 4.3.

In contrast, the correlation estimates via MLE were subpar with largely biased point estimates, large SEs, and wide 95% CIs. The consistently large SEs and resultant wide CIs highlight the instability of MLE in simultaneously estimating the copula and marginal distribution parameters. The MLE results were relatively better for correlated datasets constructed by the Clayton copula, although the results of our proposed method had smaller SEs and narrower 95% CIs in this case as well.

Table 4.3: Correlation shrinkage or enlargement in bootstrap sample means of bivariate competing risks data (T, C) [a] with Weibull marginal distributions, where T and C are linked via the Normal, Clayton, Frank, and Gumbel copulas

Copula	True Kendall's tau	Simulated data avg. correl.	Bootstrap sample mean avg. correl.	95% CI of bootstrap sample mean correl.	Shrinkage or Enlargement
Normal	0.3	0.302	0.258	(0.178, 0.343)	-14.6%
	0.5	0.498	0.447	(0.366, 0.519)	-10.2%
	0.8	0.802	0.755	(0.719, 0.784)	-5.9%
Clayton	0.3	0.301	0.415	(0.328, 0.502)	37.9%
	0.5	0.500	0.632	(0.564, 0.683)	26.4%
	0.8	0.799	0.844	(0.823, 0.862)	5.6%
Frank	0.3	0.303	0.205	(0.122, 0.290)	-32.3%
	0.5	0.498	0.348	(0.230, 0.425)	-30.1%
	0.8	0.800	0.591	(0.491, 0.675)	-26.1%
Gumbel	0.3	0.298	0.193	(0.124, 0.265)	-35.2%
	0.5	0.499	0.359	(0.278, 0.439)	-28.1%
	0.8	0.800	0.674	(0.398, 0.726)	-15.8%

Abbreviations: avg., average; CI, confidence interval; correl., correlation
a 200 datasets, where each consists of 1000 bootstrap sample mean observations from the originally simulated (T, C) data with a sample size of 500

Table 4.4: Simulation results of correlation (dependence) estimation in bivariate competing risks data (T, C) [a] with the proposed method, compared to those with MLE, where the underlying marginal distributions are Weibull distributed

Copula	True tau	Estimated Kendall's tau with proposed method			Estimated Kendall's tau with MLE		
		Point estimate	Bootstrap[b] SE	Bootstrap 95% CI	Point estimate	Bootstrap SE	Bootstrap 95% CI
Indep.	0	0.005	0.077	(-0.067, 0.186)	0.268	0.280	(0.096, 0.891)
Normal	0.3	0.249	0.085	(0.127, 0.471)	0.610	0.197	(0.096, 0.749)
	0.5	0.450	0.085	(0.347, 0.679)	0.654	0.202	(0.165, 0.815)
	0.8	0.720	0.033	(0.694, 0.824)	0.891	0.205	(0.289, 0.891)
Clayton	0.3	0.277	0.037	(0.265, 0.399)	0.373	0.078	(0.180, 0.473)
	0.5	0.549	0.040	(0.505, 0.652)	0.597	0.062	(0.388, 0.651)
	0.8	0.832	0.016	(0.700, 0.891)	0.921	0.093	(0.638, 0.955)
Frank	0.3	0.452	0.078	(0.247, 0.538)	0.498	0.227	(0.070, 0.909)
	0.5	0.554	0.075	(0.299, 0.557)	0.546	0.165	(0.340, 0.918)
	0.8	0.728	0.053	(0.613, 0.790)	0.669	0.191	(0.279, 0.963)
Gumbel	0.3	0.340	0.083	(0.247, 0.572)	0.013	0.139	(0.043, 0.641)
	0.5	0.479	0.076	(0.284, 0.568)	0.547	0.162	(0.076, 0.935)
	0.8	0.749	0.032	(0.602, 0.853)	0.910	0.163	(0.535, 0.972)

Abbreviations: CI, confidence interval; Indep., independence copula; SE, standard error
a Simulated (T, C) data sample sizes were 500 or 1000 for Kendall's tau of zero or 0.8, and 2000, 3000, or 4000 for Kendall's tau of 0.3 or 0.5
b 50 bootstrap samples of the originally simulated (T, C) data were used for the calculations

4.2 Part 2: Estimation of marginal survival and the effect of a binary treatment variable in bivariate competing risks survival data

4.2.1 Subsequent estimation of marginal survival following the estimation of correlation in bivariate competing risks survival data

- *Simulation settings*

The study objective in biomedical research given some bivariate competing risks survival data may be the unbiased estimation of the marginal survival probability (or hazard rate, with a 1-to-1 correspondence between the two) over time for an event of interest. In addition, the researcher would likely be interested in the effect (or regression coefficient) of some treatment or exposure on the marginal hazard rate of the event of interest. Therefore, we simulated these situations that would subsequently follow after the estimation of correlation (dependence) in bivariate competing risks data.

Two simulation settings were devised for the estimation of marginal survival probability over time for an event of interest, considering the possible correlation of a competing event.

First, 200 samples from the Log-Normal marginal distributions simulated in the previous section with mean and standard deviation for each of log(T) and log(C) as $T \sim LogNormal(2.2, 1.0)$, $C \sim LogNormal(2.0, 0.25)$, were linked via a Gumbel copula for a

simulated Kendall's tau = 0.339, and the marginal survival curve of T was subsequently plotted by

(i) using the proposed method to estimate the correlation between T and C as 0.304,

(ii) incorrectly assuming zero correlation or independence (cause-specific hazards),

(iii), (iv) incorrectly specifying the correlation as 0.5, 0.8, respectively.

The copula-graphic estimator by Zheng and Klein (1994, 1995) introduced in Ch. 2 and its implementation via the *compound.Cox* package in R (Emura and Chen, 2016), together with the *survival* package in R, were used to plot the marginal survival curves of T.

Second, 500 samples from Weibull and Exponential marginal distributions of $T \sim Weibull(2, 0.25)$ and $C \sim Exponential(0.2)$, were linked via a Clayton copula for a simulated Kendall's tau = 0.791, and the marginal survival curve of T was subsequently plotted by

(i) using the proposed method to estimate the correlation between T and C as 0.769,

(ii) incorrectly assuming zero correlation or independence (cause-specific hazards),

(iii), (iv) incorrectly specifying the correlation as 0.3, 0.9, respectively.

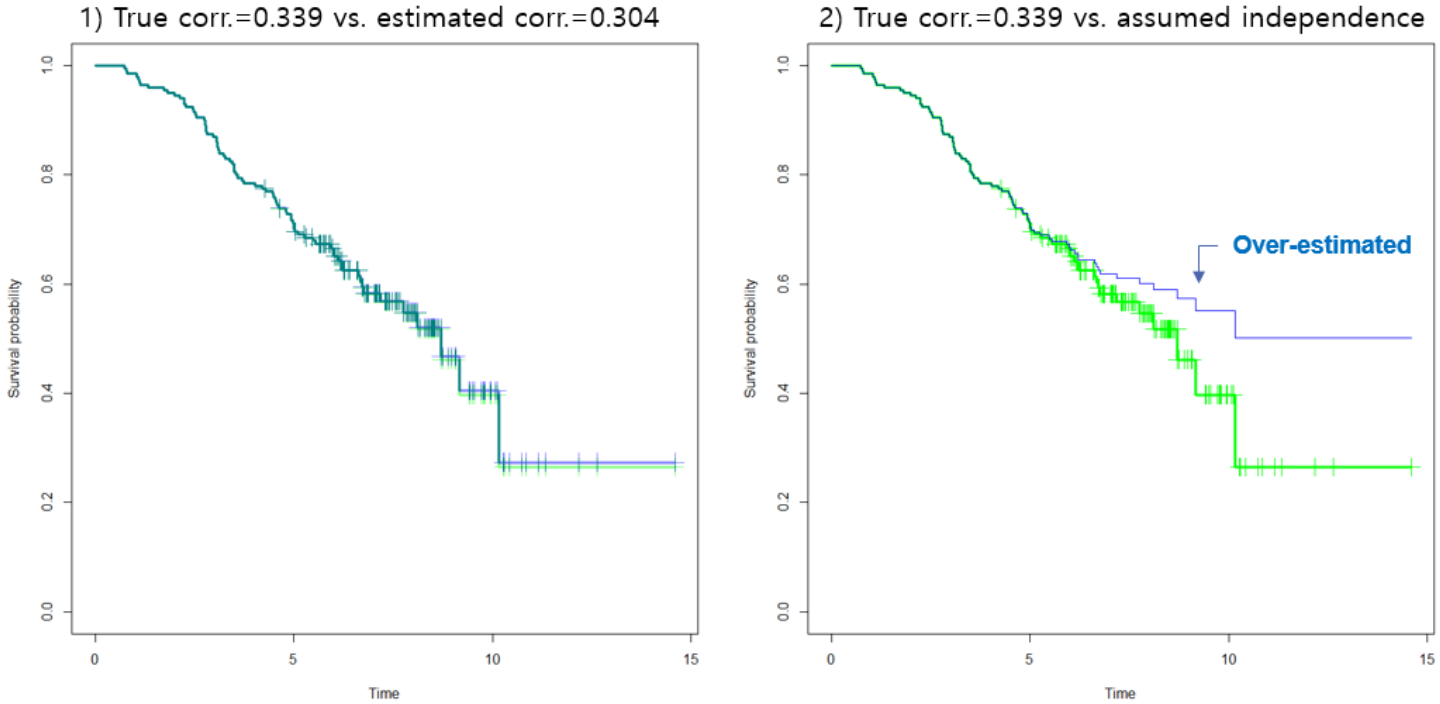
In the next section 4.2.2, this example expands to a hypothetical randomized clinical trial (RCT) setting with a single binary treatment variable, for which the estimation of its efficacy (regression coefficient and its statistical significance) is the study objective.

- *Simulation results*

The main purpose of Figures 4.1.1 and 4.1.2 are to visually demonstrate the unbiased marginal survival curve when the correlation (dependence) between T and C is correctly estimated via our proposed method, compared to the biased marginal survivals under incorrectly specified correlations of 0, 0.5, or 0.8. The well-known K-M survival curve assumes independent censoring, corresponding to the largely over-estimated survival curve in the right of Figure 4.1.1. For positively (negatively) correlated time-to-events T and C, the K-M curve will always over-estimate (under-estimate) the marginal survival curve as shown. Naturally, if the correlation is wrongly assumed as larger (smaller) than the actual positive correlation, this will result in under-estimation (over-estimation) of the marginal survival probability over time, as in Figure 4.1.2.

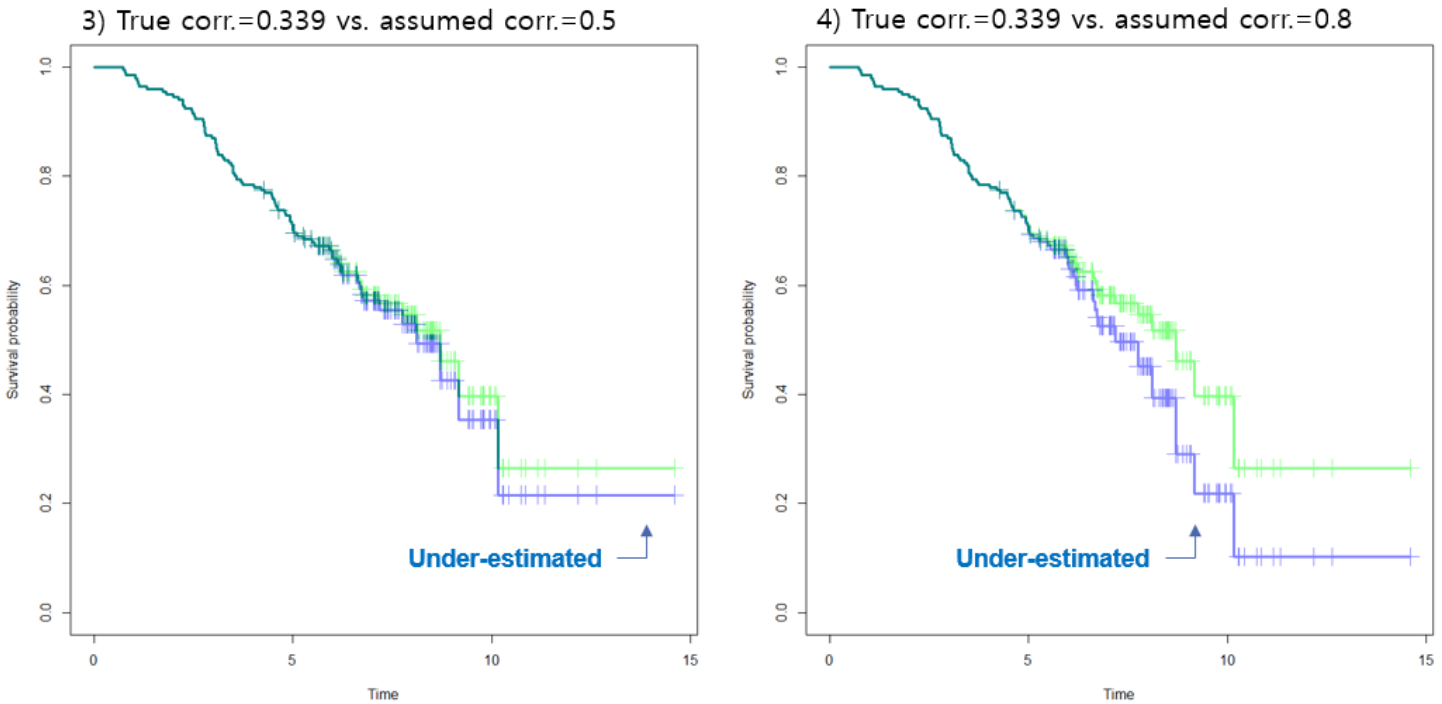
Although the biasedness of the marginal survival of T is less pronounced in Figures 4.2.1~4.2.2, the over-estimation under incorrectly smaller correlations (right of Figure 4.2.1, left of Figure 4.2.2) and under-estimation under an incorrectly larger correlation (right of Figure 4.2.2) can still be seen. Under the previously mentioned hypothetical RCT scenario of a new treatment to a disease, the time to event of interest T may be the overall survival (OS) time of a patient, while the time to competing event C may be the observed time until patient withdrawal from the trial due to deteriorating health or adverse effects of the new treatment. Biasedness of the marginal survival of T, especially that of the universal K-M curve, then equates to the patients' marginal OS probability over time under the new treatment, or efficacy of the new treatment, being incorrectly estimated.

Figure 4.1.1: Simulation results of marginal survival curves of the time to event of interest T in bivariate competing risks data (T, C) [a], where the "true" marginal survival curve of T with Kendall's tau = 0.339 is plotted in green, and the marginal survival curves after either estimating the correlation between (T, C) via the proposed method or assuming the correlation to be zero (independence) are plotted in blue



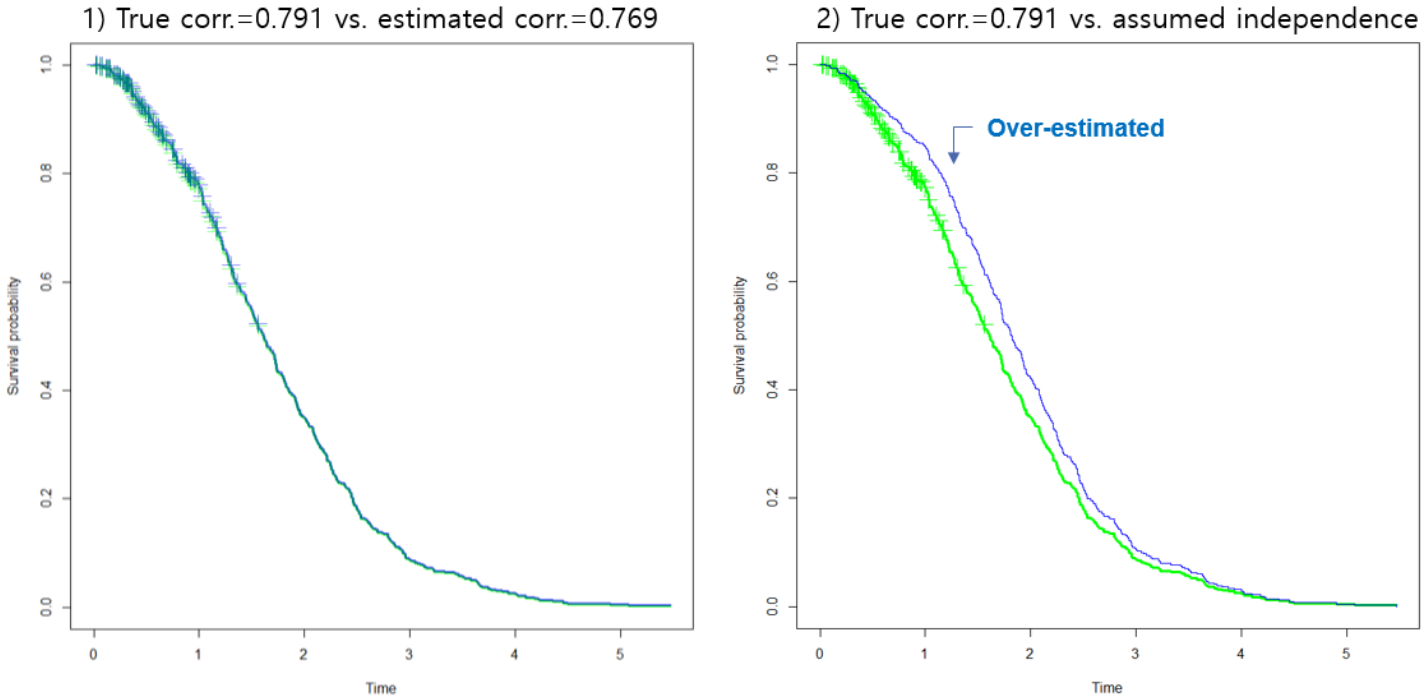
a The underlying marginal distributions are Log-Normally distributed and the copula linking the marginal distributions is the Gumbel copula

Figure 4.1.2: Simulation results of marginal survival curves of the time to event of interest T in bivariate competing risks data (T, C) [a], where the "true" marginal survival curve of T with Kendall's tau = 0.339 is plotted in green, and the marginal survival curves after assuming the correlation between (T, C) to be either 0.5 or 0.8 are plotted in blue



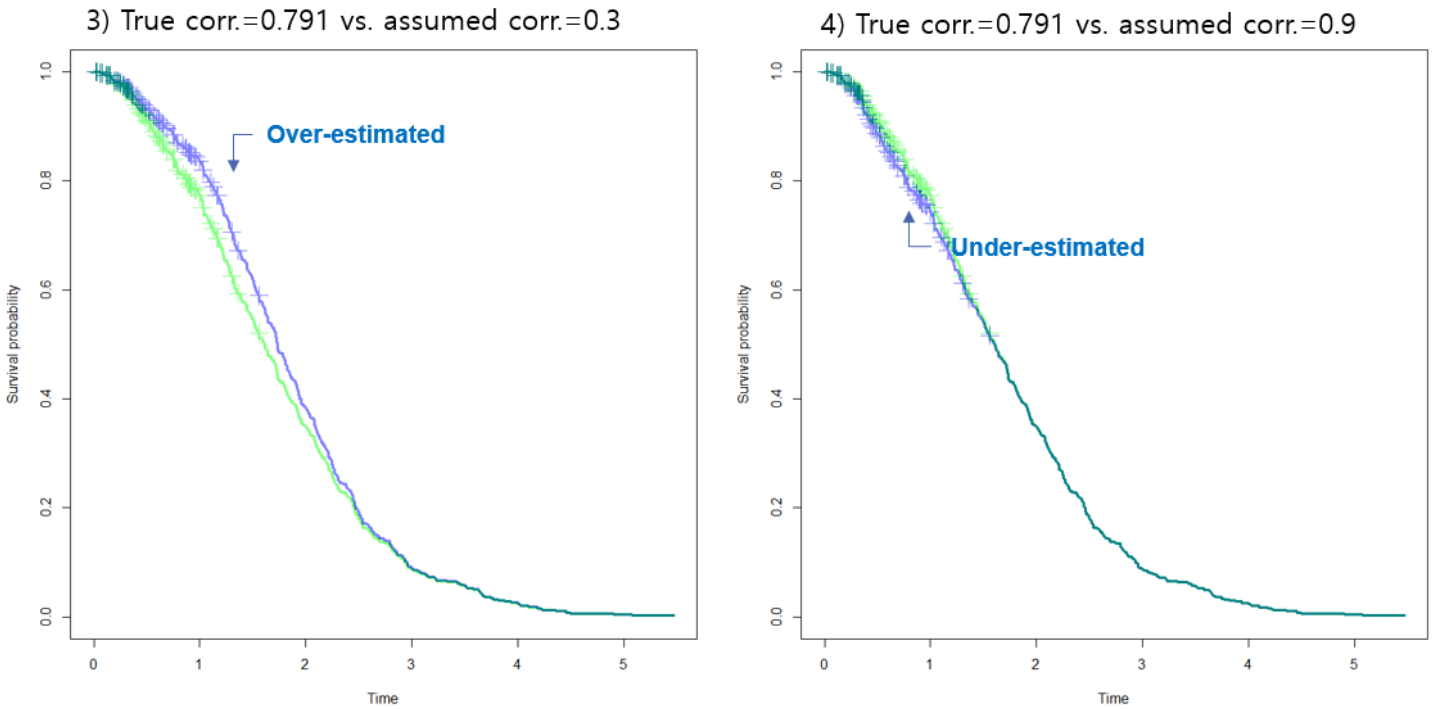
a The underlying marginal distributions are Log-Normally distributed and the copula linking the marginal distributions is the Gumbel copula

Figure 4.2.1: Simulation results of marginal survival curves of the time to event of interest T in bivariate competing risks data (T, C) [a], where the "true" marginal survival curve of T with Kendall's tau = 0.791 is plotted in green, and the marginal survival curves after either estimating the correlation between (T, C) via the proposed method or assuming the correlation to be zero (independence) are plotted in blue



a The underlying marginal distributions are Weibull and Exponentially distributed for T and C, respectively, and the copula linking the marginal distributions is the Clayton copula

Figure 4.2.2: Simulation results of marginal survival curves of the time to event of interest T in bivariate competing risks data (T, C) [a], where the "true" marginal survival curve of T with Kendall's tau = 0.791 is plotted in green, and the marginal survival curves after assuming the correlation between (T, C) to be either 0.3 or 0.9 are plotted in blue



a The underlying marginal distributions are Weibull and Exponentially distributed for T and C, respectively, and the copula linking the marginal distributions is the Clayton copula

4.2.2 Subsequent estimation of the effect of a binary treatment variable following the estimation of correlation in bivariate competing risks survival data

- *Simulation settings*

The second example of the previous section 4.2.1 is expanded and continued here. 500 samples from Weibull and Exponential marginal distributions were generated as T ~ *Weibull(2, 0.25)* and C ~ *Exponential(0.2)*, linked via a Clayton copula for a simulated Kendall's tau = 0.791 (true tau = 0.8). In a hypothetical RCT setting of a new treatment to a disease, the time to event of interest T is the OS time of a patient, while the time to competing event C is the dependently censored time until patient withdrawal from the trial due to deteriorating health or adverse effects of the new treatment. Thus, a strongly positive correlation between T and C is hypothesized. The new treatment variable 'Trt' was generated as Trt ~ *Bernoulli(0.5)* with equal probability of either the new treatment or a placebo. As the treatment assignment was completely randomized, all other covariates such as patient age, gender, and disease characteristics were assumed to be well-balanced in the treatment and control groups, i.e., no other adjustment for covariates was needed for.

A semi-parametric Cox PH model was assumed as

$\mathrm{h}(\mathrm{x}|\mathrm{Trt}) = h_0(x) \cdot \exp\left[\beta^T \cdot Trt\right]$, X (= x) = min(T, C)

with baseline hazards of T and C as $h_{0,T}(t) = \alpha_T \lambda_T t^{\alpha-1}$ and $h_{0,C}(c) = \lambda_C$, respectively. Therefore, the hazard functions were set as

$h_T(t|Trt) = 2 \cdot 0.25 \cdot t^{(2-1)} \cdot \exp(\beta_T \cdot Trt) = 0.5 \cdot t \cdot \exp(\beta_T \cdot Trt)$,

$h_C(c|Trt) = 0.2 \cdot \exp(\beta_C \cdot Trt)$,

and the correlated survival times (T, C) were generated as

$T = [-\frac{\log(u)}{\{0.25 \cdot \exp(\beta_T \cdot Trt)\}}]^{1/2}$, $C = -\frac{\log(v)}{\{0.2 \cdot \exp(\beta_C \cdot Trt)\}}$ (section 4.1.1),

for $C_\alpha(u, v) = (u^{-\alpha} + v^{-\alpha} - 1)^{-\frac{1}{\alpha}}$ from a Clayton copula with copula dependence parameter α = 8, corresponding to a Kendall's tau = $\alpha/(2+\alpha)$ = 0.8 (Table 2.1). The true effect sizes or regression coefficients of the new treatment on T and C, β_T and β_C, were set as -0.5 and 0.2, respectively. These beta values correspond to a HR = exp(-0.5) = 0.61 of the new treatment on T (= OS time), and HR = exp(0.2) = 1.22 of the new treatment on C (= time to patient withdrawal). That is, the new treatment is largely effective in prolonging patient survival, while having a slightly detrimental effect on, or increasing the hazard of, patient withdrawal from the trial. No other censoring or competing events were assumed, as one can always combine all events other than the event of interest into a single dependent competing or censoring event.

200 bootstrap samples of the initially generated (T, C) data were taken, and the regression coefficients β_T and β_C of the treatment variable Trt were estimated under several correlation scenarios:

(i) use the proposed method to estimate the mean correlation between T and C as 0.769, and utilize the estimated correlation in the marginal Cox regression analyses,

(ii) incorrectly assume zero correlation or independent censoring (cause-specific hazards), (iii), (iv) incorrectly specify the correlation as 0.3, 0.9, respectively.

The dependCox.reg() function from the *compound.Cox* package in R, which enables univariate Cox regression under dependent censoring (Emura & Chen, 2016), was used for (i), (iii), and (iv), and the conventional coxph() function from the *survival* package for (ii).

- *Simulation results*

Tables 4.5.1~4.5.2 show the simulation results of regression coefficients estimation for a univariate Cox regression model of a binary treatment variable with dependent time-to-events (T, C). Under our simulation settings, the average proportions of T and C were 77.2% and 22.8%, respectively.

First of all, the proposed method of correlation estimation performed well with a mean estimate of 0.769 and standard error of 0.055, compared to the simulated Kendall's tau of 0.791 or true Kendall's tau of 0.8. The approaches of cause-specific hazards and two incorrectly specified correlations do not have standard errors for tau since it was not considered as a parameter for estimation.

Second, the mean estimates of the regression coefficients β_T and β_C were evidently less biased using the proposed method of correlation estimation, compared to the other three scenarios. The mean estimates of $\hat{\beta}_T$ = -0.482 and $\hat{\beta}_C$ = 0.187 were close to the true values of -0.5 and 0.2, compared to the largely over-estimated values under independence

and an incorrectly smaller or weaker correlation = 0.3, as the mean percentage error (MPE) values clearly demonstrate. The beta estimates under an incorrectly larger or stronger correlation = 0.9 between the survival endpoints became under-estimated as $\hat{\beta}_T$ = -0.436 and $\hat{\beta}_C$ = -0.130, with the direction of association even being reversed in the estimation of β_C. Absolute deviation from the true regression coefficients became larger as the assumed size or strength of the correlation deviated further from the true correlation, which emphasizes the importance of accurately estimating the correlation in competing risks or dependently censored survival data when conducting regression analyses.

Table 4.5.1: Simulation results of estimated correlation 0.769 and assumed independence, and their subsequent regression coefficients estimations for a univariate Cox regression model with bivariate competing risks data (T, C) [a]

Param.	True value	Proposed method of correlation estimation			Cause-specific hazards (Independence, $\hat{\tau} = 0$)		
		Mean estimate	MPE	Bootstr. SE	Mean estimate	MPE	Bootstr. SE
β_T	-0.5	-0.482	3.58%	0.096	-0.718	-43.6%	0.109
β_C	0.2	0.187	6.58%	0.191	0.545	-172%	0.197
τ	0.8	0.769	3.56%	0.055	0.000	-100%	-

Table 4.5.2: Simulation results of assumed correlations of 0.3 and 0.9, and their subsequent regression coefficients estimations for a univariate Cox regression model with bivariate competing risks data (T, C) [a]

Param.	True value	Incorrectly assumed correlation ($\hat{\tau} = 0.3$)			Incorrectly assumed correlation ($\hat{\tau} = 0.9$)		
		Mean estimate	MPE	Bootstr. SE	Mean estimate	MPE	Bootstr. SE
β_T	-0.5	-0.619	-23.8%	0.105	-0.436	12.8%	0.088
β_C	0.2	0.517	-159%	0.193	-0.130	164.9%	0.117
τ	0.8	0.300	-62.5%	-	0.900	12.5%	-

Abbreviations: Bootstr., bootstrap; MPE, mean percentage error; Param., parameter for estimation; SE, standard error
a The marginal distributions are Weibull and Exponentially distributed, and the copula linking the marginal distributions is the Clayton copula

Chapter 5

Real Data Analysis

5.1 Real data 1: Acute lymphoblastic leukemia (ALL) data

5.1.1 ALL data: Description and preparation

A real-world dataset of 2,279 patients with acute lymphoblastic leukemia (ALL) from the European Group for Blood and Marrow Transplantation was available via the *dynpred* package in R (van Houwelingen & Putter, 2012). The patients with ALL received an allogeneic hematopoietic stem cell transplant (AHSCT) from an HLA-identical sibling after first complete remission during the time period of 1985 – 1998. The clinically relevant events recorded were the incidence of acute graft versus host disease (AGvHD), the recovery of platelet counts to normal level, relapse of the disease, and all-cause death. AGvHD refers to an acute manifestation of the grafted or transplanted immune cells attacking its host, the patient who received AHSCT. For these documented events, the time-to-event and indicator of event occurrence (yes/no) were recorded. The median follow-up of patients was 6.6 years (van Houwelingen & Putter, 2012). Additional prognostic variables that were recorded were donor-recipient gender mismatch (yes/no), T-cell depletion prophylaxis for GvHD prevention (yes/no), age of patient at transplant (≤20, 20-40, >40), and year of transplant (1985-89, 1990-94, 1995-98).

A major marker of treatment success in ALL patients who receive AHSCT is being relapse-free. The occurrence of AGvHD, although life-threatening if very severe, is also known to have a beneficial “graft vs. leukemia” effect, where the transplanted healthy immune cells attack any residual cancerous blood cells that may remain in the host (Baron et al., 2023; Nordlander et al., 2004; Katsahian et al., 2004). Focusing on this issue, the outcome of interest was set as the time to relapse, where the main exposure or treatment is the occurrence of AGvHD. Thus, our research question was whether AGvHD incidence has a beneficial effect on the time to relapse to reduce its hazard among patients with ALL. An important issue here is the possible dependence between the endpoints of relapse and all-cause death before relapse, as it is clinically likely that patients with a longer/shorter time to relapse would also have a longer/shorter time to all-cause death, and vice-versa. The possible correlation (dependence) between these survival endpoints should thus be estimated and taken into account to unbiasedly estimate the regression coefficient and statistical significance of the exposure of interest, AGvHD incidence. The survival time of patients was observed only up to their time to relapse, as data on survival after relapse is not always reliable (van Houwelingen & Putter, 2012). Therefore, the observed survival time was the minimum of either the time to relapse or the time to all-cause death or end-of-study censoring but never both, inducing a “classical” competing risks situation.

Several other factors were considered for data preparation. First, the study population was limited to those whose platelet count recovered to normal levels, as this is an important indicator of successful “engraftment”, where the transplanted hematopoietic stem cells

initially take root and establish themselves in the host's body. The biological assumption here was that initial engraftment was necessary for AGvHD to occur later on. To avoid immortal time bias (Zhang et al., 2022), a landmark time of 100 days since transplant was used (Dafni, 2011) to define platelet recovery (yes/no), as well as AGvHD status (yes/no) by its very definition. Hence, only those who were followed up for at least 100 days, together with platelet recovery, were considered for further analysis. This resulted in a study population of 1,083 ALL patients with platelet recovery within 100 days of AHSCT. Second, we envisioned a quasi-RCT setting where the “treatment” is whether or not a patient experienced AGvHD. As such, the distribution of other available prognostic variables (gender mismatch, GvHD prevention prophylaxis, age of patient, and year of transplant) were checked to see if they were well-balanced by AGvHD status (yes/no). As these prognostic variables were all categorical, a non-significant difference in the proportions of a variable by AGvHD status with a Chi-squared test P-value >0.05 was considered as sufficient balance for each variable.

Regarding the possible correlation (dependence) between the two outcomes of time to relapse and time to either all-cause death or end-of-study censoring, three analysis scenarios were considered in assessing the effect of AGvHD on the time to relapse: 1) estimation of correlation with the proposed method, 2) assumed independence, and 3) assumed correlation of Kendall’s tau = 0.3. After either estimating or assuming the correlation between the two survival outcomes, the marginal survival probability over time of disease relapse and the regression coefficient of AGvHD incidence for the time to relapse

in a univariate Cox regression model were estimated according to each correlation scenario. The marginal survival probability over time was plotted using the original study population dataset (N=1,083) defined above, while the beta coefficient of the treatment arms variable was estimated in the original study population as well as in its 200 bootstrap samples for additional bootstrap SE and P-value calculations. The Wald statistic was used for the bootstrap P-values. The CG.Clayton() function in the *compound.Cox* R package (Emura & Chen, 2016) was used for marginal survival curve plotting, and the dependCox.reg() function of the same R package for univariate Cox regression with dependent censoring. The conventional coxph() function in the *survival* R package was used for analysis scenario 2) of assumed independence.

5.1.2 ALL data: Analysis results of applying the proposed method

Table 5.1 shows the baseline characteristics of the study population prepared for the analysis of the ALL dataset. AGvHD occurred in 52.4% of patients with ALL, and the Chi-squared test or ANOVA P-values were >0.05 for all covariates considered, indicating an adequate balance of covariates by the main exposure or “treatment” of AGvHD, in terms of the outcome of time to relapse. For disease relapse against AGvHD incidence, the proportion of relapse among patients with AGvHD was 17.6% compared to 20.2% of relapse among those without AGvHD, indicating a possible protective association of AGvHD with disease relapse. The one-sided P-value of this univariate assessment was 0.327/2=0.164. In addition, 17.5 % of patients who received GvHD prevention experienced AGvHD compared to 22.1% among those without such prophylaxis, with the P-value of

0.066 showing borderline significance. Thus, further matching by GvHD prevention status may result in a better balance of covariates between the AGvHD incidence groups (yes/no).

According to the three analysis scenarios detailed in the previous section 5.1.1, we either estimated the correlation (dependence) between the time to relapse and the time to all-cause death or end of study censoring via our proposed method, or assigned assumed correlations of zero (independence) or Kendall's tau=0.3. The correlation estimations by our proposed method, for the original real-world data and its 200 bootstrap samples, are shown in Table 5.2. The point estimate of the original data's correlation was Kendall's tau=0.673, while the mean estimate of its 200 bootstrap samples was 0.781, which is a very strong positive correlation. This is to be expected, as disease relapse and overall survival are known to correlate strongly in patients with ALL. Figure 5.1 depicts the survival probability curves under the estimated correlation being the mean estimate of 0.781 (in green) vs. the assumed independence (left, in blue) and assumed correlation of 0.3 (right, in blue) scenarios. It is evident that the resulting survival curves are very different according to the underlying correlation assumption between the time to event of interest and the time to its competing event(s). Therefore, an accurate estimation of correlation between the survival endpoints is important and necessary, rather than resorting to the independent censoring assumption and the conventional K-M survival curve (Figure 5.1; left, in blue).

Table 5.2 displays the univariate Cox regression results of the beta coefficient estimation regarding the potentially beneficial effect of AGvHD incidence on the time to relapse of ALL. Results according to the three analysis scenarios of either an estimated or

assumed correlation between the survival endpoints are shown in separate columns. Results of a single analysis of the original dataset and those of the 200 bootstrap samples of the original dataset are shown in separate rows. The SE and P-value calculations of the original dataset analysis were done by the dependCox.reg() function of the *compound.Cox* package in R (Emura & Chen, 2016). Comparing the mean estimate of the bootstrap samples and the single estimate of the original study population, the beta coefficient, SE, and P-value estimates are mostly in agreement, albeit the size of the AGvHD effect being slightly larger for the single analysis of the original data under the proposed method of correlation estimation (-0.122 vs. -0.104).

Comparing the results by the three scenarios of an estimated or assumed correlation in Table 5.2, all are in agreement that the AGvHD effect on the hazard of disease relapse is not statistically significant at the 0.05 level. However, under the one-sided test of H_0: AGvHD has a null or detrimental effect on the hazard of relapse vs. H_a: AGvHD has a protective effect on the hazard of relapse, the results under the estimated correlation of 0.781 or 0.673 estimation show statistical significance at the liberal 0.2 level with P-values of 0.177 or 0.145, while those of assumed independence or assumed Kendall's tau=0.3 do not. This is mainly due to the smaller SEs of the estimated beta coefficient under the proposed method of correlation estimation. Overall, the hypothesized protective effect of AGvHD incidence upon the subsequent relapse of ALL reaches statistical significance more closely when the apparently strong correlation between the survival outcomes is taken into consideration.

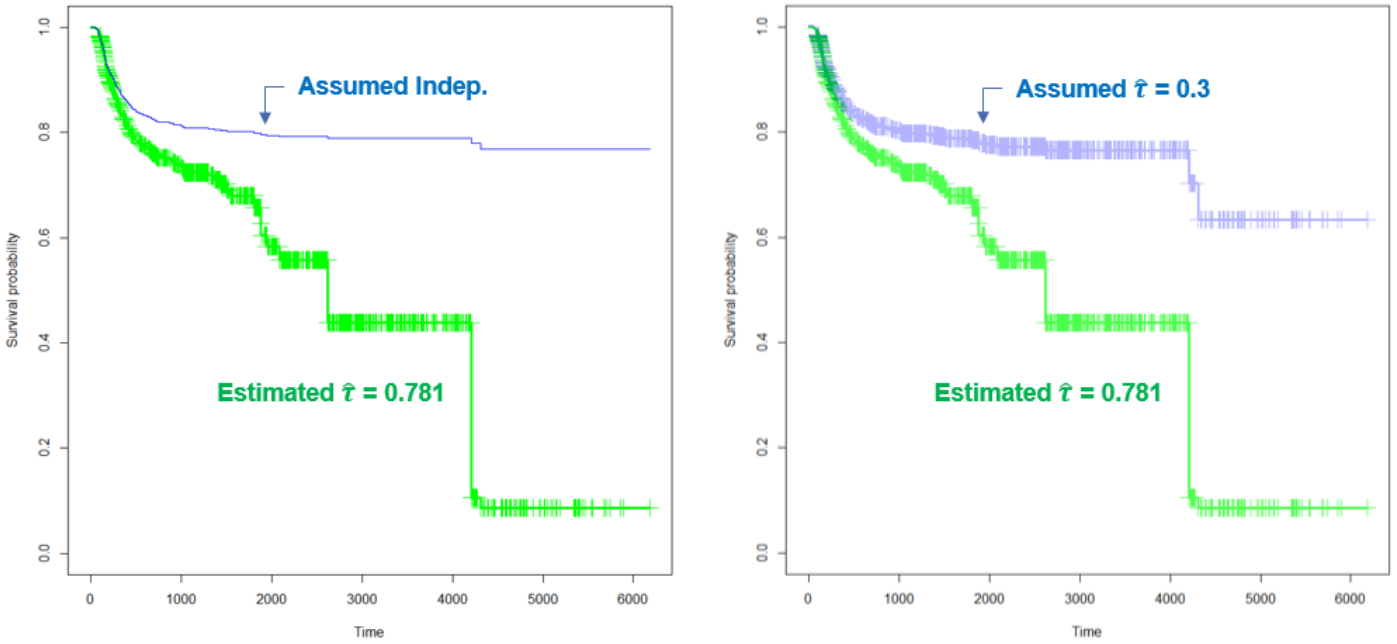
Table 5.1: Baseline characteristics of the ALL study population (N=1,083) in total and by incident AGvHD status

Variables	Total	AGvHD incidence		P-value [a]
		Yes	No	
Overall N (%)	1,083(100)	567 (52.4)	516 (47.6)	-
Year of transplant, N (%)				0.865
1985 – 1989	213 (19.7)	108 (19.0)	105 (20.3)	
1990 - 1994	442 (40.8)	233 (41.1)	209 (40.5)	
1995 – 1998	428 (39.5)	226 (39.9)	202 (39.2)	
Age at transplant, N (%)				0.536
≤ 20	275 (25.4)	138 (24.4)	137 (26.6)	
20 - 40	533 (49.2)	278 (49.0)	255 (49.4)	
> 40	275 (25.4)	151 (26.6)	124 (24.0)	
Donor-recipient mismatch, N (%)				0.267
No mismatch	835 (77.1)	429 (75.7)	406 (78.7)	
Gender mismatch	248 (22.9)	138 (24.3)	110 (21.3)	
GvHD prevention, N (%)				0.066
No prophylaxis	870 (80.3)	468 (82.5)	402 (77.9)	
T-cell depletion prophylaxis	213 (19.7)	99 (17.5)	114 (22.1)	
Disease relapse, N (%)				0.327
Yes	204 (18.8)	100 (17.6)	104 (20.2)	
No	879 (81.2)	467 (82.4)	412 (79.8)	
Follow-up time (years), mean (SD)	4.91 (3.91)	4.83 (3.97)	5.01 (3.84)	0.453

Abbreviations: AGvHD, acute graft vs. host disease; ALL, acute lymphoblastic leukemia; GvHD, graft vs. host disease; SD, standard deviation

[a] P-values were calculated with ANOVA for continuous variables and the Chi-squared test for categorical variables

Figure 5.1: Marginal survival curves of the time to relapse in the ALL study population (N=1,083) under the estimated correlation via the proposed method (green), under independence (left, in blue), and under an assumed correlation of 0.3 (right, in blue), where the copula linking the time to relapse and the time to other endpoints (all-cause death or end of study censoring) is the Clayton copula



Abbreviations: ALL, acute lymphoblastic leukemia

Table 5.2: Results of correlation (dependence) estimation with the proposed method and subsequent regression coefficient estimation for a univariate Cox regression model of AGvHD occurrence on the time to relapse in the ALL study population (N=1,083), where the copula linking the time to relapse and the time to other endpoints (all-cause death or end of study censoring) is the Clayton copula

	Proposed method of correlation estimation			Cause-specific hazards (Independence, $\hat{\tau} = 0$)			Assumed correlation ($\hat{\tau} = 0.3$)		
Params.	Mean [a] estimate	Bootstr. [b] SE	One-side P-value [c]	Mean estimate	Bootstr. SE	One-side P-value	Mean estimate	Bootstr. SE	One-side P-value
β_T	-0.104	0.113	0.177	-0.118	0.143	0.204	-0.115	0.143	0.211
τ	0.781	0.069	<0.001	0	-	-	0.3	-	-
Params.	Orig. data[d] estimate	Orig. data SE	One-side P-value	Orig. data estimate	Orig. data SE	One-side P-value	Orig. data estimate	Orig. data SE	One-side P-value
β_T	-0.122	0.115	0.145	-0.112	0.140	0.211	-0.110	0.142	0.218
τ	0.673	-	-	0	-	-	0.3	-	-

Abbreviations: AGvHD, acute graft vs. host disease; ALL, acute lymphoblastic leukemia; Bootstr., bootstrap; Orig., original; Params., parameters for estimation; SE, standard error

a The mean value of 200 regression coefficients estimated from the 200 bootstrap samples of the original ALL study population

b The standard error (empirical standard deviation) of 200 regression coefficients estimated from the 200 bootstrap samples

c A one-sided P-value calculated from the Wald statistic = $\hat{\beta}_T$ / SE($\hat{\beta}_T$)

d The single regression coefficient, SE, and P-value estimated from the original ALL study population

5.2 Real data 2: AIDS Clinical Trials Group (ACTG) Study 175 data

5.2.1 ACTG 175 data: Description and preparation

A real-world dataset of a double-blind RCT among adults infected with the human immunodeficiency (HIV) virus whose CD4 T-cell counts were 200~500/mm^3 was obtained from the *speff2trial* package in R. The study objective was to compare the efficacy of monotherapy with either zidovudine (also known as AZT) or didanosine vs. the combination therapies of AZT plus didanosine or AZT plus zalcitabine, resulting in a total of four treatment arms. The primary endpoint of the study was ≥50% decline in CD4 T-cell count, progression of HIV to AIDS, or all-cause death, whichever came first. Early patient withdrawal due to deteriorating health or toxic effects of the drug occurred during the trial, which is strongly indicative of dependent censoring that is positively correlated with the primary endpoint. Therefore, the estimation of correlation (dependence) between the primary endpoint and the competing event of patient withdrawal is necessary to unbiasedly estimate and compare the efficacy of the treatment arms. The ACTG 175 data has been analyzed by several previous studies similar to our study, focusing on the possible correlation between the time to primary endpoint and time to patient withdrawal (Deresa & Van Keilegom; 2021, Chen; 2010, Huang & Zhang; 2008).

Among the four treatment arms initially included in the data, only the two treatment arms of AZT alone (N=532) and AZT plus didanosine (N=522) were considered for a total

of 1,054 patients. Among the 1,054 patients, 284 patients experienced the primary endpoint of a decline in CD4 T-cells, progression to AIDS, or death, while 381 patients withdrew from the trial and 389 were administratively censored at the end of study. The possible dependent censoring of patient withdrawal and administrative end-of-study censoring were combined into one competing event (against that of the primary endpoint) as 381+389 = 770 patients. Since the study was a double-blind RCT with randomized treatment allocation, we expected the eight clinically relevant covariates of age, gender, race, intravenous drug use, hemophilia, baseline CD4 T-cell count, prior antiretroviral history, and disease symptoms indicator (Deresa & Van Keilegom; 2021, Chen; 2010, Huang & Zhang; 2008) to be well-balanced between the two treatment arms. Either ANOVA for continuous covariates or the Chi-squared test for categorical covariates were used to test for the covariates' sufficient balance by treatment arms at a significance level of 0.05.

Regarding the possible correlation (dependence) between the two outcomes of time to the primary endpoint and time to either withdrawal or end-of-study censoring, three analysis scenarios were considered in comparing the efficacy of treatment arms: 1) estimation of correlation with the proposed method, 2) assumed independence, and 3) assumed correlation of Kendall's tau = 0.8. After either estimating or assuming the correlation between the survival endpoints, the marginal survival probability over time of the primary endpoint and the regression coefficient of the treatment arms variable for the time to the primary endpoint in a univariate Cox regression model were estimated according to each correlation scenario. The marginal survival probability over time was

plotted using the original ACTG 175 dataset study population, while the beta coefficient of the treatment arms variable was estimated in the original ACTG 175 dataset as well as in its 200 bootstrap samples for additional bootstrap SE and P-value calculations. The Wald statistic with bootstrap SEs were used for the bootstrap P-values. The CG.Clayton() function in the *compound.Cox* R package (Emura & Chen, 2016) was used for marginal survival curve plotting, and the dependCox.reg() function of the same R package for univariate Cox regression with dependent censoring. The conventional coxph() function in the *survival* R package was used for analysis scenario 2) of assumed independence.

5.2.2 ACTG 175 data: Analysis results of applying the proposed method

Baseline characteristics of the ACTG 175 data study population (N=1,054) by the two treatment arms are shown in Table 5.3. A similar number of patients were allocated to each treatment (532 patients for monotherapy, 522 for combination therapy), and as expected from a double-blind RCT, the P-values show that all relevant covariates were well-balanced between the treatment arms. The outcome variables of the primary endpoint indicator (yes/no) and mean follow-up time are very different between the two treatments (P-values <0.001), with patients in the combination treatment arm showing a smaller proportion of the primary endpoint and a longer mean follow-up time.

Figure 5.2 displays the marginal survival curves of the primary endpoint by the three estimated or assumed correlation scenarios. For plotting the survival curve by our proposed method of correlation estimation, we used the mean correlation estimate of 200 bootstrap

samples of the original ACTG 175 dataset (Table 5.4). Compared to the survival curve plotted under the estimated correlation of Kendall's tau = 0.313 (in green), the survival curve of assumed independence between the survival endpoints over-estimates (left, in blue), and that of assumed Kendall's tau = 0.8 under-estimates the marginal survival of the primary endpoint (right, in blue). An accurate estimation of the marginal survival probability over time is important in describing or predicting the patients' status if patient withdrawal from the trial or any other dependent censoring were not to occur, thus enabling us to compare the two treatment arms more objectively. In this aspect, the largely differing survival curve under an assumed correlation of 0.8 (right, in blue) shows how the estimated marginal survival probabilities can deviate to a large degree under different correlations between the survival endpoints.

Table 5.4 compares the estimated regression coefficient of the combination treatment vs. monotherapy for the time to the primary endpoint in a univariate Cox regression model among the three scenarios of an estimated or assumed correlation between the primary endpoint and other endpoints (patient withdrawal or end of study censoring). Also, in separate rows are the mean estimate of 200 bootstrap samples of the original ACTG 175 dataset and the single estimate of the original dataset. As expected, the estimated regression coefficients, SEs, and P-values of the original dataset and those of its bootstrap samples are similar.

Comparing column-wise by the three estimated or assumed correlations, the ordering of the relative sizes of the estimated regression coefficients shows that the effect estimate

is largest under assumed independence between the survival endpoints (-0.701), slightly smaller under the estimated correlation of 0.313 by the proposed method (-0.688), and smallest under an assumed correlation of 0.8 (-0.376). This is in agreement with the previous studies by Huang and Zhang (2008) and Chen (2010), while difficult to directly compare with Deresa and Van Keilegom (2021) due to their use of a linear regression model. Overall, the combination treatment is clearly superior over monotherapy in terms of the primary endpoint with protective HRs of exp(-0.688) = 0.50, exp(-0.701) = 0.49, or exp(-0.376) = 0.69 across all three correlation scenarios, which is also in agreement with the aforementioned studies.

A point of note is the difference in the estimated correlation between our study and Deresa and Van Keilegom (2021). First, we estimated a mean correlation of 0.313 between the time to primary endpoint and time to other endpoints (patient withdrawal or end of study censoring), while Deresa and Van Keilegom (2021)'s correlation estimation of 0.458 (=2/pi*arcsine(0.659)) was between the time to primary endpoint and time to patient withdrawal, treating end of study censoring as administrative independent censoring. This is clearly reasonable, while we believe that combining all other events into one dependent or competing event to estimate its correlation with the event of interest is also a reasonable approach. Second, since the ACTG study 175 was a double-blind RCT, we confirmed that all eight covariates were balanced between the two treatment arms and considered them to be adjusted for, and proceeded with a univariate Cox regression of the treatment arm variable upon the time to primary event. The previous studies differ with ours in that they

again adjusted for the aforementioned covariates in the regression model, resulting in different beta coefficient estimates compared to ours. However, the relative effect sizes by the estimated correlation, assumed independence, or assumed correlation of 0.8 are in agreement among all studies including ours. Conclusively, the better efficacy of combination therapy (vs. monotherapy) under the possibly correct estimation of correlation between the survival endpoints is not as large compared to that of assumed independence, but larger than that under an incorrectly assumed correlation of Kendall's tau = 0.8.

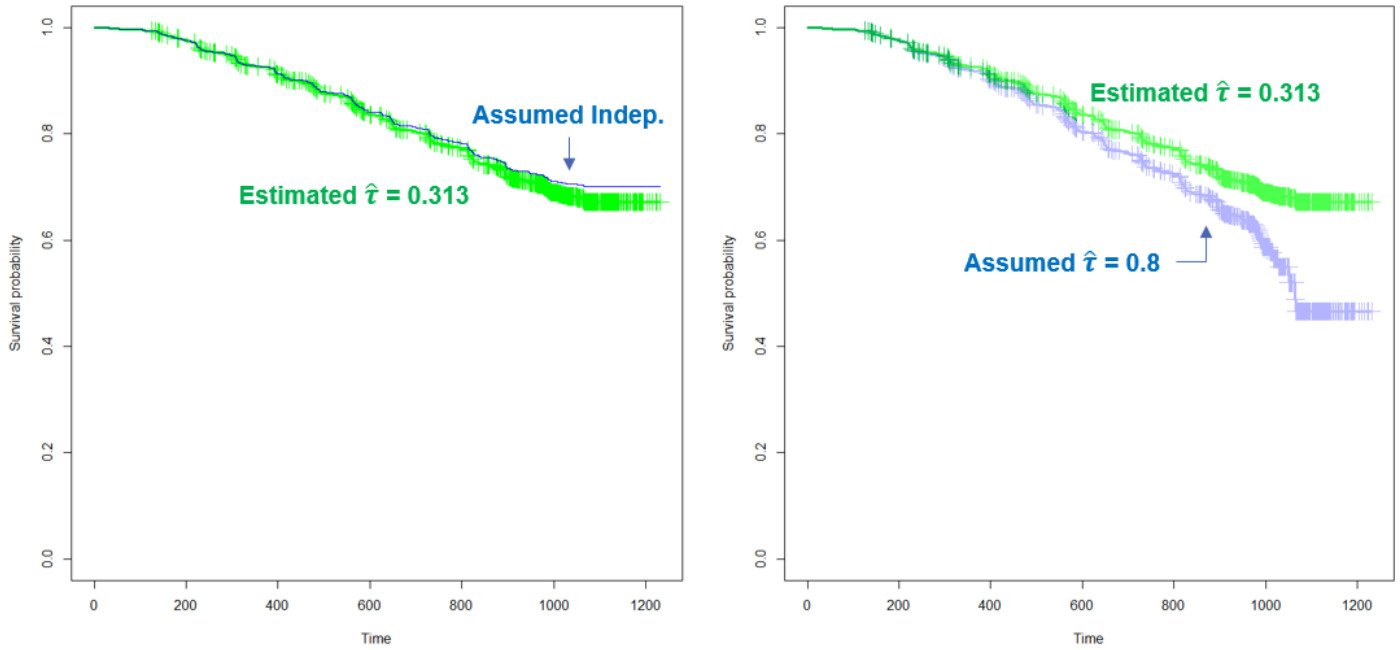
Table 5.3: Baseline characteristics of the ACTG 175 dataset study population (N=1,054) in total and by treatment arms

Variables	Total	Treatment arms		P-value [a]
		Mono	Combo	
Overall N (%)	1,054(100)	532 (50.5)	522 (49.5)	-
Age, mean (SD)	35.2 (8.77)	35.2 (8.85)	35.2 (8.70)	0.994
Gender, N (%)				0.458
Male	866 (82.2)	432 (81.2)	434 (83.1)	
Female	188 (17.8)	100 (18.8)	88 (16.9)	
Race, N (%)				0.329
White	760 (72.1)	376 (70.7)	384 (73.6)	
Other	294 (27.9)	156 (29.3)	138 (26.4)	
History of IV drug use, N (%)				0.344
Yes	136 (12.9)	63 (11.8)	73 (14.0)	
No	918 (87.1)	469 (88.2)	449 (86.0)	
Hemophilia, N (%)				0.927
Yes	85 (8.1)	42 (7.9)	43 (8.2)	
No	969 (91.9)	490 (92.1)	479 (91.8)	
Baseline CD4 count, mean (SD)	351 (122)	349 (130)	353 (114)	0.552
Prior antiretroviral therapy, N (%)				0.761
Yes	618 (58.6)	309 (58.1)	309 (59.2)	
No	436 (41.4)	223 (41.9)	213 (40.8)	
Disease symptoms, N (%)				0.530
Yes	185 (17.6)	89 (16.7)	96 (18.4)	
No	869 (82.4)	443 (83.3)	426 (81.6)	
Primary endpoint, N (%)				<0.001
Yes	284 (26.9)	181 (34.0)	103 (19.7)	
No	770 (73.1)	351 (66.0)	419 (80.3)	
Follow-up time (years), mean (SD)	858.2 (302.9)	801.2 (326.9)	916.2 (264.2)	<0.001

Abbreviations: ACTG, AIDS Clinical Trials Group Study; Combo, combination therapy of AZT plus didanosine; IV, intravenous; Mono, AZT monotherapy; SD, standard deviation

[a] P-values were calculated with ANOVA for continuous variables and the Chi-squared test for categorical variables

Figure 5.2: Marginal survival curves of the time to the primary endpoint in the ACTG 175 dataset (N=1,054) under the estimated correlation via the proposed method (green), under independence (left, in blue), and under an assumed correlation of 0.8 (right, in blue), where the copula linking the time to the primary endpoint and the time to other endpoints (withdrawal from the trial or end of study censoring) is the Clayton copula



Abbreviations: ACTG, AIDS Clinical Trials Group Study

Table 5.4: Results of correlation (dependence) estimation with the proposed method and subsequent regression coefficient estimation for a univariate Cox regression model of mono vs. combination treatment arms on the time to the primary endpoint in the ACTG 175 dataset (N=1,054), where the copula linking the time to the primary endpoint and the time to other endpoints (patient withdrawal or end of study censoring) is the Clayton copula

	Proposed method of correlation estimation			Cause-specific hazards (Independence, $\hat{\tau} = 0$)			Assumed correlation ($\hat{\tau} = 0.8$)		
Params.	Mean [a] estimate	Bootstr. [b] SE	Bootstr. P-value [c]	Mean estimate	Bootstr. SE	Bootstr. P-value	Mean estimate	Bootstr. SE	Bootstr. P-value
β_T	-0.688	0.118	<0.001	-0.701	0.116	<0.001	-0.376	0.085	<0.001
τ	0.313	0.073	<0.001	0	-	-	0.8	-	-
Params.	Orig. data[d] estimate	Orig. data SE	Orig. data P-value	Orig. data estimate	Orig. data SE	Orig. data P-value	Orig. data estimate	Orig. data SE	Orig. data P-value
β_T	-0.675	0.122	<0.001	-0.704	0.123	<0.001	-0.372	0.082	<0.001
τ	0.370	-	-	0	-	-	0.8	-	-

Abbreviations: ACTG, AIDS Clinical Trials Group Study, Bootstr., bootstrap; Orig., original; Params., parameters for estimation; SE, standard error
a The mean value of 200 regression coefficients estimated from the 200 bootstrap samples of the original ACTG 175 dataset
b The standard error (empirical standard deviation) of 200 regression coefficients estimated from the 200 bootstrap samples
c A two-sided P-value calculated from the Wald statistic = $\hat{\beta}_T$ / SE($\hat{\beta}_T$)
d The single regression coefficient, SE, and P-value estimated from the original ACTG 175 dataset

Chapter 6

Discussion and Conclusion

6.1 Points of Discussion

The current study proposed a novel method of estimating the possible correlation or dependence in bivariate competing risks survival data (T, C), where only the minimum of the time-to-events is observed and never both. We essentially addressed the "non-identifiability" dilemma in competing risks data by establishing a connection between any parametric bivariate competing risks survival data and the identifiable BVN distribution via the bivariate CLT (Figure 3.2). Simulations across a range of marginal distributions, copulas, and correlations showed our proposed method to accurately and precisely estimate the true correlation of (T, C) (Tables 4.1~4.2). Further estimation of the marginal survival curve of the time-to-event of interest T also demonstrated the importance of estimating the possible dependence between T and C by our proposed method, in contrast to the biasedness of the K-M curve and its default assumption of independent censoring (Figures 4.1~4.2). In addition, the necessity of correlation estimation between the survival outcomes was shown in a Cox PH regression model of estimating the effect of a treatment or exposure on the marginal hazard of T (Table 4.5). The beta coefficients were accurately estimated when the correlation of T and C was estimated by our proposed method and included in the model, while the coefficient estimates became biased under the "cause-specific" hazards

analysis of independent censoring or when the correlation was incorrectly assumed. To verify our proposed method's real-world applicability, we used two real-world datasets that were strongly indicative of dependence between the survival time endpoints. Here, we were able to re-confirm the proposed method's usefulness in accounting for correlated survival outcomes when estimating the effect of an exposure or treatment in disease etiology research or RCTs (Tables 5.2 and 5.4).

Compared to the previous literature on dependent competing risks or dependently censored survival data, the current study presents a novel approach to explicitly estimate the correlation, or copula dependence parameter α, between the bivariate survival time endpoints. First of all, our proposed method is a significant advancement compared to the beginning works by Zheng and Klein (1995, 1994) and Huang and Zhang (2008), where the copula parameter α and the copula's functional form were assumed to be completely known for the subsequent estimation of marginal survival or hazard functions. However, the contributions of Zheng and Klein must be acknowledged in proving the identifiability of marginal distributions of multivariate survival times once their dependence structure is known, and in developing the copula-graphic survival estimator under dependence, which reduces to the K-M survival curve under independence. Huang and Zhang also provided the framework of a sensitivity analysis under various possible correlation scenarios for a Cox regression model with dependent censoring, and their simulation approaches have been closely followed in the current study. The mainstream consensus in copula-based dependence modeling of survival data up to the works of Emura and Chen (2016, 2018)

was that the copula parameter α must be “assumed”, due to the non-identifiability of competing risks data and the likelihood function providing little information regarding the value of α (Chen, 2010; Emura & Chen, 2016; Michimae & Emura, 2022). Emura and Chen (2016) did propose a novel way of indirectly estimating α by using cross-validation of a survival prediction model. Their approach was to choose α that maximizes the cross-validated Harrell’s c-index, under the rationale that the α value resulting in the best prediction of actual survival times would be the true α. However, this approach relies on the existence of highly predictive covariates within the given dataset and was deemed infeasible after its application to our simulations (results not shown). The most recent works of Deresa and Van Keilegom (2019, 2020, 2021, 2022-1, 2022-2) and Czado and Van Keilegom (2023) deserve much acknowledgement, as their approach of BVN-distributed error terms after data transformation for the identifiability of competing risks survival data (Deresa & Van Keilegom, 2019, 2020) was the starting point of the current study. Czado and Van Keilegom (2023) expanded parametric identifiability from the BVN distribution and its Normal copula to other widely used marginal distributions and copulas, such as the Log-Normal and Weibull marginal distributions and the Archimedean copulas of Clayton, Frank, and Gumbel. Their theorems essentially state that bivariate competing risks or dependently censored survival data from the parametric marginal distributions and copulas noted above are “identifiable” from the usual likelihood construction and subsequent MLE. However, we empirically verified the instability of MLE in the case of Weibull marginal distributions linked via Normal, Clayton, Frank, or Gumbel copulas (Table 4.4), and the

disagreement between our results and the statements of Czado and Van Keilegom (2023) seems to require further study.

We also propose some other topics for future study. First, our numerical estimation algorithm may be used to further develop a test of dependence in bivariate competing risks or dependently censored survival data. The test may be formulated in the lines of a null hypothesis stating independence of T and C, where the estimated correlation not being statistically significantly different from zero would imply independence only for a Normal (Gaussian) copula (Hogg, McKean, & Craig, 2013). For resampled data statistics of empirical CIs, the proportion of null hypothesis rejection would be used to calculate the P-value of the test. Second, the phenomenon of correlation shrinkage or enlargement in the bootstrap sample means of Archimedean copulas should be further investigated. Although no direct search results were found, the previous works of Fermanian et al. (2004), Genest and Segers (2010), and Segers (2012) regarding the asymptotics or convergence of copula processes may provide theoretical groundwork in explaining this phenomenon. Third, as the proposed method's global search with simulated annealing required comparatively long runtimes, other potentially faster algorithms such as 'differential evolution' which is known to work well for continuous numerical optimization (Cortez, 2021) may be worth exploring.

6.2 Study Conclusion

In conclusion, the current study proposed a novel method to explicitly estimate the

correlation (dependence) in bivariate competing risks survival data, subsequently enabling an unbiased estimation of the marginal survival or hazard functions of the event of interest when the independent censoring assumption does not hold. Simulations showed that the proposed method works well over various marginal distributions, copulas, and sizes of the correlation. Our study provides a potential contribution to the existing literature in that the proposed method is applicable to any parametric bivariate competing risks data, requires no covariate information to estimate the correlation, and shows accurate and precise results where the conventional MLE fails to do so. We expect the current study to have further applications in biomedical time-to-event analyses where dependence between the survival endpoints exist more often than not, especially in disease etiology research and RCTs of drug efficacy.

Bibliography

Abbring JH, van den Berg GJ. The Identifiability of the Mixed Proportional Hazards Competing Risks Model. Journal of the Royal Statistical Society. Series B (Statistical Methodology). 2003;65(3):701-710.

Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: possibilities and pitfalls. Int J Epidemiol. 2012;41(3):861-870.

Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation. 2016;133(6):601-609.

Baron F, Labopin M, Tischer J, et al. GVHD occurrence does not reduce AML relapse following PTCy-based haploidentical transplantation: a study from the ALWP of the EBMT. J Hematol Oncol. 2023;16(1):10.

Chen Y-H. Semiparametric Marginal Regression Analysis for Dependent Competing Risks Under an Assumed Copula. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2010;72(2):235-251.

Cortez P. Modern Optimization with R. Springer Cham; 2021.

Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;34(2):187-202.

Cox DR. The Analysis of Exponentially Distributed Life-Times with Two Types of Failure. 1959;21(2):411-421.

Crowder M. Identifiability Crises in Competing Risks. International Statistical Review. 1994;62(3):379-391.

Czado C, Van Keilegom I. Dependent censoring based on parametric copulas. Biometrika. 2022.

Dafni U. Landmark analysis at the 25-year landmark point. Circ Cardiovasc Qual Outcomes. 2011;4(3):363-371.

Deresa NW, Keilegom IV. Copula Based Cox Proportional Hazards Models for Dependent Censoring. Journal of the American Statistical Association. 2023:1-11.

Deresa NW, Van Keilegom I, Antonio K. Copula-based inference for bivariate survival data with left truncation and dependent censoring. Insurance: Mathematics and Economics. 2022;107:1-21.

Deresa NW, Van Keilegom I. A multivariate normal regression model for survival data subject to different types of dependent censoring. Computational Statistics & Data Analysis. 2020;144:106879.

Deresa NW, Van Keilegom I. Flexible parametric model for survival data subject to dependent censoring. Biom J. 2020;62(1):136-156.

Deresa NW, Van Keilegom I. On semiparametric modelling, estimation and inference for survival data subject to dependent censoring. Biometrika. 2020;108(4):965-979.

Efron B. The Two-Sample Problem with Censored Data. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967;Vol.4:831-852.

Emura T, Chen YH. Analysis of Survival Data with Dependent Censoring: Copula-Based Approaches. Springer Singapore; 2018.

Emura T, Chen Y-H. Gene selection for survival data under dependent censoring: A copula-based approach. Statistical Methods in Medical Research. 2014;25(6):2840-2857.

Emura T, Shih JH, Ha ID, and Wilke RA. Comparison of the marginal hazard model and the sub-distribution hazard model for competing risks under an assumed copula. Statistical Methods in Medical Research. 2019 Dec;29(8):2307-27.

Fermanian JD, Radulović D, Wegkamp M. Weak Convergence of Empirical Copula Processes. Bernoulli. 2004;10(5):847-860.

Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association. 1999 Jun;94(446):496-509.

Genest C, Segers J. On the covariance of the asymptotic empirical copula process. Journal of Multivariate Analysis. 2010;101(8):1837-1845.

Givens GH, Hoeting JA. Computational Statistics. John Wiley & Sons; 2005.

Gray RJ. A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. The Annals of Statistics. 1988 Sep;16(3):1141-54.

Gumbel EJ. Bivariate Exponential Distributions. Journal of the American Statistical Association. 1960;55(292):698-707.

HECKMAN JJ, HONORÉ BE. The identifiability of the competing risks model. Biometrika. 1989;76(2):325-330.

Hofert M, Kojadinovic I, Mächler M, Yan J. Elements of Copula Modeling with R. Springer Cham; 2018.

Hogg RV, McKean JW, Craig AT. Introduction to Mathematical Statistics, 7th Edition. Pearson Education; 2013.

Hsu JY, Roy JA, Xie D, et al. Statistical Methods for Cohort Studies of CKD: Survival Analysis in the Setting of Competing Risks. Clinical Journal of the American Society of Nephrology. 2017;12(7):1181-1189.

Huang X, Zhang N. Regression survival analysis with an assumed copula for dependent censoring: a sensitivity analysis approach. Biometrics. 2008;64(4):1090-1099.

Katsahian S, Porcher R, Mary JY, Chevret S. The graft-versus-leukaemia effect after allogeneic bone-marrow transplantation: assessment through competing risks approaches. Stat Med. 2004;23(24):3851-3863.

Kirkpatrick S, Gelatt CD Jr., Vecchi MP. Optimization by Simulated Annealing. Science. 1983;220(4598):671-680.

Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data, 2nd Edition. Springer-Verlag; 2003.

Klein JP. Competing risks. WIREs Computational Statistics. 2010;2(3):333-339.

Lau B, Cole SR, Gange SJ. Competing Risk Regression Models for Epidemiologic Data. American Journal of Epidemiology. 2009;170(2):244-256.

López-Ibáñez M, Dubois-Lacoste J, Pérez Cáceres L, Birattari M, Stützle T. The irace package: Iterated racing for automatic algorithm configuration. Operations Research Perspectives. 2016;3:43-58.

Meller M, Beyersmann J, Rufibach K. Joint modeling of progression-free and overall survival and computation of correlation measures. Statistics in Medicine. 2019;38(22):4270-4289.

Michimae H, Emura T. Likelihood Inference for Copula Models Based on Left-Truncated and Competing Risks Data from Field Studies. Mathematics. 2022;10(13):2163.

Moeschberger ML, Klein JP. Statistical methods for dependent competing risks. Lifetime Data Anal. 1995;1(2):195-204.

Moeschberger ML. Life Tests Under Dependent Competing Causes of Failure. Technometrics. 1974;16(1):39-47.

Nash JC. On Best Practice Optimization Methods in R. J Stat Softw. 2014;60(2):1 - 14.

Nelsen RB. An Introduction to Copulas, Second Edition. Springer New York, NY; 2006.

Nordlander A, Mattsson J, Ringden O, et al. Graft-versus-host disease is associated with a lower relapse incidence after hematopoietic stem cell transplantation in patients with acute lymphoblastic leukemia. Biol Blood Marrow Transplant. 2004;10(3):195-203.

Pintilie M. Competing Risks: A Practical Perspective. Wiley; 2006.

Prentice RL, Kalbfleisch JD, Peterson AV, Jr., Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. Biometrics. 1978;34(4):541-554.

Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med. 2007;26(11):2389-2430.

Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics. 2000;56(3):779-788.

Schwarz M, Jongbloed G, Van Keilegom I. On the identifiability of copulas in bivariate competing risks models. The Canadian Journal of Statistics. 2013;41(2):291-303.

Segers J. Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. Bernoulli. 2012;18(3):764-782.

Sklar A. Fonctions de répartition à n dimensions et leurs marges. Publications de l'Institut de Statistique de L'Université de Paris. 1959;Vol.8:229-231.

Sorrell L, Wei Y, Wojtys M, Rowe P. Estimating the correlation between semi-competing risk survival endpoints. Biom J. 2022;64(1):131-145.

Tsiatis A. A nonidentifiability aspect of the problem of competing risks. Proc Natl Acad Sci U S A. 1975;72(1):20-22.

van Houwelingen H, Putter H. Dynamic Prediction in Clinical Survival Analysis. CRC Press; 2012.

Weber EM, Titman AC. Quantifying the association between progression-free survival and overall survival in oncology trials using Kendall's tau. Stat Med. 2019;38(5):703-719.

Willems S, Schat A, van Noorden MS, Fiocco M. Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator. Stat Methods Med Res. 2018;27(2):323-335.

Wolbers M, Koller MT, Stel VS, et al. Competing risks analyses: objectives and approaches. European Heart Journal. 2014;35(42):2936-2941.

Xiang Y, Gubian S, Martin F. Generalized Simulated Annealing. Computational Optimization in Engineering - Paradigms and Applications. IntechOpen; 2017.

Xiang Y, Gubian S, Suomela B, Hoeng J, Generalized Simulated Annealing for Global Optimization: The GenSA Package. The R Journal. 2013;5(1).

Zhang HS, Yang Y, Lee S, Park S, Nam CM, Jee SH. Metformin use is not associated with colorectal cancer incidence in type-2 diabetes patients: evidence from methods that avoid immortal time bias. Int J Colorectal Dis. 2022;37(8):1827-1834.

Zheng M, Klein JP. A self-consistent estimator of marginal survival functions based on dependent competing risk data and an assumed copula. Communications in Statistics - Theory and Methods. 1994;23(8):2299-2311.

ZHENG M, KLEIN JP. Estimates of marginal survival for dependent competing risks based on an assumed copula. Biometrika. 1995;82(1):127-138.

Appendix

Table A1. Correlation shrinkage or enlargement in bootstrap sample means of bivariate competing risks data (T, C) [a] with Exponential marginal distributions, where T and C are linked via parametric copulas such as the Normal, Clayton, Frank, and Gumbel copulas

Copula	True Kendall's tau	Simulated data avg. correl.	Bootstrap sample mean avg. correl.	95% CI of bootstrap sample mean correl.	Shrinkage or Enlargement
Normal	0.3	0.299	0.268	(0.207, 0.328)	-10.4%
	0.5	0.500	0.464	(0.418, 0.510)	-7.2%
	0.8	0.800	0.781	(0.755, 0.806)	-2.4%
Clayton	0.3	0.301	0.407	(0.343, 0.469)	35.2%
	0.5	0.500	0.625	(0.582, 0.667)	25.0%
	0.8	0.800	0.873	(0.858, 0.887)	9.1%
Frank	0.3	0.301	0.220	(0.094, 0.281)	-26.9%
	0.5	0.498	0.368	(0.100, 0.438)	-26.1%
	0.8	0.799	0.628	(0.171, 0.693)	-21.4%
Gumbel	0.3	0.299	0.212	(0.140, 0.279)	-29.1%
	0.5	0.500	0.381	(0.149, 0.453)	-23.8%
	0.8	0.800	0.713	(0.676, 0.750)	-10.9%

Abbreviations: avg, average; CI, confidence interval; correl, correlation
a 200 datasets, where each consists of 1000 bootstrap sample mean observations from the original (T, C) dataset with a sample size of 1,000.

Table A2. Correlation shrinkage or enlargement in bootstrap sample means of bivariate competing risks data (T, C) [a] with Log-Normal marginal distributions, where T and C are linked via parametric copulas such as the Normal, Clayton, Frank, and Gumbel copulas

Copula	True Kendall's tau	Simulated data avg. correl.	Bootstrap sample mean avg. correl.	95% CI of bootstrap sample mean correl.	Shrinkage or Enlargement
Normal	0.3	0.299	0.240	(0.180, 0.304)	-19.7%
	0.5	0.500	0.398	(0.334, 0.450)	-20.4%
	0.8	0.800	0.608	(0.532, 0.662)	-24.0%
Clayton	0.3	0.301	0.357	(0.303, 0.412)	18.6%
	0.5	0.500	0.518	(0.461, 0.560)	3.6%
	0.8	0.800	0.642	(0.525, 0.692)	-19.8%
Frank	0.3	0.301	0.190	(0.069, 0.257)	-36.9%
	0.5	0.498	0.310	(0.085, 0.381)	-37.8%
	0.8	0.799	0.501	(0.129, 0.587)	-37.3%
Gumbel	0.3	0.299	0.191	(0.119, 0.253)	-36.1%
	0.5	0.500	0.333	(0.157, 0.409)	-33.4%
	0.8	0.800	0.571	(0.491, 0.629)	-28.6%

Abbreviations: avg, average; CI, confidence interval; correl, correlation
a 200 datasets, where each consists of 1000 bootstrap sample mean observations from the original (T, C) dataset with a sample size of 1,000.

국문요약

이변량 경쟁위험 생존자료에서 상관성과 주변부 분포 추정을 위한 통합된 모수적 추정 방법 제안

둘 중 먼저 발생한 사건까지의 시간만 알 수 있는 이변량 경쟁위험 생존자료에서 두개 사건 발생 시간 간의 상관성은 식별 불가능한 것으로 알려져 있다. 두개의 사건 발생 시간 간의 상관성 또는 영이 아닌 상관계수가 존재한다면, 독립적 중도절단 가정 하의 특정 원인별 위험 (cause-specific hazards) 분석이나 잘못 가정된 상관계수 하의 분석은 편향을 야기한다. 이러한 상관성이 존재할 때 가장 중요하게 추정되어야 할 모수는 사건 발생 시간 간의 상관계수로 볼 수 있다. 이 경우에 최대우도추정법은 추정치가 편향되고 분산 또한 큰 것으로 알려져 있고, 정확한 상관계수 추정을 위한 실용적인 방법은 아직 없는 실정이다.

이변량 정규분포를 따르는 이변량 경쟁위험 자료에서는 원래의 분포 모수가 식별 가능함에 착안하여, 본 연구는 이변량 중심극한정리를 연결고리로 주어진 이변량 경쟁위험 자료와 식별 가능한 이변량 정규분포를 연결하는 통합된 모수적 접근법을 제안하였다. 즉, 같은 표본 평균 정보

를 갖는 이변량 정규분포의 상관계수 모수가 주어진 자료에서 추정하고자 하는 상관계수이며 이것을 반복적인 수치 알고리즘으로 추정 가능함을 보였다. 상관계수의 정확한 추정은 이후의 주변부 생존 또는 위험 함수의 비편향적 추정을 또한 가능케 한다.

본 연구는 기존 연구들 대비 광범위한 모수적 이변량 경쟁위험 자료에 적용이 가능하고, 상관계수 추정을 위한 공변량 정보가 필요 없으며, 최대우도추정법이 경쟁위험 자료의 상관계수 추정에 사용될 수 없음을 보완하는 측면에서 잠재적 기여를 할 수 있을 것으로 보인다. 아울러 보건의료적 관점에서, 사건 발생까지의 시간 간에 상관성이 존재할 수 있는 질병의 인과성 연구 또는 신약 평가 임상시험 등에 추가적인 응용 사례가 있기를 기대한다.

핵심되는 말: 경쟁위험 생존분석, 상관계수, 종속성, 식별성 문제, 이변량 중심극한정리