RELAXING THE INDEPENDENCE ASSUMPTION IN RELATIVE SURVIVAL ANALYSIS

Reuben Adatorwovor

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2020

Approved by:

Jason P. Fine

Feng-Chang Lin

John S. Preisser Jr

Chirayath M. Suchindran

David B. Richardson

ABSTRACT

Reuben Adatorwovor : Relaxing the Independence Assumption in Relative Survival Analysis
(Under the direction of Jason P. Fine)

Quantifying credible cancer survival in competing risk population-based studies is generally done by disease-specific survival analysis when reliable cause of death information is available. Relative survival analysis may be used to estimate disease-specific survival when cause of death is missing and or subject to misspecification and not reliable for practical usage. This method is popular for population-based cancer survival studies using registry data and does not require cause of death information. The standard estimator under the independence assumption is the ratio of all-cause survival in the cancer cohort group to the known expected survival from a healthy reference population. Disease-specific death competes with other causes of mortality, potentially creating dependence among the causes of death. The standard ratio estimate is only valid when death from disease and death from other causes are independent. To relax the independence assumption, we formulate dependence using a copula-based model. Likelihood-based, nonparametric and parametric regression methods are implemented to fit a parametric, a nonparametric and a regression model to the distribution of disease-specific death respectively without the need for cause of death information. We assumed that the copula is known and the distribution of other cause of mortality is derived from the reference population. Since the dependence structure for disease related and other-cause mortality is nonidentifiable and unverifiable from the observed data, we propose a sensitivity analysis, where the analysis is conducted across a range of assumed dependence structures. We demonstrate the practical utility of our method through simulation studies and an application to French breast cancer data.

This work has being dedicated to my family especially Dayana, Solomon, Shalom, and Shaina.

ACKNOWLEDGEMENTS

I am thankful to Dr. Jason P. Fine for his insight and helpful suggestions and the rest of my advising committee members for reading of the original version of this paper and providing very useful comments that improved the presentation of this manuscript. I am also grateful to the UNC CH graduate school for providing funds for me through the dissertation completion fellowship which enabled me to have undivided attention during my last couple of months of dissertation work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AFT	Accelerated Failure Time
$b-Gumbel$	Bivariate Gumbel
$C-Copula$	Clayton copula
$S_{T_1}(t)$	Disease-specific survival
EMP	Empirical
EST	Estimate
$G-Gumbel$	Gumbel copula
$ModB$	Model Based
$Mod-B$	Model Based
T_j	Latent Failure time j
K-M	Kaplan-Meier
SE	Standard Error
$S_R(t)$	Relative Survival

CHAPTER 1: Introduction

1.1 Introduction

Cancer patient survival in competing risk settings is a fundamental problem for cancer researchers and physicians. Improvement in cancer therapeutics comes with understanding of cancer prognostic measures which is seemingly simple but often lacking and or confusing. Additionally, understanding the underlying underpinnings of the current estimators for these prognostic measures is complicated especially without complete and accurate cause of event information. Appropriate estimators under reasonable assumptions are nonexistent for disease-specific survival estimation. Quantifying cancer survival in population-based cancer registries not only provides information to patients and their families to understand prognosis and make decisions on the type of treatment to seek but also provides useful guidance to their physicians in making decisions about the type of treatment regimen to deploy. This dissertation delves into the constraints of the existing estimators for disease-specific survival analysis. In particular, three estimators were proposed to mitigate against unsubstantiated assumptions for the current estimators that assume that time to disease-specific mortality is independent of time to competing mortality for disease-specific survival when cause of event is either missing or subject to misspecification.

In Chapter 2, we propose a parametric method for modelling disease-specific survival for competing risk registry data. We formulate the dependence between the latent failure times distributions for death from disease and death from competing causes using copula models. The copula model captures the nonlinear scale invariant dependence inherent in competing risk data. It takes as input the marginal distribution of the minimum event time where the distribution of other cause mortality is assumed known and extracted from the reference population with the usual assumption that disease-specific death is negligible in this reference population. Due to the nonlinearity of the mortality trends observed in the cancer registry data, two Archimedean copulas:

the Gumbel and the Clayton copulas have been discussed and implemented. The usual maximum likelihood estimation procedure was implemented.

Chapter 3 relaxes the parametric assumption introduced earlier in chapter 2. Instead of relying on traditional endpoints such as the disease-specific hazard and disease-specific cumulative incidence functions for disease-specific survival estimation, we employ a function of the Kaplan-Meier estimator. First, we formulate the dependence by using the Gumbel copula. The all-cause mortality is estimated using the Kaplan-Meier estimator which is a function of the disease-specific survival and the excess survival. Next, the nonlinear function is inverted and solved for disease-specific survival using a variant of Newton-Raphson algorithm. The usual regularity conditions hold.

We discuss the risk factors that influence cancer survival in chapter 4. These factors may include but are not limited to age, sex, calendar period using the competing risk dependence regression method. Assuming a similar formulation in chapter 2 above, we incorporate the covariates using the Accelerated Failure Time (AFT) model. In this case, a transformation of the latent failure time is implemented to obtain an Extreme Value Distribution (EVD) for the minimum of the event times. As usual, the distribution of the competing mortality iss derived from the background reference population and together with the distribution of the latent failure times modelled using the Gumbel copula. Likelihood inference and interpretation is proposed.

Extensive simulations were implemented to assess the performance of our methodologies. Due to the identifiability constraints and the unverifiable nature of the dependence between the latent failure times in the observed registry data, a sensitivity analysis was proposed where we estimated disease-specific survival across a spectrum of dependence structures. We demonstrated the utility of our methods through an application to French breast cancer data obtained from Institut Curie breast cancer database, France.

Chapter 5 concludes the dissertation. In that we proposed three novel methods for estimating relative survival for cancer registries. These methods are the first step in incorporating dependence in relative survival analysis. The key measures are the so called ”net survival” under the independence assumption and ”crude survival” or ”crude probability of death” under the dependence assumption. These prognostic measures for cancer-specific survival are not only of great interest to patients and their love ones for end of life decision making, or for clinicians for clinical decision making, or for researchers in understanding therapeutics, but also for policy makers in making decisions that impact us all.

CHAPTER 2: Relaxing the Independence Assumption in Relative Survival Analysis: A Parametric Approah

2.1 Introduction

Cancer patients including breast, prostate, endometrial and thyroid cancer are at higher risk of dying from heart disease and stroke than the general population. As the number of cancer survivors increases, so is the rate of cardiovascular deaths (Sturgeon et al., 2019). Such medical research frequently yields multiple event times which may consist of a terminal and or a non-terminal event, including landmarks of the disease process. The practical concern for physicians is patient survival, suggesting an analysis based on the distribution of these event times or the disease-specific hazard and or cumulative incidence function. There is often scientific interest in understanding disease-specific mortality in the absence of failure types other than the disease of interest, a quantity which is sometimes controversial but meaningful to many practitioners or researchers. Other researchers prefer the latter quantity in understanding disease-specific mortality in the presence of other competing causes. Understanding these quantities helps inform researchers in the analysis of the biological efficacy of treatment regimen rendered to patients to assess patient survival.

Survival probability is an important measure not only for clinicians in determining prognosis and treatment regimen but also for patients and their families for decision making. With improvement in medical treatment and long follow-up in population-based disease registries, there is a potential for lost to follow-up during which patients may either experience disease-specific death or death from non-disease related causes (Brinkhof et al., 2010). In such competing risk settings where one death type precludes the occurrence of other types, standard methodology assumes that cause of death is known (Gichangi and Vach, 2005).

In the analysis of competing risks events from registry data, accurate documentation of death is essential (Percy et al., 1981; Welch and Black, 2002; Mieno et al., 2016). A challenge is that documentation either may not be available, or may be incomplete or incorrect for cause of death, resulting in problems distinguishing disease and non-disease related mortality. The issue is pronounced in Europe, where comparison of disease-specific survival across countries is of interest. The World Health Organization (Organization et al., 1977) defines cause of death as ”the disease or injury which initiated the train of morbid events leading directly to death”. However, population-based disease registries may not be harmonized across countries, leading to imprecise cause of death definitions and different levels of documentation of cause of death information. Often, the underlying cause of death may be unclear as hospital coding of cancer death may not agree with the death certificate coding. As an example, (Welch and Black, 2002) reported that 41% of deaths that occurred (within one month diagnosis and cancer directed surgery) were not attributable to the coded cancer in the registry. When reliable cause of death information is available, it is often located in separate databases, which may be costly to obtain and difficult to link with registry data.

Suppose that $T = \min\{T_k : \ k = 1, 2, 3, \cdots, K\}$ is the potentially observable failure time and $\varepsilon = \{k : T = T_k\}$ the failure type where $T_1, \cdots, T_K$, with $K \in \mathbf{N}$ are the latent failure times associated with the K failure types. In registry data, $K = 2$ and $\varepsilon = 1$ implies death from cancer and $\varepsilon = 2$ implies death from other competing causes. Standard methods for independently right censored survival data without competing risks cannot generally be used to make inference about disease-specific survival. Under dependent competing risks, where T_1 and T_2 are dependent, the Kaplan-Meier (Kaplan and Meier, 1958) curve estimates a function of the cause-specific hazard function, defined in Section 2.2. The logrank test (Bland and Altman, 2004) assesses group differences between the cause-specific hazard function, while the standard proportional hazards model (Cox, 1972) formulates the effects of covariates on the cause-specific hazard function. The cumulative incidence function, defined in Section 2.2, gives disease-specific survival in the presence of competing events. This quantity has been widely adopted in applications, with the Aalen-Johanson estimator (Aalen and Johansen, 1978), Gray’s test (Gray et al., 1988), and the Fine-Gray model (Fine and Gray, 1999), providing analogs to the Kaplan-Meier curve, the logrank test, and the proportional hazards model for the cumulative incidence function. Without cause of death information, these methods are not applicable.

To address disease-specific survival without cause of death information, relative survival methods have been proposed. Relative survival, $S_R(t)$ is the ratio of the observed survival rate in a group of cancer patients, during a specified period, to the expected survival rate in a healthy reference population (Ederer, 1961). Mathematically,

$$S_R(t) = \frac{S_O(t)}{S_P(t)} \tag{2.1}$$

where at time t, $S_O(t)$ is the survival probability for an individual in the registry and $S_P(t)$ is the expected survival from mortality tables. Existing literature has focused exclusively on the estimation of $S_R(t)$ under the independence assumption, $T_1 \perp T_2$. Under independence, $S_O(t) = S_{T_1}(t) \cdot S_{T_2}(t),\ S_P(t) = S_{T_2}(t)$ which implies $S_R(t) = S_{T_1}(t)$ where $S_{T_1}(t)$ and $S_{T_2}(t)$ are the survival probabilities corresponding to T_1 and T_2 respectively. The relationship (2.1) can be rewritten in terms of hazard functions as $\lambda_O(t) = \lambda_E(t) + \lambda_P(t)$ (Cronin and Feuer, 2000), where $\lambda_O(t)$ is the hazard in the disease registry, $\lambda_E(t)$ is the so called excess hazard among the cancer cohort, and $\lambda_P(t)$ is the hazard from mortality tables. Under independence, $\lambda_E(t) = \lambda_{T_1}(t)$ and $\lambda_P(t) = \lambda_{T_2}(t)$, where $\lambda_{T_j}(t) = \frac{-dlogS_{T_j}(t)}{dt}, j = 1, 2$, are the net hazard functions for cancer and other cause mortality. The disease-specific survival probability $S_{T_1}(t)$ under the independence assumption is the target of relative survival analysis and corresponds to a hypothetical population in which death from competing causes does not exist. It differs from the cumulative incidence function which is commonly used to quantify disease-specific survival in analyses with cause of death information. $S_R(t)$ has an excess hazard (Suissa, 1999) interpretation and is no longer a survival probability when formulated as in 2.1.

Relative survival based on independence methods was pioneered by Berkson and Gage (1950), and Ederer (1961) for nonparametric estimation of $S_{T_1}(t)$. A variant of this method was proposed by Hakulinen (1982) to address the bias due to heterogeneity of patient withdrawal within subgroups. Perme et al. (2012) demonstrated that these classical methods may be biased under certain censoring patterns. For example, in population comparisons, such bias may arise from unmeasured covariates affecting the cancer cohort group and the reference population from which rates of expected mortality are drawn. Rebolj Kodre and Pohar Perme (2013) studied biases associated with censoring and age distribution (at the time of cancer diagnosis) and proposed weighting corrections. Nixon et al. (1994) documented that event times and censoring times are dependent on the age of the patients in a cancer study. Stratified methods Sasieni and Brentnall (2017) based on age standardization of relative survival ratios may reduce such biases. Hakulinen et al. (2011) and Perme et al. (2012)

developed alternative estimators valid under weaker assumption. However, the above estimation methods for $S_R(t)$ all require independence of death from cancer and death from competing causes (Hakulinen et al., 2011; Perme et al., 2012).

To relax the independence assumption (de Lacerda et al., 2019; Makkar et al., 2018), we formulate the dependence between the latent failure times distributions for death from disease and death from competing causes using copula models (Deheuvels, 1978). The copula function generates a joint distribution for the two event times, taking as input their marginal distributions. Copulas allow a broad range of dependence structures and have been employed widely in survival analysis, including bivariate event times (Oakes, 1982), competing risks with known cause of failure (Heckman and Honoré, 1989), and semi-competing risks where one event time censors the other but not vice versa (Fine et al., 2001). We employ such models with competing risks data from disease registries where cause of death information is either not reliable or not available. Because the joint distribution of the latent failure times is nonparametrically nonidentifiable (Tsiatis, 1975), we treat the copula function as known. The marginal distribution of the time to disease-specific death is modelled parametrically with the distribution of death from other causes drawn from the reference population. Likelihood-based inference is proposed. Because the joint distribution is unidentifiable nonparametrically and unverifiable from the observed registry data, a sensitivity analysis is suggested in which disease-specific survival is estimated across a range of rich dependence structures, specified via the copula function. To our knowledge, this is the first attempt in accommodating dependence in relative survival analysis.

The rest of this paper proceeds as follows. In section 2.2, we present the data and copula model formulation for competing risks data. Section 2.3 describes the likelihood estimation and inference procedure without cause of death information, as well as the proposed sensitivity analysis. In section 2.4, we present the numerical illustrations including simulation results and application to French breast cancer data. Section 2.5 discusses and concludes the paper.

2.2 Data and Model

We begin by defining traditional endpoints for competing risk data with known cause of death. The cause-specific hazard, $\lambda_k(t)$ is the instantaneous failure rate for occurrence of event $\varepsilon = k$ at time t (Prentice et al., 1978),

$$\lambda_k(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t, K = k | T > t)}{\delta t} \tag{2.2}$$

and the cumulative incidence function $C_k(t)$ is the proportion of patients who died from cause k by time t in the presence of patients who might die from other causes. The disease-specific failure probability can be expressed as $C_k(t) = P(T \leq t : \ \varepsilon = k) = \int_0^t \lambda_k(s) \cdot S(s) ds$ where $S(t) = P(T > t)$ is the overall survival probability. Standard competing risks methods with known cause of failure focus on estimation of $\lambda_k(t)$ and $C_k(t)$.

Without cause of death information, the registry data is simply time to death from any cause, T, which may be right censored by lost to follow up. Let C be the time to right censoring, with the common assumption being that T and C are independent. The observed data consist of $X_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$, where T_i and C_i are the failure and censoring times on individual $i = 1, 2, 3, \cdots, n$. Relative survival methods employing such data do not focus on the traditional competing risks endpoints $\lambda_k(t)$ and $C_k(t)$ but rather on the latent failure time distributions with the corresponding survival functions $S_{T_1}(t)$ and $S_{T_2}(t)$.

To capture the dependence between T_1 and T_2, we employ copula models, which completely describe the dependence structure and provide scale invariant measures of association (Venter, 2002; Müller, 1996; Bäuerle and Müller, 1998; Denuit et al., 1999). Suppose ψ is a function defined such that $\psi : [0, 1] \to [0, +\infty]$ with independent marginal distributions, $u_j = P(T_j \leq t_j) = F_{T_j}(t_j) = 1 - S_{T_j}(t_j) \ \forall j \in (1, 2)$. Then, the copula model for distributions of T_1 and T_2 (Cherubini et al., 2004; Joe, 1997; McNeil et al., 2009) is:

$$C(u_1, u_2) = P(T_1 \leq t_1, T_2 \leq t_2) = \psi\left(\psi^{-1}(u_1) + \psi^{-1}(u_2)\right) = F_{T_1,T_2}(t_1, t_2)$$

where ψ^{-1} is the inverse of ψ and ψ satisfies the Laplace-Stiltjes transform and Bernstein et al. (1929) theorem. McNeil et al. (2009) showed that the generator function ψ is completely monotone for non-negative random variables with $\psi(0) = 1, \ \psi'(\cdot) < 0$ and $\psi''(\cdot) < 0$.

The most widely used scale invariant measures of association to characterize dependence are Spearman's rho (ρ_S) and Kendall's tau (τ_k) correlation coefficients. The connection between the latter and the copula generator function was shown by Genest and MacKay (1986) as:

$$\tau_k = 1 + 4\int_0^1 \frac{\psi^{-1}(u)}{\psi^{-1}(u)'}du = 1 - 4\int_0^\infty u(\psi(u))^2 du$$

with $\psi^{-1'}$ being the derivative of ψ^{-1}. While in theory, any copula may be used to link the marginal distributions of T_1 and T_2, in this paper, we focus on two popular Archimedean copulas, indexed by a single dependence parameter θ having simple interpretations. The Gumbel copula:

$$C(u_1, u_2) = \exp\left[-\{(-log(u_1))^\theta + (-log(u_2))^\theta\}^{\frac{1}{\theta}}\right] \tag{2.3}$$

with $\theta \in (1, +\infty)$ and the Clayton copula:

$$C(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-\frac{1}{\theta}} \tag{2.4}$$

with $\theta \in (0, +\infty)$. A special case of product copula model: $C(u_1, u_2) = u_1 \cdot u_2$ is obtained when $\theta = 1$ and when $\theta \to 0$ for Gumbel and Clayton copulas respectively, which gives independence of T_1 and T_2. When $\theta > 0$, the Clayton copula is bounded by: $C(u_1, u_2) \leq \theta(1 - u_1 - u_2) + (1 + \theta)u_1u_2$. As dependence increases, that is $\theta \to +\infty$, the Clayton copula approximates the Fréchet-Hoeffding (Fréchet, 1951; Hoeffding, 1940) upper bound, giving perfect positive dependence.

2.3 Likelihood Estimation and Inference

We first formulate our model without covariates for the potentially dependent latent failure times T_1 and T_2. The survival function for all-cause mortality time, $T = \min(T_1, T_2)$ at time t, is:

$$\begin{aligned} S_T(t) &= S_{T_1}(t) + S_{T_2}(t) - 1 + F_{T_1,T_2}(t, t) \\ &= 1 - F_{T_1}(t) - F_{T_2}(t) + F_{T_1,T_2}(t, t) \end{aligned} \tag{2.5}$$

with the corresponding density function of T equalling

$$f_T(t) = f_{T_1}(t) + f_{T_2}(t) - f_{T_1,T_2}(t, t) \tag{2.6}$$

where $f_{T_j}(t) = \frac{dF_{T_j}(t)}{dt}$, and $f_{T_1,T_2}(t) = \frac{dF_{T_1,T_2}(t,t)}{dt}$.

If censoring of T by C is noninformative, then the likelihood contribution for individual i is:

$$L_i = f_{X_i,\Delta_i}(X_i, \delta_i) = [f_T(X_i)]^{\delta_i}[S_T(X_i)]^{1-\delta_i} \tag{2.7}$$

From equation (2.7), the full log-likelihood function based on independent observations is:

$$\begin{aligned} l(\mathbf{X}, \Delta) &= \sum_{i=1}^{n} (\delta_i * \log f_T(X_i) + (1-\delta_i) * \log S_T(X_i)) \\ &= \sum_{i=1}^{n} \delta_i * \log\left[f_{T_1}(X_i) + f_{T_2}(X_i) - f_{T_1,T_2}(X_i, X_i)\right] \\ &\quad + \sum_{i=1}^{n} (1-\delta_i) * \log\left[S_{T_1}(X_i) + S_{T_2}(X_i) - 1 + F_{T_1,T_2}(X_i, X_i)\right] \end{aligned} \tag{2.8}$$

where $(\mathbf{X}, \Delta) = (X_i, \Delta_i, i = 1, 2, 3, \cdots, n)$. We specify a parametric model for $F_{T_1}(t)$, with parameter of interest η.

The general form of the probability density function of T_1 at time t is $f_{T_1}(t|\eta)$ with survival probability $S_{T_1}(t|\eta) = 1 - F_{T_1}(t|\eta) = \int_t^\infty f_{T_1}(s|\eta)ds$. The distribution of T_2 is assumed known and extracted from the reference population with the usual assumption that disease-specific death is negligible in this reference population (Ederer, et al. 1961). This is illustrated in the French breast cancer data analysis in section 4.2. The copula distribution linking $F_{T_1}(t)$ and $F_{T_2}(t)$ may be specified using simple parametric copula models such as the Archemedean copulas. The parameters in the copula model may be chosen for a pre-specified dependence between T_1 and T_2, for example, Kendall's tau (τ_k). In the numerical illustrations, T_1 was assumed to follow a Weibull distribution with parameter $\eta = (\lambda, \alpha)$ and probability density function $f_{T_1}(t|\eta) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1} \exp\left\{-\left(\frac{t}{\lambda}\right)^{\alpha}\right\}$ because of its versatility to accommodate a wide range of hazard shapes. We consider the Gumbel and the Clayton copulas in sections 2.3 and 2.4 for the joint distribution of T_1 and T_2 as both copulas exhibit tail behaviours that mimic the mortality trend observed in the cancer registry data. The bivariate joint distribution and density functions for the Gumbel copula are:

$$
\begin{aligned}
F_{T_1,T_2}(t,t|\eta) &= \exp\left\{ -\left((-log\,(u_1))^{\theta} + (-log(u_2))^{\theta} \right)^{\frac{1}{\theta}} \right\} \\
f_{T_1,T_2}(t,t|\eta) &= F_{T_1,T_2}(t,t|\eta) \cdot \left(\left(-\log\,(u_1)^{\theta} \right) + \left(-log\,(u_2)^{\theta} \right) \right)^{\frac{1}{\theta}-1} \\
&\times \left(\left(-\log(u_1)^{\theta-1} \cdot \frac{f_{T_1}(t|\eta)}{u_1} \right) + \left(-log\,(u_2)^{\theta-1} \cdot \frac{f_{T_2}(t|\eta)}{u_2} \right) \right), \qquad (2.9)
\end{aligned}
$$

while under the Clayton copula, the bivariate joint distribution and density functions are:

$$
\begin{aligned}
F_{T_1,T_2}(t,t|\eta) &= \left(u_1^{-\theta} + u_2^{-\theta} - 1 \right)^{-\frac{1}{\theta}} \\
f_{T_1,T_2}(t,t|\eta) &= \frac{F_{T_1,T_2}(t,t|\eta)}{\left(u_1^{-\theta} + u_2^{-\theta} - 1 \right)} \cdot \left(\frac{f_{T_1}(t|\eta)}{u_1^{\theta+1}} + \frac{f_{T_2}(t|\eta)}{u_2^{\theta+1}} \right) \qquad (2.10)
\end{aligned}
$$

where $u_1 = F_{T_1}(t|\eta), u_2 = F_{T_2}(t)$.

The maximum likelihood estimator (MLE) of η can be obtained by maximizing the log-likelihood function in (2.8) using Nelder-Mead algorithm (Nelder and Mead, 1965). Parameter estimation was sensitive to the choice of initial parameter values when $\tau_k \in (0.6, 0.9)$ for small sample sizes with larger ($> 50\%$) censoring proportions. Because the model is highly nonlinear, computing may be unstable, particularly with small sample sizes and high censoring proportions. We suggest using multiple starting values wherever possible and taking the MLE to be the maximizer giving the largest value of the log likelihood across all starting values. The usual regularity conditions for the MLE hold, given that the estimator converges in probability, that is $\hat{\eta} \xrightarrow{P} \eta$ and is asymptotically normal, $\widehat{\eta} \sim N\left(\eta, I_O(\eta)^{-1}\right)$ with variance estimated using the inverse of the observed information matrix $(I_O(\eta)^{-1})$ evaluated at the MLE, $\hat{\eta}$. The observed information matrix is:

$$
\begin{aligned}
I_O(\eta) &= \frac{\partial^2 l(\eta|\mathbf{X},\Delta)}{\partial\eta\partial\eta^T} \\
&= \sum_{i=1}^{n}\left\{\frac{\delta_i\cdot[f_T(X_i)]\cdot\left\{\frac{\partial}{\partial\eta}f_T(X_i)\right\}^T\left\{\frac{\partial}{\partial\eta}\left[f_T(X_i)\right]\right\}}{[f_T(X_i)]^T[f_T(X_i)]}\right\}+ \\
&\quad \sum_{i=1}^{n}\left\{\frac{(1-\delta_i)\cdot[S_T(X_i)]\cdot\left\{\frac{\partial}{\partial\eta}[S_T(X_i)]\right\}^T\left\{\frac{\partial}{\partial\eta}[S_T(X_i)]\right\}}{[S_T(X_i)]^T[S_T(X_i)]}\right\}
\end{aligned}
\tag{2.11}
$$

Since the dependence structure for time to disease mortality (T_1) and time to other competing mortality (T_2) is nonidentifiable and unverifiable from the observed registry data, we propose a sensitivity analysis, where the analysis is conducted across a range of assumed dependence structures. The levels of dependence represent the varying levels of dependent competing mortality possible in the observed registry data. For each copula dependence structure, we estimate η with $\hat{\eta}$ and compute $F_{T_1}(t|\hat{\eta})$ to estimate relative survival. The corresponding standard errors are obtained as the square root of the Delta method variance: $Var(\widehat{S_{T_1}(X)}) = g(\widehat{S_{T_1}(X)})\cdot I_O(\hat{\eta})^{-1}\cdot g^T(\widehat{S_{T_1}(X)})$ where $g(\eta)$ is the derivative of $S_{T_1}(t|\eta)$ with respect to η. Due to the complex nature of the likelihood, numerical approximation is used to estimate the information matrix in the numerical illustrations in Section 2.4.

In the presence of informative censorship where T and C are dependent, we propose conditioning on additional covariates Z in F_{T_2}, (Sasieni and Brentnall, 2017; Perme et al., 2012), where $F_{T_2}(t|Z)$ is the conditional distribution of T_2 given Z. Such covariates might include age, sex, period, as well as other relevant demographic variables. Let Z_i be the covariate observed on individual $i = 1,\cdots,n$. The log-likelihood function (2.8) is easily modified, where the likelihood contribution for individual i $(= 1,\cdots,n)$ is (2.7) with $F_{T_2}(t|Z_i)$ replacing $F_{T_2}(t)$ in $f_T(X_i)$ and $S_T(X_i)$. Here, we estimate η in $F_{T_1}(t|\eta)$ unconditionally on Z to mitigate against the bias associated with these covariates (Sasieni and Brentnall, 2017; Perme et al., 2012). The usual likelihood regularity conditions continue to hold, with the resulting estimator $\hat{\eta}$ being consistent and asymptotically normal with variance which may be estimated using the inverse of the observed information matrix evaluated at $\hat{\eta}$.

2.4 Numerical Illustrations

2.4.1 Simulation Studies

To evaluate the performance of our proposed method, we simulated data to mimic the French breast cancer data set for sample sizes; 1000, 2500 and 5000 with 500 replications. The latent failure times for $T_j \sim Weibull(\alpha_j, \lambda_j)$ with probability density function defined above in section 2.3. The parameters for the Weibull distribution for T_1 were $\lambda_1 = 0.182$ and $\alpha_1 = 1.609$, while those for T_2 were $\lambda_2 = 0.742$ and $\alpha_2 = 0.693$. In the estimation of $\lambda_1,\ \alpha_1$ for T_1, $\lambda_2,\ \alpha_2$ are assumed known for T_2 and vice versa for estimation of λ_2 and α_2. Noninformative censoring times were generated from a uniform distribution $(0, \gamma)$, where γ was chosen for $10,\ 30$ and 50% censoring. We consider the Gumbel copula with Kendall's tau, $\tau_k = 1 - \frac{1}{\theta} = 0,\ 0.25,\ 0.50,$ and $\ 0.75$. Initial parameter values were randomly chosen from uniform distributions, with multiple starting values wherever possible as described in section 2.3. We also simulated data from the Clayton copula. The results are similar to those for the Gumbel copula and are described in the appendix 1, table $A.12$. Appendix 1 also show the distribution, density and contour plots for T_1 treating T_2 as competing event for both Gumbel and Clayton copulas for independence and strong dependence structures.

Appendix 1 also shows the dependency level possible in breast cancer registry data. The simulation reveal the trend observed from time since diagnosis and provide useful insight into understanding possible informative censorship. These figures: A.6, A.7 and A.8 suggests the levels of competing mortality during the course of treatment and may also provide information to physicians as to which time of treatment that may be more potent.

Tables 2.1, and 2.2 show the results for estimation of the model for T_1 treating T_2 as a competing event and for T_2 treating T_1 as a competing event. The bias is small decreasing to zero as the sample size increases for each of the censoring levels. The empirical variance and the model based variance tend to agree and the coverage is close to the nominal 0.95 level, particularly at larger sample sizes. The empirical variance decreases as the sample size increases at roughly the expected root n rate.

Table 2.1: Estimated parameters of the model for T_1 across samples sizes (N), dependence levels (τ_k) and levels of censoring (C) treating T_2 as a competing event and vice versa.

	C		0.10					0.30					0.50				
τ_k	N	$\hat{\eta}$	Mean	Bias^a	ModB^a	EMP ^{a}	CP	Mean	Bias^a	ModB ^{a}	EMP ^{a}	CP	Mean	Bias^a	ModB^a	EMP^a	CP
0.00	1000	$\hat{\lambda_1}$	0.182	-0.080	0.090	0.090	0.940	0.182	-0.220	0.110	0.110	0.946	0.182	-0.290	0.150	0.170	0.928
		$\hat{\alpha_1}$	1.610	0.790	1.420	1.560	0.938	1.610	0.830	1.810	1.970	0.936	1.611	1.980	2.620	2.720	0.950
	2500	$\hat{\lambda_1}$	0.182	-0.260	0.030	0.030	0.964	0.182	-0.270	0.040	0.040	0.952	0.182	-0.010	0.060	0.050	0.960
		$\hat{\alpha_1}$	1.609	-0.450	0.570	0.590	0.940	1.610	0.130	0.720	0.730	0.952	1.609	-0.200	1.040	1.060	0.954
	5000	$\hat{\lambda_1}$	0.182	-0.050	0.020	0.020	0.948	0.182	0.120	0.020	0.020	0.944	0.182	0.050	0.030	0.030	0.958
		$\hat{\alpha_1}$	1.610	0.520	0.280	0.280	0.954	1.610	0.640	0.360	0.370	0.962	1.610	0.290	0.520	0.530	0.956
	1000	$\hat{\lambda_2}$	0.748	5.980	9.940	10.270	0.936	0.748	6.430	11.760	12.130	0.940	0.746	3.650	14.410	15.320	0.922
		$\hat{\alpha_2}$	0.694	0.840	6.790	7.020	0.944	0.694	1.300	7.510	7.710	0.958	0.697	4.070	8.490	0.010	0.948
	2500	$\hat{\lambda_2}$	0.742	-0.320	3.770	3.750	0.928	0.742	0.510	4.440	4.150	0.950	0.744	2.300	5.490	5.350	0.940
		$\hat{\alpha_2}$	0.694	0.730	2.680	2.700	0.946	0.694	0.360	2.960	2.820	0.958	0.693	-0.480	3.360	3.280	0.960
	5000	$\hat{\lambda_2}$	0.743	0.630	1.870	1.640	0.962	0.743	0.770	2.200	1.920	0.956	0.743	1.000	2.700	2.460	0.968
		$\hat{\alpha_2}$	0.693	0.100	1.340	1.210	0.962	0.693	0.120	1.480	1.310	0.964	0.693	0.280	1.680	1.530	0.958
0.25	1000	$\hat{\lambda_1}$	0.182	-0.480	0.080	0.080	0.948	0.182	-0.710	0.110	0.110	0.942	0.182	-0.450	0.140	0.130	0.954
		$\hat{\alpha_1}$	1.610	1.490	1.310	1.340	0.952	1.612	3.030	1.710	1.730	0.952	1.613	3.960	2.480	2.840	0.934
	2500	$\hat{\lambda_1}$	0.182	-0.320	0.030	0.030	0.958	0.182	-0.180	0.040	0.040	0.944	0.182	0.040	0.060	0.060	0.962
		$\hat{\alpha_1}$	1.608	-1.020	0.530	0.500	0.948	1.608	-1.210	0.690	0.670	0.950	1.608	-1.840	1.000	0.970	0.936
	5000	$\hat{\lambda_1}$	0.182	-0.050	0.020	0.020	0.956	0.182	-0.120	0.020	0.020	0.936	0.182	-0.130	0.030	0.030	0.940
		$\hat{\alpha_1}$	1.609	-0.250	0.260	0.240	0.956	1.610	0.520	0.340	0.320	0.958	1.610	0.590	0.500	0.460	0.950
	1000	$\hat{\lambda_2}$	0.753	10.920	14.700	14.430	0.938	0.756	13.950	17.140	16.970	0.946	0.760	18.430	20.460	20.810	0.946
		$\hat{\alpha_2}$	0.690	-3.170	8.240	7.930	0.954	0.689	-4.270	9.070	8.890	0.952	0.687	-5.930	10.080	9.900	0.944
	2500	$\hat{\lambda_2}$	0.744	2.420	5.490	5.230	0.952	0.745	3.270	6.320	6.120	0.946	0.746	3.910	7.380	6.580	0.952
		$\hat{\alpha_2}$	0.692	-0.780	3.240	3.110	0.948	0.692	-0.940	3.550	3.470	0.950	0.692	-0.770	3.930	3.680	0.956
	5000	$\hat{\lambda_2}$	0.741	-0.950	2.690	2.530	0.950	0.741	-0.870	3.090	3.000	0.952	0.742	0.260	3.620	3.580	0.964
		$\hat{\alpha_2}$	0.695	1.940	1.610	1.480	0.958	0.696	2.370	1.770	1.670	0.952	0.695	2.260	1.960	1.940	0.948

$\hat{\eta}$: estimated parameters, ModB: model-based variance, EMP: empirical variance, CP: 95% coverage probability. ^{a} : $\times 10^{-3}$

Table 2.2: Continuation: Estimated parameters of the model for T_1 across samples sizes (N), dependence levels (τ_k) and levels of censoring (C) treating T_2 as a competing event and vice versa.

	C				0.10					0.30					0.50		
τ_k	N	$\hat{\eta}$	Mean	Bias^a	ModB^a	EMP ^{a}	CP	Mean	Bias^a	ModB ^{a}	EMP ^{a}	CP	Mean	Bias^a	ModB^a	EMP^a	CP
0.50	1000	$\hat{\lambda}_1$	0.182	-0.030	0.070	0.070	0.956	0.182	-0.150	0.090	0.100	0.956	0.182	-0.260	0.120	0.120	0.954
		$\hat{\alpha}_1$	1.611	2.270	1.270	1.300	0.948	1.612	2.710	1.680	1.780	0.938	1.613	3.840	2.380	2.710	0.928
	2500	$\hat{\lambda}_1$	0.182	-0.170	0.030	0.030	0.952	0.182	-0.090	0.040	0.040	0.936	0.183	0.190	0.050	0.050	0.948
		$\hat{\alpha}_1$	1.612	2.270	1.270	1.300	0.948	1.609	-0.690	0.670	0.650	0.952	1.608	-1.050	0.950	0.920	0.952
	5000	$\hat{\lambda}_1$	0.182	0.010	0.010	0.020	0.946	1.823	-0.050	0.020	0.020	0.934	1.824	0.040	0.020	0.030	0.956
		$\hat{\alpha}_1$	1.609	-0.340	0.250	0.240	0.954	1.610	0.420	0.330	0.330	0.946	1.610	0.450	0.480	0.510	0.932
	1000	$\hat{\lambda}_2$	0.759	17.440	19.080	20.140	0.932	0.763	21.070	21.840	22.880	0.932	0.767	25.540	25.050	25.330	0.932
		$\hat{\alpha}_2$	0.688	-5.180	9.440	9.910	0.944	0.686	-6.950	10.280	10.840	0.942	0.684	-9.170	11.210	11.510	0.940
	2500	$\hat{\lambda}_2$	0.742	-0.050	6.860	6.120	0.942	0.741	-1.330	7.730	7.360	0.940	0.742	0.030	8.740	0.800	0.940
		$\hat{\alpha}_2$	0.694	1.320	3.670	3.340	0.954	0.696	2.750	3.980	3.790	0.948	0.695	2.040	4.310	4.000	0.950
	5000	$\hat{\lambda}_2$	0.740	-1.580	3.360	3.380	0.944	0.741	-0.460	3.820	3.840	0.950	0.744	1.620	4.340	4.660	0.946
		$\hat{\alpha}_2$	0.695	1.720	1.820	1.870	0.936	0.694	1.290	1.980	2.040	0.946	0.692	-0.750	2.150	2.270	0.944
0.75	1000	$\hat{\lambda}_1$	0.182	-0.200	0.060	0.070	0.956	0.182	-0.110	0.070	0.070	0.954	0.182	-0.260	0.100	0.100	0.948
		$\hat{\alpha}_1$	1.610	0.490	1.060	1.520	0.936	1.611	1.320	1.360	1.590	0.928	1.612	2.660	2.090	2.370	0.942
	2500	$\hat{\lambda}_1$	0.182	-0.160	0.020	0.030	0.930	0.182	-0.120	0.030	0.030	0.940	0.182	0.050	0.040	0.040	0.934
		$\hat{\alpha}_1$	1.609	-0.580	0.420	0.450	0.954	1.609	-0.850	0.540	0.580	0.944	1.608	-1.280	0.840	0.890	0.936
	5000	$\hat{\lambda}_1$	0.182	0.050	0.010	0.010	0.948	0.182	0.010	0.010	0.020	0.944	0.182	-0.020	0.020	0.020	0.952
		$\hat{\alpha}_1$	1.609	-0.010	0.210	0.210	0.944	1.610	0.250	0.270	0.280	0.946	1.609	-0.120	0.420	0.450	0.948
	1000	$\hat{\lambda}_2$	0.760	17.780	20.190	20.360	0.938	0.763	20.780	22.630	22.390	0.938	0.766	24.200	26.040	25.790	0.952
		$\hat{\alpha}_2$	0.689	-4.600	9.870	10.440	0.946	0.687	-5.800	10.610	10.900	0.948	0.685	-7.510	11.580	11.590	0.956
	2500	$\hat{\lambda}_2$	0.745	3.240	7.290	7.390	0.940	0.743	0.670	7.980	8.150	0.936	0.743	1.420	9.090	9.270	0.938
		$\hat{\alpha}_2$	0.693	-0.030	3.830	3.760	0.958	0.695	2.180	4.080	4.060	0.954	0.694	1.770	4.450	4.390	0.958
	5000	$\hat{\lambda}_2$	0.742	-0.190	3.540	3.770	0.946	0.743	0.980	3.900	4.060	0.946	0.743	0.830	4.440	4.540	0.950
		$\hat{\alpha}_2$	0.694	0.550	1.900	1.990	0.930	0.693	-0.150	2.030	2.090	0.944	0.693	0.350	2.210	2.280	0.944

$\hat{\eta}$: estimated parameters, ModB: model-based variance, EMP: empirical variance, CP: 95% coverage probability. ^{a} : $\times 10^{-3}$

2.4.2 Application to French Breast Cancer Data

In this section we analyze data from women between the ages of 18 and 96 years surviving breast cancer in France from 1980 to 2011. The data were obtained from the Institut Curie breast cancer database. This database contains records from $24,458$ nonmetastatic breast cancer patients treated at the Institut Curie. Out of the $24,458$ breast cancer patients, $9,885$ (40.4%) died while $14,573$ were alive and administratively censored on December 31^{st} 2011. Five age group categories were considered for the estimation of relative survival. $3,970$ were between the ages of $15-44$, $6,895$ between the ages of $45-54$, $6,420$ between the ages of $55-64$, $4,675$ between the ages of $65-74$ and $2,498$ were in the $75-99$ age group category. We individually matched the observed death or censoring time in the disease cohort group with a corresponding time in the healthy reference population on age, sex, and year (date of diagnosis and the date of death or censored) for each participant and for each follow-up period. The background mortality data from the Human Mortality Database (https://www.mortality.org) was last modified on June 28, 2018. Within each follow-up year, we assumed that $\lambda_P(t)$ is piecewise constant (Dickman, et al., 2004) for each period up to time X. The cumulative hazard for each period based on $\lambda_P(t)$ is calculated from the background survival function at the beginning and end of the period. The cumulative hazard is then used to obtain $\lambda_P(t)$ under the piecewise constant assumption. The goal of matching in determining $\lambda_{T_2} = \lambda_P$ is to mitigate the impact of age and calendar year on potentially dependent censoring by C (Perme et al., 2012). We estimate $2, 5, 10$, and $15-$year relative survival assuming a Weibull distribution for T_1 and a Gumbel copula model with differing levels of dependence to specify the joint distribution for the distributions of T_1 and T_2. We compared estimates from our parametric estimator to the estimates of the estimator of Perme et al. (2012), which require independence of T_1 and T_2 and employ $S_{T_2}(t)$ from the same reference population.

Tables 2.3 and 2.4 show the estimates of $S_{T_1}(t)$ for cancer mortality both overall and stratified by age. The parametric estimates under independence are similar to those from the Pohar-Perme method suggesting that the Weibull assumption is a reasonable fit to the data. One observes that as dependence increases, cancer survival generally decreases. For a fixed dependence level, younger women tend to have higher cancer survival rates than do older women, with marked reductions for the 65-74 and 75-99 age groups. There is some instability in survival estimates at 15 years, especially for the older age groups, as evidenced by the large standard errors. Perhaps, this may be due to small numbers of patients at risk at longer follow-up times.

The relative survival function under the independence assumption corresponds to an ideal world where the only cause of death is breast cancer. This quantity can only be estimated under unverifiable dependence assumptions between T_1 and T_2 using disease registry data. To account for uncertainty in dependence, we recommend reporting a range of probabilities corresponding to differing levels of dependence. For example, using results from table 2.3, the overall 5 year breast cancer survival from $1980 - 2011$ is estimated to be between 84.0-87.4% under dependence ranging from Kendall's tau equal to 0 (independence) to 0.75 (strong dependence). These cancer survival probabilities may be meaningfully compared with those in other populations having different background mortality rates and different dependence levels between T_1 and T_2.

Table 2.3: 2, 5, 10 and 15-yr overall relative survival for French women diagnosed with breast cancer between 1980 and 2011.

τ_k	0.00			0.25		0.50		0.75	
Year	PP^a	$S_{T_1}(t)^a$	SE^b	$S_{T_1}(t)^a$	SE^b	$S_{T_1}(t)^a$	SE^b	$S_{T_1}(t)^a$	SE^b
2	95.6	96.0	6.99	95.8	6.96	95.4	7.23	94.7	7.74
5	84.8	87.4	9.01	86.6	9.10	85.5	9.31	84.0	9.53
10	71.0	72.8	11.01	71.4	10.99	69.8	10.91	68.0	10.67
15	59.5	59.5	12.22	57.9	12.08	56.3	11.74	54.9	11.19

a : $\times 10^{-2}$, b : $\times 10^{-3}$, τ_k: dependence, PP: Pohar-Perme, $S_{T_1}(t)^a$: parametric relative survival estimate at year t, SE: standard error for the relative survival estimate.

The results of a sensitivity analysis was conducted across different levels of dependence structures each representing different competing mortality observed in the registry data. Figure 2.4.2 shows the 2,5,10 and 15-yr overall breast survival plots across a spectrum of dependence structures for women between the ages of 18 and 96-yr living in France during 2008 and 2011.

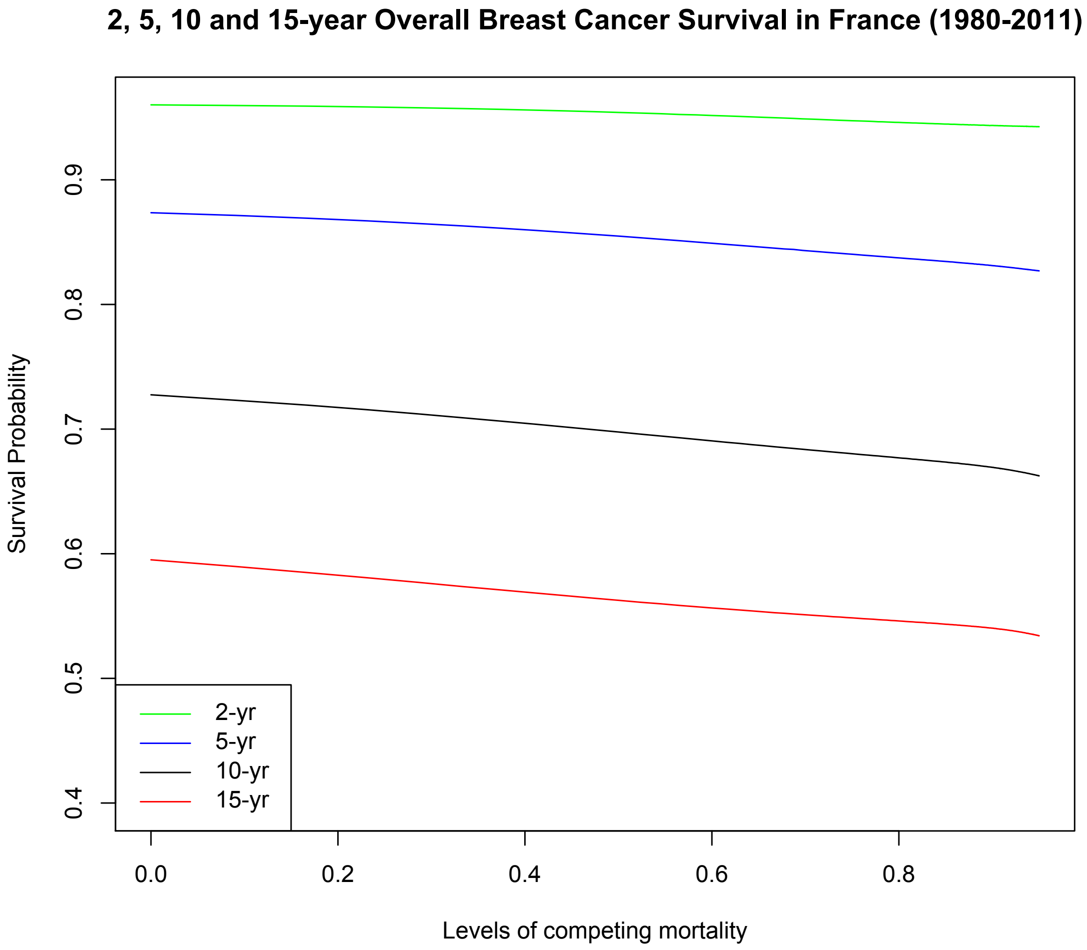

Figure 2.1: Although the graphics looks like a straight line, these are actually survival curves spanning the spectrum of dependence structures (0-0.9) each representing the levels of competing mortality.

Table 2.4: 2, 5, 10 and 15-yr age group specific relative survival for French women diagnosed with breast cancer between 1980 and 2011.

τ_k			0.00		0.25		0.50		0.75	
Year	Agegp	PP^a	$S_{T_1}(t)^a$	SE^b	$S_{T_1}(t)^a$	SE^b	$S_{T_1}(t)^a$	SE^b	$S_{T_1}(t)^a$	SE^b
2	15-44	95.8	94.9	20.90	94.9	20.73	94.8	20.73	94.8	20.68
	45-54	97.1	96.6	16.44	96.5	16.13	96.3	16.27	96.2	16.40
	55-64	95.7	96.1	13.72	96.0	13.49	95.7	13.70	95.3	14.12
	65-74	95.1	97.0	08.50	96.8	08.54	96.2	09.61	95.1	11.60
	75-99	91.5	96.5	07.94	95.6	08.93	93.4	12.44	89.9	17.16
5	15-44	85.1	86.9	23.70	86.8	23.64	86.7	23.62	86.7	23.35
	45-54	88.6	90.4	19.39	90.1	19.36	89.8	19.45	89.7	19.28
	55-64	85.8	88.1	17.72	87.6	17.71	86.9	17.87	86.6	17.66
	65-74	84.1	86.9	16.71	85.8	17.01	84.2	17.71	82.5	18.09
	75-99	72.3	77.1	24.21	72.7	24.85	67.1	25.08	61.7	24.00
10	15-44	71.9	74.4	26.88	74.2	26.84	74.0	26.75	74.1	26.62
	45-54	78.3	80.1	22.83	79.6	22.80	79.2	22.73	79.2	22.34
	55-64	73.4	74.5	22.03	73.5	21.97	72.7	21.74	72.7	21.08
	65-74	68.4	67.2	25.38	65.0	25.32	63.0	24.72	62.3	23.20
	75-99	44.6	43.1	34.83	37.0	32.55	33.0	28.61	31.1	24.35
15	15-44	62.5	63.2	29.03	63.0	28.96	62.9	28.83	63.0	28.72
	45-54	70.8	70.5	25.31	69.8	25.24	69.4	25.00	69.6	24.51
	55-64	63.5	61.9	24.81	60.7	24.62	59.9	24.11	60.3	23.20
	65-74	50.3	48.7	30.06	46.2	29.46	44.7	27.92	45.3	25.47
	75-99	19.9	20.27	35.56	15.9	32.33	14.5	27.98	15.4	23.33

a : $\times 10^{-2}$, b : $\times 10^{-3}$, τ_k: dependence, Agegp: Age group, PP: Pohar-Perme, $S_{T_1}(t)^a$: parametric relative survival estimate at year t, SE: standard error for the relative survival estimate.

2.5 Discussion and Conclusion

Our model formulation for competing risk data without cause of failure information is general, permitting arbitrary but known copula functions. The distribution of other cause mortality is obtained from external reference data (Sarfati et al., 2010; Perme et al., 2012; Sasieni and Brentnall, 2017). We have undertaken preliminary investigations of simultaneous estimation of the dependence parameter and the parameter in the disease-specific survival distribution. There is evidence of instability, with care needed in the model specification to aid identifiability. This is expected, as there are similar identifability issues even when the cause of failure is known. The proposed sensitivity analysis is a practical solution to this issue, providing a range of estimates across different dependence levels not requiring simultaneous estimation of the dependence parameter. The parametric model for disease-specific mortality is restrictive but may be flexible enough for applications where the hazard is smooth over time, which is the case in cancer registry data. To relax the parametric assumption, nonparametric techniques are currently being developed which should be valuable in settings with more complex failure patterns.

The focus of relative survival analysis is the distribution of the latent event time for death from disease. This endpoint has been advocated by many practitioners (Slud et al., 1988; Reason, 1990; Louzada et al., 2015), as it removes the impact of other cause mortality on the risk of disease-specific mortality, permitting comparisons across populations with different background mortality. As an alternative, other work has considered estimation of the crude disease-specific survival, $C_k(t)$, using the relative survival estimates and the known reference hazard for other cause mortality (Cronin and Feuer, 2000). An analogous procedure could be implemented using our copula based estimate of the distribution of T_1 and would provide an assessment of the sensitivity of the estimator of C_k under independence of T_1 and T_2. Such procedure would be of interest to individuals who prefer crude disease-specific mortality to net disease-specific mortality. This is a topic for future research.

In conclusion, and unlike the Perme et al. (2012) and Cronin and Feuer (2000) methods which focused exclusively on the estimators for net survival and crude survival or crude probability of death measures respectively under the independence of competing mortality, our estimator provide both estimates for net and crude survival measures regardless of the independent competing mortality.

CHAPTER 3: A Nonparametric Method for Dependent Competing Risk in Relative Survival Analysis

3.1 Introduction

Missing data is a common problem in most biomedical studies including population-based cancer registries with competing events where the occurrence of one event type impedes other event types (Putter et al., 2007; Lau et al., 2009; Austin et al., 2016). In such registries, the cause of event is one of the most important variables documented for disease-specific survival estimation, which is used in comparison of disease-specific survival among groups or populations under different competing risk setting and is often of great interest to physicians for determining prognosis and effectiveness of treatment regimen. The standard assumption in such competing risk studies is that cause of disease-specific event is known (Gichangi and Vach, 2005).

Credible disease-specific analysis for competing risk data require accurate documentation of cause of death (Percy et al., 1981; Welch and Black, 2002; Mieno et al., 2016; Tan et al., 2019). A challenge is that cause of death information may be missing or subject to misspecification (James and Bull, 1996; Maudsley and Williams, 1996; Platell and Semmens, 2004; Lambert et al., 2010) in the registries making it impossible to distinguish disease and non-disease related events. For example, Welch and Black (2002) raised concern that cancer death rates are systematically misclassified, in that 41% of cancer patients who died as a result of cancer directed surgery (within one month of diagnosis) do not have cancer recorded as the underlying cause of death. Without reliable cause of death information, disease-specific analysis using classical methods is difficult and practically impossible (Percy, 1989; Hoel et al., 1993; Ederer et al., 1999; Begg and Schrag, 2002). With imprecise definitions and different levels of cause of death documentations, the World Health Organization (Organization et al., 1977) defines cause of death as ”the disease or injury which initiated the train of morbid events leading directly to death”. This was to aid harmony in cause of death definition across countries.

However, substantial disharmony in registries still exists not only in Europe but also in the United States of America. Sometimes, the underlying cause of death may be unclear as hospital cancer coding may not agree with the death certificate coding. Even when reliable cause of death information is available, it is often inaccessible and located in separate databases, which may be costly to obtain and difficult to link with registry data. Although much effort has been directed to link vital statistics with cancer registries in the United States of America (German et al., 2011), substantial disharmony (for example; with varying levels across all 79 sites for the Surveillance Epidemiology, and End Results (SEER) program in the United States of America) exists in determining cause of death information.

In competing risk studies with known cause of death information, standard methodologies for disease-specific survival assumes that time to disease-specific event is independent of time to non-disease-specific event (Fermanian, 2003; Gichangi and Vach, 2005; Austin et al., 2016). It is not uncommon to find competing risk studies where this assumption is grossly violated as most clinical research (Austin and Fine, 2017; D'Amico et al., 2018), often have competing causes (Austin et al., 2020). Austin et al. (2016) showed that majority ($> 77\%$) of randomized control trials with potential competing risks were ignored during statistical analyses.

In analyzing competing risk data with missing or unreliable cause of death information, classical methods like Kaplan-Meier method (Kaplan and Meier, 1958), logrank test (Bland and Altman, 2004) and standard proportional hazard model (Cox, 1972) are inapplicable for estimating disease-specific survival. Without competing events, these methods are useful for estimating disease-specific survival, comparing survival among groups and assessing the effect of covariates respectively. In the presence of competing risks with reliably known cause of death information, the analog of these methods (Aalen-Johanson estimator, Grey's test, and Fine-Gray model) are applicable. Ignoring the issues of missingness and or unreliable cause of death in competing risk setting, disease-specific survival analysis using the above methods are inadmissible (Austin et al., 2016) and (Adatorwovor et al., 2020) as they may introduce unintended biases, distort the accuracy of statistical inference and provoke misleading results.

While models based on independent timings of event and competing event are ubiquitous when analysing registry data without missing cause of death information, dependent models are relegated and not available for disease-specific survival analysis (Tan et al., 2019; de Lacerda et al., 2019). To address this knowledge gap, we consider dependence models for estimating disease-specific survival via copula for registry data without the need for missing and or unreliable cause of death information. Earlier attempt has being made to model the dependence in competing risk studies under parametric assumption (Adatorwovor et al., 2020).

Suppose the time to potential unobservable failure time $T = \min\{T_k : \ k = 1, 2, 3, \cdots, K\}$ and failure type $\varepsilon = \{k : T = T_k\}$ with $T_1, \cdots, T_K, \ K \in \mathbf{N}$ being the latent failure times associated with the K failure types. With $\varepsilon \geq 2$ implying competing risk setting (Dignam et al., 2012). In the absence of reliable cause of death information, relative survival methods have being proposed. Relative survival, $S_R(t)$ is the ratio of the observed survival rate in a group of cancer patients, during a specified period, to the expected survival rate in a healthy reference population (Ederer, 1961). At time t,

$$S_R(t) = \frac{S_O(t)}{S_P(t)} \tag{3.12}$$

where $S_O(t)$ is the survival probability for an individual in the registry and $S_P(t)$ is the expected survival from mortality tables. Existing methods focused exclusively on the estimation of $S_R(t)$ under the independence of T_1 and T_2 where $S_O(t) = S_{T_1}(t) \cdot S_{T_2}(t), \ S_P(t) = S_{T_2}(t)$ which implies $S_R(t) = S_{T_1}(t)$ with $S_{T_1}(t)$ and $S_{T_2}(t)$ being the survival probabilities corresponding to T_1 and T_2 respectively. Equation (3.12) can be rewritten in terms of disease-specific hazard function as $\lambda_O(t) = \lambda_E(t) + \lambda_P(t)$ (Cronin and Feuer, 2000), where $\lambda_O(t)$ is the hazard in the disease registry, $\lambda_E(t)$ is the so called excess hazard among the cancer cohort, and $\lambda_P(t)$ is the hazard from mortality tables. Under independence, $\lambda_E(t) = \lambda_{T_1}(t)$ and $\lambda_P(t) = \lambda_{T_2}(t)$, where $\lambda_{T_j}(t) = \frac{-dlogS_{T_j}(t)}{dt}, j = 1, 2,$ are the net hazard functions for cancer and other cause mortality. When T_1 is independent of T_2, the disease-specific survival probability $S_{T_1}(t)$ which is the target of relative survival analysis corresponds to a hypothetical population in which competing mortality is non-exist and differs from the cumulative incidence function which is commonly used to quantify disease-specific survival in competing risk analyses without missing cause of death information. Under dependence of T_1 and T_2, $S_{T_1}(t)$ is of interest to some practitioners who prefer crude survival (survival experienced in a real world where competing mortality exist simultaneously with disease-specific mortality) or crude probability death for disease-specific mortality rates to net probability of death.

Berkson and Gage (1950), Ederer and Heise (1959) and (Ederer, 1961) pioneered nonparametric relative survival method under the independence of T_1 and T_2. The Ederer II method (Ederer and Heise, 1959) was recommended (Hakulinen, Seppä, and Lambert, 2011) as the gold standard for estimating relative survival because its estimates are approximately close to the estimates of age-standardised relative survival ratio. Hakulinen (1982) proposed a variant of this method to address the bias due to heterogeneity of patient withdrawal within subgroups. A modification of Hakulinen (1982) method was proposed by Nixon et al. (1994) to address issues related to the dependence of patients' age on event occurrence or censorship. Stratified method based on age standardization of relative survival ratios was proposed to reduce biases associated with age (Corazziari et al., 2004). Perme et al. (2012) demonstrated that these classical methods may be biased under certain censoring patterns in population comparisons. Such bias may originate from unmeasured covariates affecting the cancer cohort group and the reference population from which rates of expected mortality are drawn. Rebolj Kodre and Pohar Perme (2013) proposed weighting corrections to address biases associated with censoring and age distribution (at the time of cancer diagnosis). Hakulinen, Seppä and Lambert (2011) and (Perme et al., 2012) developed estimators which are only valid under questionable assumption of independence of competing causes of event. However, the above estimation methods for $S_{T_1}(t)$ all require independence of T_1 and T_2 which cannot be substantiated in practical application settings.

We relaxed the independence assumption by formulating the dependence between the latent failure times for death from disease and mortality due to competing causes using copula (Deheuvels, 1978). A bivariate copula distribution for the latent failure times T_k (with $k = 2$) was generated taking as input their marginal distributions with a single dependence structure. Dependence models with copula have been widely utilized in survival analysis, including bivariate event times (Oakes, 1982), competing risks with known cause of failure (Heckman and Honoré, 1989), and semi-competing risks where one event time censors the other but not vice versa (Fine et al., 2001).

We employ such models for competing risk disease registry data with missing or unreliable cause of death information. Due to identifiability constraint of dependence for the joint distribution (Tsiatis, 1975), for the observed registry data, we treat the copula function as known. We nonparametrically modelled the marginal distribution of the time to disease-specific death with the distribution of other cause mortality drawn from the reference population. A variant of the Newton-Raphson procedure is used to solve the nonlinear function for disease-specific survival. Because the joint distribution is unidentifiable nonparametrically and unverifiable from the observed registry data, a sensitivity analysis is proposed where disease-specific survival is estimated across varying dependence structures, specified via the copula function. To our knowledge, this is the first attempt in accommodating dependence nonparametrically through the use of copula functions in estimating relative survival.

The rest of this paper proceeds as follows. In section 3.2, we present the data and copula model formulation for competing risks data. Section 3.3 describes the nonparametric estimation and inference procedure with missing cause of death information, bootstrap variance estimation as well as the proposed sensitivity analysis. In section 3.4, we present the numerical illustrations including simulation results and application to French breast cancer data. Section 3.5 discusses and concludes the paper.

3.2 Data and Model Formulation

Unlike the traditional endpoints, $\lambda_k(t)$ and $C_k(t)$ defined in Adatorwovor et al. (2020) for competing risk data with known cause of death, we focus on a function of the Kaplan-Meier (K-M) estimator (Kaplan and Meier, 1958) defined in (3.13) for all-cause survival probability estimation. Relative survival methods with missing or unreliable cause of death information focuses on the distribution of the latent failure times, $T_1 = \min(T_k)$ and T_2 (distribution derived from background population). With missing cause of death information, the observed data is simply time to event from any disease, T, which may be right censored by time to lost to follow up C. Under the standard assumption that T is independent of C, and for an individual i, the observed data consist of $X_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$, where T_i and C_i are the unobservable failure and censoring times respectively. The conventional representation of Kaplan-Meier estimator is:

$$S(X_i) = \prod_{i:\ X_i \leq x} \left(1 - \frac{d_i}{n_i}\right) \tag{3.13}$$

where for an individual i, d is the number of participants who died up until the mininmum time X, and n is the number of individuals known to have survived at time X.

We utilize copula models to capture the dependence between the distributions of T_1 and T_2. Copulas completely describe the dependence structure and provide scale invariant measures of association (Müller, 1996; Bäuerle and Müller, 1998; Denuit et al., 1999; Venter, 2002). Suppose ψ is a function defined such that $\psi : [0, 1] \rightarrow [0, +\infty]$ with independent marginal distributions, $u_j = P(T_j \leq t_j) = F_{T_j}(t_j) = 1 - S_{T_j}(t_j)\ \forall j \in (1, 2)$. Then, the copula model for the distributions corresponding to T_1 and T_2 (Cherubini et al., 2004; Joe, 1997; McNeil et al., 2009) is:

$$C(u_1, u_2) = P(T_1 \leq t, T_2 \leq t) = \psi\left(\psi^{-1}(u_1) + \psi^{-1}(u_2)\right) = F_{T_1,T_2}(t, t)$$

where ψ^{-1} is the inverse of ψ and ψ satisfies the Laplace-Stiltjes transform and (Bernstein et al., 1929) theorem, and $F_{T_1,T_2}(t, t)$ is the bivariate copula distribution function for the latent times T_1 and T_2 at time t. The generator function ψ is completely monotone for non-negative random variables with $\psi(0) = 1,\ \psi'(\cdot) < 0$ and $\psi''(\cdot) < 0$ (McNeil et al., 2009).

In theory, any scale invariant measure of association can be used to characterize dependence between the distributions of T_1 and T_2. The connection between Kendall's tau (τ_k) correlation coefficients and the generator function ψ has being shown (Genest and MacKay, 1986) as:

$$\tau_k = 1 + 4\int_0^1 \frac{\psi^{-1}(u)}{\psi^{-1}(u)'}du = 1 - 4\int_0^\infty u(\psi(u))^2 du$$

with $\psi^{-1'}$ being the derivative of ψ^{-1} and τ_k can be simplified to $1 - \frac{1}{\theta}$ for Gumbel. In this paper, we present our proposed method based on the Gumbel copula (G-copula) indexed by a single dependence parameter θ (having simple interpretations) to link the marginal distributions of T_1 and T_2. Thus:

$$C(u_1, u_2) = \exp\left[-\{(-log(u_1))^\theta + (-log(u_2))^\theta\}^{\frac{1}{\theta}}\right] \tag{3.14}$$

with $\theta \in (1, +\infty)$ and $u_j = F_{T_j}(X) = 1 - S_{T_j}(X)$ being the distribution function corresponding T_1 and T_2 respectively. When $\theta = 1$, T_1 and T_2 are independent implying that $C(u_1, u_2) \leq \theta(1 - u_1 - u_2) + (1 + \theta)u_1u_2$ but with $\theta > 1$ implying that T_1 and T_2 are dependent. The general bivariate survival function at time t for any copula function is:

$$\begin{aligned} S_T(t, t) &= S_{T_1}(t) + S_{T_2}(t) - 1 + \psi\left(\psi^{-1}(1 - S_{T_1}(t)) + \psi^{-1}(1 - S_{T_2}(t))\right) \\ &= S_{T_1}(t) + S_{T_2}(t) - 1 + C(u_1, u_2) \end{aligned} \tag{3.15}$$

where $S_T(t,t)$ is the all-cause survival probability, $S_{T_1}(t)$ and $S_{T_2}(t)$ are the corresponding survival distribution functions for T_1 and T_2 respectively. Under the independence of T_1 and T_2, (3.15) becomes $S_T(t,t) = S_{T_1}(t) \cdot S_{T_2}(t)$. Estimation of $S_{T_1}(t)$ is achieved by inversion of the survival function defined in (3.15) as:

$$S_T(t,t)^{-1} = S_{T_1}(t) - S_{T_2}(t) + 1 - C(u_1, u_2) \tag{3.16}$$

Unlike the upper tail dependence exhibited by the G-copula, the Clayton copula (C-copula) in contrast exhibit lower tail behaviour which also mimics the mortality trend in the observed registry data. The bivariate joint distribution function at the time X for the C-copula for distributions of T_1 and T_2 is:

$$F_{T_1,T_2}(X,X) = \Big(F_{T_1}(X)^{-\theta} + F_{T_2}(X)^{-\theta} - 1\Big)^{-\frac{1}{\theta}} \tag{3.17}$$

When T_1 and T_2 are dependent ($\theta > 0$) the bivariate dependence survival function is:

$$S_T(X,X) = S_{T_1}(X) + S_{T_2}(X) - 1 + \Big(F_{T_1}(X_i)^{-\theta} + F_{T_2}(X_i)^{-\theta} - 1\Big)^{-\frac{1}{\theta}} \tag{3.18}$$

with the inversion formula as:

$$S_T^{-1}(X,X) = S_{T_1}(X) - S_{T_2}(X) + 1 - \Big(F_{T_1}(X)^{-\theta} + F_{T_2}(X)^{-\theta} - 1\Big)^{-\frac{1}{\theta}} \tag{3.19}$$

3.2.1 Monotonicity

The derivative of the estimating equation defined in 3.20 can be established for

$$g(S_{T_1}(t)) = S_{T_1}(t) - S_{T_1}(t) - S_{T_2}(t) + 1 - C(u_1, u_2) = 0 \tag{3.20}$$

is:

$$g'(S_{T_1}(t)) = -1 - C'(u_1, u_2) = -(1 + C'(u_1, u_2)) < 0 \tag{3.21}$$

where $C'(u_1, u_2) = \psi'(\psi^{-1}(u_1) + \psi^{-1}(u_2)) \cdot \Big(\psi^{-1\prime}(u_1) + \psi^{-1\prime}(u_2)\Big)$

3.3 $S_{T_1}(t)$ Estimation and Inference

Relative survival under the independence of T_1 and T_2 is given by $S_{T_1}(X) = \frac{S_T(X,X)}{S_{T_2}(X)}$. Under dependence, relative survival estimates are based on the inversion formula in (3.16). In section 3.4, T_1 was assumed to follow a Weibull distribution with parameter $\eta = (\lambda, \alpha)$ and probability density function $f_{T_1}(t|\eta) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1} \exp\left\{-\left(\frac{t}{\lambda}\right)^{\alpha}\right\}$ because of its versatility to accommodate varying hazard shapes while the distribution of T_2 is derived from the background population using the piecewise exponential constant function; which rely on the cumulative hazard function defined for the Nelson-Aalen (N-A) estimator (Nelson, 1972; Aalen and Johansen, 1978). The N-A estimator $\Lambda(X)$, is an estimator used to estimate the cumulative number of expected event(s) and is $\Lambda(X) = \sum\limits_{X \leq x} \dfrac{d_i}{n_i}$. The relationship between K-M estimator and N-A estimators can be established as $\Lambda(X_i) = -\log(S(X_i))$, where $S(X_i)$ is defined in (3.13).

In the estimation of disease-specific survival $S_R(t)$ which is $S_{T_1}(t)$, we rely on a function of the Kaplan-Meier estimator for all-cause survival probability for an individual i surviving beyond the time point X. From equation (3.13), we estimate the all-cause survival probability at time X and denote it by $\widehat{S_T(X,X)}$. In order to estimate $S_{T_1}(X)$, we replaced the estimator $S_T(X,X)$ in equation (3.15) with $\widehat{S_T(X,X)}$. $S_{T_2}(X)$ is derived from a healthy reference background population and together with $\widehat{S_T(X,X)}$ is substituted into the inversion formula in (3.16) for the estimation of $S_{T_1}(X)$.

Due to the complex nature of the nonlinear equation (3.16), a variant of Newton-Raphson algorithm (Hasselman, 2009) was implemented to obtain a numerical solution for $S_{T_1}(X)$. The Kaplan-Meier (K-M) estimator for all-cause survival $S_{T_1}(X_i)$ may be subject to monotonicity constraint (Fine et al., 2001) especially for C-copula. In such scenarios, the estimate for the estimator $S_{T_1}(X_i)$ for an individual i is $\widehat{S^*_{T_1}(X_i)} = \min\limits_{X_i \leq x}\left(\widehat{S_{T_1}(X_i)}, \widehat{S_{T_1}(X_i - 1)}\right)$ wherever possible with $S_{T_1}(X_i - 1)$ being the survival probability corresponding to the previous time point. While estimating $S_{T_1}(X)$ in the presence of a fixed dependence parameter θ, appropriate choice of the initial values for the parameters escapes the monotonicity constraints for the G-copula.

In assessing the performance of our method, we estimated the bias of the estimator at time X for each of the simulation studies presented in section 3.4. We showed that the estimator presented in equation (3.22) was unbiased for $S_{T_1}(X_i)$ for each individual i surviving at time X. The estimator for $S_{T_1}(X)$ is;

$$\widehat{S_{T_1}(X)} = g\left[\left(S_T\widehat{(X,X)}\right), (S_{T_2}(X))\right] \tag{3.22}$$

where $S_{T_1}\widehat{(X,X)}$ is given in equation (3.13) and is the K-M estimator for the bivariate copula survival function and g is a monotone function define for the distribution functions of T_1 and T_2 respectively.

Under the usual regularity conditions, $\widehat{S_{T_1}(X)}$ is asymptotically normal and consistent. As $n \to \infty$, $n^{\frac{1}{2}}|\widehat{S_{T_1}(X)} - S_{T_1}(X)|$ converges to a Gaussian process with mean zero and variance, $\widehat{Var}\left(\widehat{S_{T_1}(X)}\right)$ that was deduced using a nonparametric bootstrap variance estimation method, $\forall\ X \geq 0$. Nonparametric bootstrap procedure (describe below) is implemeted for variance estimation corresponding to the estimated quartile time X. A consistent estimator for the variance of $S_T\widehat{(X,X)}$ is given by Greenwood formula described in the appendix 2. The variance estimation of $S_{T_1}(X)$ was achieved by the bootstrap method implemented using the following procedure where $B = 500$ bootstrapped samples:

1. Draw B samples of size n with replacement from the original data set.

2. Calculate $\hat{\eta}$ for each of the samples from step 1. That is, we now have $\hat{S}_{T_{1_1}}, \cdots, \hat{S}_{T_{1_B}}$

3. We calculate the standard error from the B estimates of $\hat{S}_{T_1}$ by using the standard formulas for standard errors, $se(\hat{S}_{T_1}) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{S_{T_1}}_i - \bar{\hat{S_{T_1}}})^2}$, with $\bar{\hat{S_{T_1}}} = \frac{1}{B}\sum_{i=1}^{B}\hat{S_{T_1}}_i$.

Clarke et al. (2009) showed that

$$\frac{1}{n}\sum_{i=1}^{n}\hat{S}_n(X) \xrightarrow{p} S(X) \tag{3.23}$$

where $\hat{S}_n = 1 - \hat{F}_n$ and $\hat{F}_n$ being the enpirical CDF and $F(X)$ is the true distribution of the estimate. In the simulation study, we showed the estimation of the variance at each of the quartile times X for the model for each sample size and for 15% censoring level. The corresponding 95% coverage probability was computed based on the estimated bootstrap variance.

3.4 Numerical Illustrations and Applications

3.4.1 Simulation Procedure

We generate competing risk data that mimicked the French breast cancer data set to evaluate our proposed method. Sample sizes; 2500, 5000 and 10000 were simulated each with 500 replications. The unobservable latent failure time T_j was allowed to follow Weibull distribution with α_j, λ_j as parameters with the probability density function defined in section 3.3. The parameters for the Weibull distribution for T_1 were $\lambda_1 = 0.182$ and $\alpha_1 = 1.609$, while those for T_2 were $\lambda_2 = 0.742$ and $\alpha_2 = 0.693$. In the estimation of $\lambda_1,\ \alpha_1$ for T_1, $\lambda_2,\ \alpha_2$ are assumed known for T_2 and vice versa for estimation of λ_2 and α_2. Noninformative censoring times were generated from a uniform distribution $(0, \gamma)$, where γ was chosen for 15% censoring. G-copula dependence was chosen with Kendall's tau, $\tau_k = 0,\ 0.25,\ 0.50,$ and $\ 0.75$. Initial parameter values were randomly chosen from uniform distributions, with multiple starting values wherever possible as described in section 3.3. Clayton copula data could be simulated based on the description in section 3.2.

3.4.2 Figures

Figures 3.4.2 and 3.4.2 show the nonparametric survival probability function for T_1 under the G-copula for both zero and moderate dependence ($\tau_k = 0.5$) and for 1000 sample size. The step function for the nonparametric estimate $\widehat{S_{T_1}(t)}$ for T_1 is close to the truth, $S_{T_1}(t)$ at the time point t. The survival probability corresponding to the lower, median and upper quartiles is presented in tables 3.7 and 3.8 of the model for T_1 treating T_2 as a competing mortality, and for T_2 treating T_1 as a competing event. We observed that bias is small decreasing to zero for increasing sample size across each level of dependence. The empirical variance and the model based variance tend to agree and the coverage is close to the nominal 0.95 level, particularly at larger sample sizes. The empirical variance decreases as the sample size increases at roughly the expected root n rate.

The following figures 3.4.2, 3.4.2, 3.4.2 and 3.4.2 compare the true estimator $S_{T_1}(X_i)$ to the estimate $\widehat{S_{T_1}(X_i)}$ with zero (independence of T_1 and T_2), 25, 50, and 75% dependence while applying 15% right censoring for the G-copula model (3.27). The figures 3.4.2 and 3.4.2 also reveal that the

Table 3.5: Estimated parameters of the model for T_1 across samples sizes (N), dependence levels (τ_k) with 20% censoring (C) treating T_2 as a competing event and vice versa.

τ_k	N	X	$\widehat{S_{T_1}(X)}$	$S_{T_1}(X)$	Bias ^{a}	$S_1^*(X)$	B-Var ^{a}	EMP ^{a}	CP
0.00	2500	0.706	0.932	0.932	39.75	0.932	2.930	2.700	0.962
		0.965	0.715	0.714	87.78	0.715	3.240	3.260	0.952
		1.173	0.409	0.409	1.67	0.409	1.820	1.800	0.956
	5000	0.716	0.927	0.927	20.11	0.927	1.860	1.74	0.956
		0.989	0.683	0.683	46.510	0.683	1.830	1.730	0.946
		1.206	0.359	0.359	-17.270	0.359	8.20	7.800	0.952
	10000	0.705	0.932	0.932	-1.630	0.932	0.730	0.700	0.956
		0.965	0.715	0.715	-13.740	0.715	0.800	0.810	0.950
		1.173	0.409	0.409	0.470	0.409	0.450	0.450	0.950
	2500	0.706	0.894	0.893	44.220	0.893	3.660	3.500	0.962
		0.965	0.810	0.810	49.930	0.810	14.660	15.650	0.936
		1.174	0.733	0.732	132.440	0.733	47.250	48.800	0.948
	5000	0.705	0.893	0.893	13.590	0.894	1.810	1.660	0.960
		0.965	0.810	0.810	46.880	0.810	7.270	7.450	0.944
		1.173	0.732	0.732	4.990	0.732	23.350	21.990	0.952
	10000	0.705	0.893	0.893	-2.080	0.893	8.900	8.000	0.952
		0.965	0.810	0.810	-8.200	0.810	3.640	3.310	0.956
		1.173	0.732	0.732	-20.360	0.732	11.430	10.740	0.944
0.25	2500	0.716	0.927	0.927	6.55	0.927	3.74	3.65	0.954
		0.989	0.683	0.683	-8.02	0.683	3.73	3.43	0.950
		1.206	0.359	0.359	-42.89	0.359	1.66	1.59	0.942
	5000	0.716	0.927	0.927	20.110	0.927	1.860	1.740	0.956
		0.989	0.683	0.683	46.510	0.683	1.830	1.730	0.946
		1.206	0.359	0.359	-17.270	0.359	0.820	0.780	0.952
	10000	0.717	0.927	0.927	14.000	0.927	9.200	8.400	0.972
		0.989	0.683	0.683	25.150	0.683	9.100	8.900	0.952
		1.206	0.358	0.359	-7.000	0.358	4.100	3.900	0.954
	2500	0.716	0.927	0.927	6.550	0.927	3.740	3.650	0.954
		0.989	0.683	0.683	-8.020	0.683	3.730	3.430	0.950
		1.206	0.359	0.359	-42.890	0.359	1.660	1.590	0.942
	5000	0.716	0.890	0.890	17.010	0.890	2.090	1.880	0.956
		0.989	0.802	0.801	56.930	0.801	12.450	12.60	0.956
		1.206	0.720	0.719	70.460	0.719	72.970	77.240	0.946
	10000	0.717	0.890	0.890	22.240	0.890	1.040	9.300	0.954
		0.990	0.802	0.801	73.300	0.801	6.110	6.170	0.946
		1.206	0.720	0.719	61.010	0.719	36.140	35.680	0.954

$\hat{\eta}$: estimated parameters, ModB: model-based variance, EMP: empirical variance, CP: 95% coverage probability. ^{a} : $\times 10^{-5}$.

Table 3.6: Estimated parameters of the model for T_1 across samples sizes (N), dependence levels (τ_k) with 20% censoring (C) treating T_2 as a competing event and vice versa.

τ_k	N	X	$\widehat{S_{T_1}(X)}$	$S_{T_1}(X)$	Bias ^{a}	$S_1^*(X)$	B-Var ^{a}	EMP ^{a}	CP
0.50	2500	0.738	0.916	0.916	58.980	0.916	5.540	5.060	0.960
		1.021	0.639	0.640	-68.590	0.640	3.760	3.670	0.954
		1.232	0.319	0.320	-27.760	0.319	1.310	1.350	0.942
	5000	0.738	0.916	0.916	11.440	0.916	2.730	2.440	0.972
		1.022	0.639	0.639	9.270	0.634	1.860	1.710	0.956
		1.232	0.319	0.319	2.600	0.319	0.640	0.640	0.950
	10000	0.738	0.916	0.916	-13.120	0.916	1.340	1.240	0.948
		1.021	0.639	0.639	1.980	0.639	0.920	0.850	0.946
		1.232	0.319	0.319	-6.310	0.319	0.320	0.340	0.944
	2500	0.737	0.884	0.884	4.680	0.884	5.320	5.280	0.952
		1.021	0.791	0.789	135.630	0.789	71.850	66.320	0.964
		1.232	0.730	0.709	2151.670	0.708	1096.580	1257.780	0.966
	5000	0.738	0.884	0.884	18.660	0.884	2.600	2.370	0.964
		1.022	0.791	0.789	172.070	0.789	33.060	30.790	0.960
		1.232	0.722	0.709	1348.090	0.709	670.130	598.940	0.926
	10000	0.738	0.884	0.884	-28.360	0.884	1.270	1.140	0.948
		1.022	0.790	0.789	41.610	0.789	15.670	14.460	0.960
		1.232	0.717	0.709	812.370	0.709	332.050	300.850	0.956
0.75	2500	0.776	0.893	0.893	-35.280	0.892	9.620	8.720	0.972
		1.050	0.598	0.598	-21.080	0.598	2.990	2.890	0.952
		1.242	0.304	0.305	-72.520	0.304	1.050	1.060	0.928
	5000	0.777	0.893	0.893	18.880	0.893	4.630	3.970	0.966
		1.051	0.597	0.597	35.440	0.597	1.470	1.320	0.954
		1.243	0.304	0.304	-6.910	0.304	0.052	0.049	0.950
	10000	0.776	0.892	0.893	-39.260	0.892	2.240	2.010	0.962
		1.051	0.597	0.597	19.520	0.597	0.730	0.680	0.956
		1.243	0.304	0.304	-0.450	0.304	0.260	0.270	0.948

$\hat{\eta}$: estimated parameters, ModB: model-based variance, EMP: empirical variance, CP: 95% coverage probability. ^{a} : $\times 10^{-5}$

estimator for $S_{T_1}(X)$ at time X closely approximate the disease-specific survival curve for both zero and 50% dependence for the G-copula model with 1000 sample sizes.

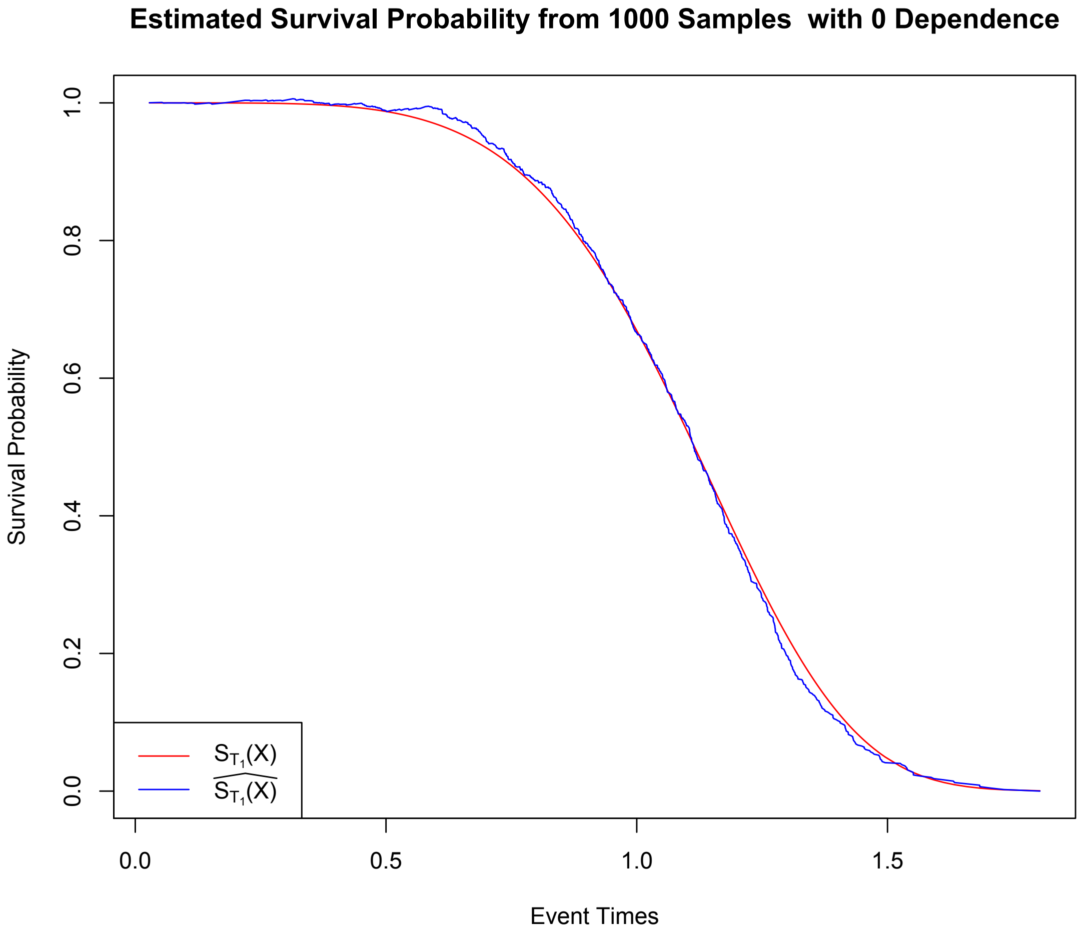

Figure 3.2: Comparison of the estimated event survival probability and the truth for 1000 samples with 15% censoring for 0% dependence structure for Gumbel copula

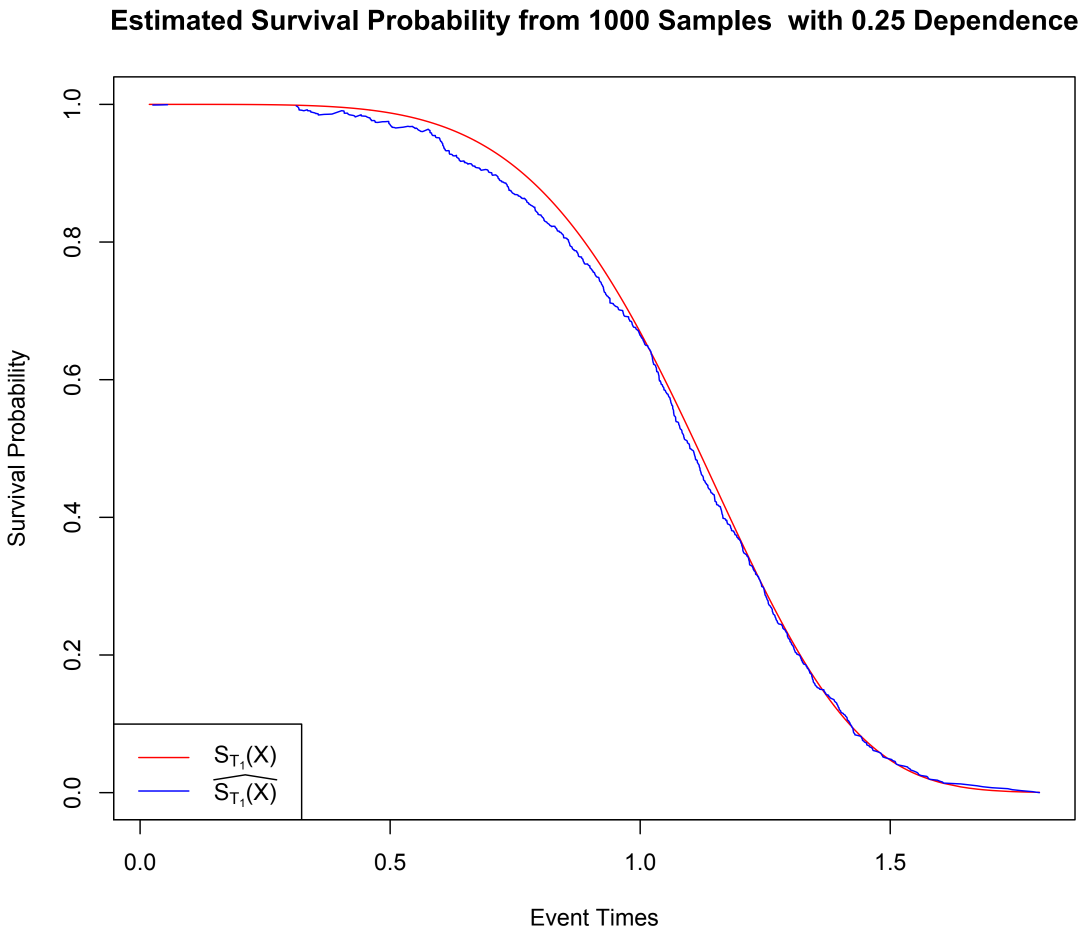

Figure 3.3: Comparison of the estimated event survival probability and the truth for 1000 samples with 15% censoring for 25% dependence structure for Gumbel copula

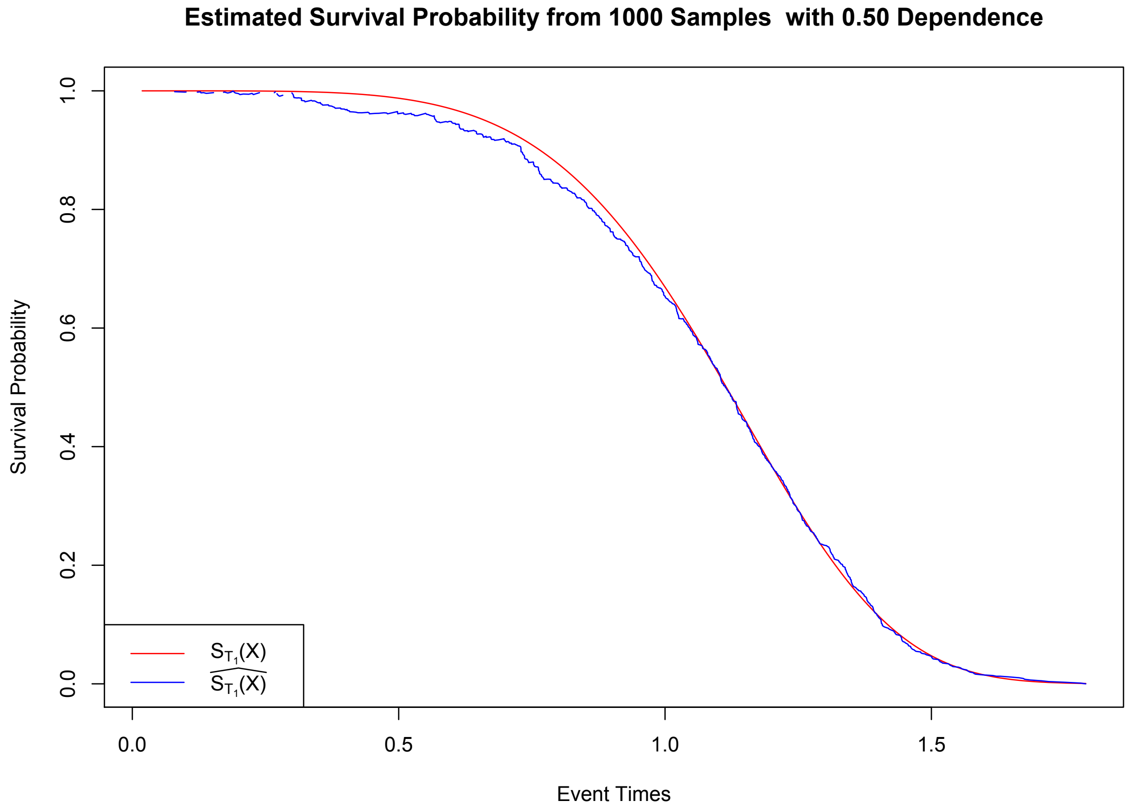

Figure 3.4: Comparison of the estimated event survival probability and the truth for 1000 samples with 15% censoring for 50% dependence structure for Gumbel copula

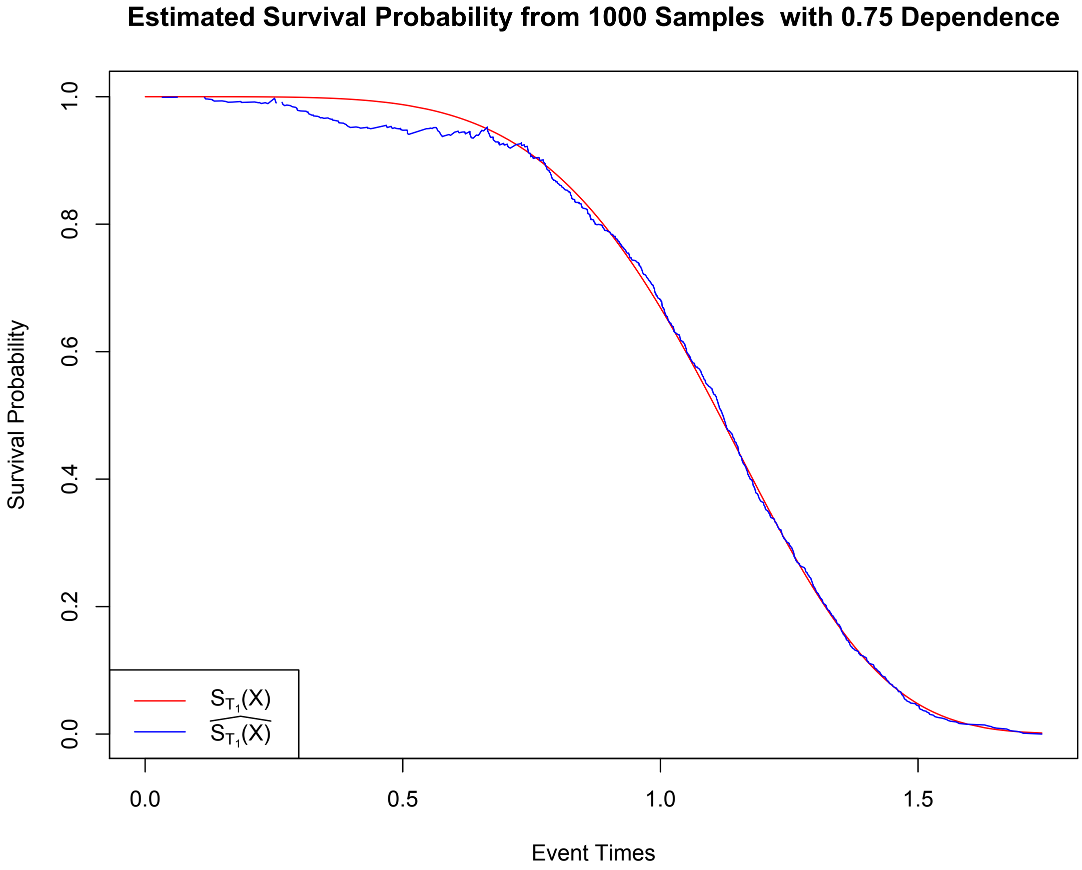

Figure 3.5: Comparison of the estimated event survival probability and the truth for 1000 samples with 15% censoring for 75% dependence structure for Gumbel copula

3.4.3 Application to French Breast Cancer Data

We analyze 24,458 nonmetastatic breast cancer patients from Institut Curie database. These women were between the ages of 18 and 96 years surviving breast cancer in France during 1980 to 2011. Out of the 24,458 breast cancer patients, 9,885 (40.4%) died while 14,573 were alive and administratively censored on December 31^{st} 2011. Five age-specific subgroups were considered for the estimation of relative survival. 3,970 were between the ages of $15-44$, 6,895 between the ages of $45-54$, 6,420 between the ages of $55-64$, 4,675 between the ages of $65-74$ and 2,498 were in the $75-99$ age subgroup category. For each participant, we matched the observed death or censoring time in the disease cohort group with a corresponding time in the healthy reference population on age, sex, and year (date of diagnosis and the date of death or censored) within each follow-up period. The background mortality data from the Human Mortality Database (https://www.mortality.org) was last modified on June 28, 2018. Within each follow-up year, we assumed that $\lambda_P(t)$ is piecewise constant (Dickman, et al., 2004) for each period up to time X. The cumulative hazard for each period based on $\lambda_P(t)$ is calculated from the background survival function at the beginning and end of the period. The cumulative hazard is then used to obtain $\lambda_P(t)$ under the piecewise constant assumption. To mitigate the impact of age and calendar year on potentially dependent censoring by C (Perme et al., 2012), we set $\lambda_{T_2}(t)$ to λ_P(t). $2, 5, 10$, and $15-$year relative survival were estimated nonparametrically for T_1 using a G-copula model with differing levels of dependence specified for the joint distribution of T_1 and T_2. We compared our results with Perme et al. (2012) estimator under the independence of T_1 and T_2.

3.4.4 Extracting $\widehat{S_{T_2}(X)}$ from the Background Population

The distribution of T_2 was extracted from the background population matching in age, dates of diagnosis and death/censored for each participant during each follow-up period. Yearly background data was last modified on June 28^{th} 2018. We assumed that the probability of death within a year is piecewise constant. $\widehat{S_{T_2}(X)} = \exp\{-\widehat{\Lambda(X)}\} = \exp\left(-\int_0^X \widehat{\lambda(s)ds}\right)$. The results for the overall and age group-specific estimates in the following tables.

Tables 3.7 and 3.8 show the estimates of $S_{T_1}(t)$ for cancer mortality both overall and age group specific. The nonparametric estimates under independence are similar to those from the

Pohar-Perme method. This suggests that our estimator is reasonable even under the independence of T_1 and T_2. Cancer survival increases with decreasing dependence. One observes that for a fixed dependence level, younger women tend to have higher cancer survival rates than do older women, with marked reductions for the 65-74 and 75-99 age group categories. There is some instability in survival estimates at 15 years, especially for the older age groups, as evidenced by the large standard errors. This may be due to small numbers of patients at risk at longer follow-up times.

Table 3.7: 2, 5, 10 and 15-yr overall relative survival for French women diagnosed with breast cancer between 1980 and 2011.

τ_k		0.00		0.25		0.50		0.75	
Year	PP^a	$S_{T_1}(t)^a$	SE^b	$S_{T_1}(t)^a$	SE^b	$S_{T_1}(t)^a$	SE^b	$S_{T_1}(t)^a$	SE^b
2	95.6	98.6	15.83	98.5	15.59	98.2	15.11	97.2	13.49
5	84.8	87.5	17.06	87.0	24.20	86.1	15.07	85.1	15.41
10	71.0	73.3	58.37	72.6	48.67	71.6	35.83	71.0	13.09
15	59.5	51.4	07.61	50.0	06.30	48.8	04.57	59.5	03.62

a : $\times 10^{-2}$, b : $\times 10^{-3}$, τ_k: dependence, PP: Pohar-Perme, $S_R(t)$: nonparametric relative survival estimate at year t, SE: standard error for the relative survival estimate.

Table 3.8: 2, 5, 10 and 15-yr age group specific relative survival for French women diagnosed with breast cancer between 1980 and 2011.

τ_k			0.00		0.25		0.50		0.75	
Year	Agegp	PP^a	$S_{T_1}(t)^a$	SE^b	$S_{T_1}(t)^a$	SE^b	$S_{T_1}(t)^a$	SE^b	$S_{T_1}(t)^a$	SE^b
2	15-44	95.8	96.3	03.40	96.3	03.23	96.1	03.34	95.9	03.41
	45-54	97.1	97.4	02.97	97.3	02.98	97.3	02.72	97.2	02.23
	55-64	95.7	97.1	03.21	97.0	03.32	96.7	03.07	96.1	02.85
	65-74	95.1	98.2	07.71	98.0	08.45	97.6	09.08	96.5	09.46
	75-99	91.5	97.8	134.89	97.5	140.56	96.6	145.69	94.4	151.40
5	15-44	81.5	85.5	05.76	85.4	05.61	85.2	05.83	85.1	05.95
	45-54	88.6	89.3	04.76	89.2	04.73	88.9	04.28	88.7	03.92
	55-64	85.8	88.5	07.23	88.0	06.67	87.1	05.77	86.1	04.74
	65-74	84.1	88.9	17.56	88.1	16.85	86.6	14.71	84.7	08.13
	75-99	72.3	83.3	67.77	81.0	72.73	77.4	72.85	73.4	101.83
10	15-44	71.9	72.7	07.82	72.4	07.35	72.1	07.67	71.9	07.42
	45-54	78.3	81.0	06.29	80.3	06.41	79.3	05.54	78.5	05.20
	55-64	73.4	77.2	7.98	76.1	7.43	74.7	06.68	73.6	05.87
	65-74	68.4	76.8	15.04	74.6	12.59	71.5	10.26	68.8	07.36
	75-99	44.6	56.1	78.25	51.4	56.39	47.0	29.79	44.9	12.58
15	15-44	62.5	63.3	08.58	63.0	08.65	62.7	08.42	62.6	08.13
	44-54	70.8	72.0	09.06	71.6	07.17	71.1	06.22	70.8	05.91
	55-64	63.5	71.0	06.22	65.1	13.15	64.1	8.42	63.6	06.55
	65-74	50.3	64.8	53.41	59.7	36.39	54.2	18.38	50.7	09.16
	75-99	19.9	22.6	44.47	20.8	21.50	20.0	12.47	19.9	10.73

a : $\times 10^{-2}$, b : $\times 10^{-3}$, τ_k: dependence, Agegp: Age group, PP: Pohar-Perme, $S_R(t)^a$: nonparametric relative survival estimate at year t, SE: standard error for the relative survival estimate.

3.5 Discussion and Conclussion

We investigated the precision of our estimator under different dependence structures and proposed sensitivity analysis as a practical solution to identifiability constraints of dependence. The nonparametric model for disease-specific mortality is reasonable for complex failure patterns and flexible dependence levels as is the case in competing risk setting. Our nonparametric estimator has a dual meaning of net survival probability under independence assumption and crude survival probability or crude probability of death under the dependence assumption. This estimator is useful to both practitioners who prefer either net survival probability or crude survival and or crude probability of death for determining prognosis. The case for covariate effects is currently being developed to understand the contribution of other risk factors in relative survival analysis.

The key point in this paper is the estimation of the distribution of latent failure time for a specific disease. This endpoint under the independence of T_1 and T_2 has been advocated by many practitioners, as it eliminates the impact of other cause mortality on the risk of disease-specific mortality assumed in a hypothetical world where mortality is due to disease of interest. It is also useful in comparison of survival across groups or populations with different background mortality. In contrast to net survival, some practitioners prefer crude probability of survival to net survival because it accommodates deaths from other causes and presents cancer survival in the real world, where the patient may experience mortality in the presence of competing causes. Our estimator not only provides a practical alternative to Perme et al. (2012) method under the independence assumption but also a useful estimator under the dependence assumption for practitioners who prefer crude survival to net survival.

CHAPTER 4: Covariate Effect for Dependence Competing Risk in Relative Survival Analysis

4.1 Introduction

Biomedical research often have competing risks where one event type censors other mutually exclusive events. Patients can potentially experience an event from any of the multiple failures particularly in cancer registry data. For example, in following patients after cancer diagnosis, a patient may commit suicide or die from the cancer under study or may die from other causes. In such competing risk setting, standard disease-specific statistical analysis and interpretation differ from survival analysis with only a single cause of failure (Dignam et al., 2012).

In determining credible prognosis for disease-specific mortality in cancer studies using registry data, accurate documentation of cause of death (Percy et al., 1981; Welch and Black, 2002; Mieno et al., 2016) and appropriate statistical methodology (Caplan et al., 1994; Gooley et al., 1999; Williamson et al., 2007; Dignam and Kocherginsky, 2008) underpinning the analysis are required when comparing groups under different populations. A challenge to the disease-specific analysis based on standard methods is that cause of death information may be missing and or unreliable for meaningful conclusions to be drawn. As a result, several modeling approaches are available to evaluate the relationship between the covariates and disease-specific failures.

The standard hazard or the cumulative incidence function for a specific failure type in competing risks analysis is used to evaluate the influence of covariate in disease-specific survival under the independence assumption (net survival or crude survival probability). Such models are unviable for competing risk analysis data without cause of death information. Existing methods for disease-specific survival analysis employing different competing risks models on the same data for hazard ratios estimation, can differ substantially and may lead to different or even seemingly contradictory inferential conclusions (Dignam et al., 2012). These methods according to (Denham et al., 1996;

Chappell, 2012) may only illuminate one important aspect of the data while possibly obscuring others.

The issue is more complicated and renders current perspective on estimated disease-specific survival probabilities useless under the current methodologies for registry data when cause of death information is missing and or unreliable because of difficulty associated with distinguishing disease-specific mortality from other cause mortality. Even within one month of cancer diagnosis and cancer directed surgery, 41% of all deaths that occurred were missclassified and not attributable to the coded cancer in the registry Welch and Black (2002). Sometimes, when reliable cause of death information is available, it is often located in separate databases, which may be costly to obtain and difficult to link with registry data.

Suppose that $T = \min\{T_k : \ k = 1, 2, 3, \cdots, K\}$ is the potentially unobservable failure time and $\varepsilon = \{k : T = T_k\}$ the failure type with $T_1, \cdots, T_K, \ K \in \mathbf{N}$ being the latent failure times associated with the K failure types. When $K = 2$ and $\varepsilon = 1$ implies death from cancer and $\varepsilon = 2$ implies death from other competing causes. Under dependent competing risks, where T_1 and T_2 are not independent, standard methods for independently right censored survival data without competing risks cannot be used to make inference about disease-specific survival. Thus, the Kaplan-Meier (Kaplan and Meier, 1958) estimator estimates a function of the disease-specific hazard function, defined in section 4.2. The logrank test (Bland and Altman, 2004) assesses group differences between the disease-specific hazard function, while the standard proportional hazards model (Cox, 1972) formulates the effects of covariates on the disease-specific hazard function. The cumulative incidence function, defined in section 4.2, gives disease-specific survival in the presence of competing events. This quantity has been widely adopted in applications, with the Aalen-Johanson estimator (Aalen and Johansen, 1978), Gray's test (Gray et al., 1988), and the Fine-Gray model (Fine and Gray, 1999), providing analogs to the Kaplan-Meier curve, the logrank test, and the proportional hazards model for the cumulative incidence function. Without cause of death information, these methods are inapplicable. This paper focused on estimating the effect of covariates on disease-specific survival analysis under both independence and dependence assumption relating to T_1 and T_2 respectively.

Without reliable cause of death information for disease-specific survival, relative survival methods have being proposed. Relative survival, $S_R(t)$ is the ratio of the observed survival rate in a group of cancer patients, during a specified period, to the expected survival rate in a healthy reference population (Ederer, 1961). Mathematically and at time t,

$$S_R(t) = \frac{S_O(t)}{S_P(t)} \tag{4.24}$$

with $S_O(t)$ being the survival probability associated with an individual in the registry and $S_P(t)$ the expected survival probability derived from mortality tables. Existing literature has focused exclusively on the estimation of $S_R(t)$ under the independence of T_1 and T_2, which implies that $S_O(t) = S_{T_1}(t) \cdot S_{T_2}(t),\ S_P(t) = S_{T_2}(t)$ and $S_R(t) = S_{T_1}(t)$ where $S_{T_1}(t)$ and $S_{T_2}(t)$ are the survival probabilities corresponding to T_1 and T_2 respectively. The relationship (2.1) can be rewritten in terms of hazard functions as $\lambda_O(t) = \lambda_E(t) + \lambda_P(t)$ (Cronin and Feuer, 2000), where $\lambda_O(t)$ is the hazard in the disease registry, $\lambda_E(t)$ is the so called excess hazard among the cancer cohort, and $\lambda_P(t)$ is the hazard from mortality tables. Under independence, $\lambda_E(t) = \lambda_{T_1}(t)$ and $\lambda_P(t) = \lambda_{T_2}(t)$, where $\lambda_{T_j}(t) = \frac{-dlogS_{T_j}(t)}{dt}, j = 1, 2$, are the net hazard functions for cancer and other cause mortality. The disease-specific survival probability $S_{T_1}(t)$ under the independence assumption is the target of standard relative survival analysis and corresponds to a hypothetical population in which death from competing causes does not exist. This quantity differs from the cumulative incidence function which is commonly used to quantify disease-specific survival in analyses with cause of death information. Under the dependence assumption, $S_{T_1}(t)$ is relative survival in the real world where competing risks exist simultaneously with the disease. Some authors called this estimator crude survival or crude probability of death.

Relative survival is employed extensively for the comparison of cancer survival in cohort groups or populations, or for evaluating changes in survival over time and for exploring potential risk factors for disease-specific mortality. Relative survival methods under the independence assumption was pioneered by Berkson and Gage (1950) and (Ederer, 1961) for nonparametric estimation of $S_{T_1}(t)$. Variants of these earlier methods were introduced by Hakulinen (1982) with the aim of addressing the bias due to heterogeneity of patient withdrawal within subgroups. Perme et al. (2012) demonstrated that these classical methods may be biased under certain censoring patterns. For example, in population comparisons, such bias may arise from unmeasured covariates affecting the cancer cohort group and the reference population from which rates of expected mortality are drawn.

In order to understand the contribution of the risk factors in registry data, (Hakulinen and Tenkanen, 1987) adapted the standard proportional hazard regression model to estimate relative survival rates via generalized linear models (GLIM). While in comparing groups under different populations, (Dickman et al., 2004) investigated the covariate effects by comparing four standard regression methods. They asserted that excess hazard estimated from the underlying population depends on characteristics such as age, sex, (supported by (Hakulinen and Tenkanen, 1987)) and period but not on other covariates such as stage or histology of the cancer. All the above methods are valid only under the independence of T_1 and T_2.

We relaxed the independence assumption (de Lacerda et al., 2019) and formulate the dependence between the distributions of the latent failure times for death from disease and death from competing causes using copula models (Deheuvels, 1978). In particular, we proposed assessing prognostic factors for disease-specific mortality via dependence competing risk regression model for estimating disease-specific survival. The standard copula function generates a joint distribution for the two event times, taking as input their marginal distributions. Copulas generally allow a spectrum of dependence structures and have been employed widely in survival analysis, including bivariate event times (Oakes, 1982), competing risks with known cause of failure (Heckman and Honoré, 1989), and semi-competing risks where one event time censors the other but not vice versa (Fine et al., 2001).

In this paper, we characterized the dependence model for competing risks data from disease registries without reliant on cause of death information which is either missing or unreliable. Because the joint distribution of the distributions of the latent failure times is nonparametrically nonidentifiable (Tsiatis, 1975), we treat the copula function as known. The marginal distribution of the time to disease-specific death is modelled via the disease-specific hazard parametrically with the distribution of other cause mortality drawn from the reference population. Likelihood-based inference and interpretation is proposed. Due to identifiability constraints and the unverifiable nature of the joint distribution of the distribution of the latent failure times, a sensitivity analysis is suggested in which disease-specific survival is estimated across a spectrum of dependence structures specified via the copula function. To our knowledge, this is the first attempt in accommodating dependence in competing risk regression model in relative survival analysis.

The rest of this paper proceeds as follows. In section 4.2, we present the data and copula model formulation for competing risks data. Section 4.3 describes the likelihood parametric regression estimation and inference procedure without cause of death information, as well as the proposed sensitivity analysis. In section 4.4, we present the numerical illustrations including simulation results and application to French breast cancer data. Section 4.5 discusses and concludes the paper.

4.2 Data and Model

In competing risk studies, the standard endpoints for disease-specific survival are the disease-specific hazard and the disease-specific cumulative function. The disease-specific hazard, $\lambda_k(t)$ is the instantaneous failure rate for occurrence of mutually exclusive events $\varepsilon = k$ at time t (Prentice et al., 1978).

$$\lambda_k(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t, K = k | T > t)}{\delta t} \tag{4.25}$$

While the cumulative incidence function $C_k(t)$ is the proportion of patients who died from cause k by time t in the presence of patients who might die from other causes. The disease-specific failure probability can be expressed as $C_k(t) = P(T \leq t: \ \varepsilon = k) = \int_0^t \lambda_k(s) \cdot S(s) ds = \int_0^t \lambda_k(s) \cdot \exp\{-\Lambda(t)\}$ where $S(t) = P(T > t)$ is the overall survival probability. Standard competing risks methods with known cause of failure focus on estimation of $\lambda_k(t)$ and $C_k(t)$.

In the absence of cause of death information, the registry data is simply time to event data from any cause, T, which may be right censored by loss to follow-up. Suppose C is the time to right censoring, with the usual assumption being that T and C are independent, then the observed data consist of (X_i, δ_i) where $X_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$ with T_i and C_i being the failure and censoring times on individual $i = 1, 2, 3, \cdots, n$. Relative survival methods employing such data do not focus on the traditional competing risks endpoints $\lambda_k(t)$ and $C_k(t)$ but rather on the latent failure time distributions $S_{T_1}(t)$ and $S_{T_2}(t)$.

The dependence between the distributions of T_1 and T_2 was modelled using copula function as it completely describe the dependence structure and provide scale invariant measures of association (Venter, 2002; Müller, 1996; Bäuerle and Müller, 1998; Denuit et al., 1999).

To formulate this dependence, suppose ψ is a generator function defined such that $\psi : [0, 1] \rightarrow [0, +\infty]$ with independent marginal distributions, $u_j = P(T_j \leq t_j) = F_{T_j}(t_j) = 1 - S_{T_j}(t_j) \ \forall j \in (1, 2)$.

Then, the copula model for the distributions of T_1 and T_2 is:

$$C(u_1, u_2) = P(T_1 \leq t_1, T_2 \leq t_2) = \psi\left(\psi^{-1}(u_1) + \psi^{-1}(u_2)\right) = F_{T_1,T_2}(t_1, t_2)$$

where ψ^{-1} is the inverse of ψ and ψ satisfies the Laplace-Stiltjes transform and Bernstein, (1929) theorem. McNeil and Nešlehová, (2009) showed that the generator function ψ is monotone for non-negative random variables with $\psi(0) = 1$, $\psi'(\cdot) < 0$ and $\psi''(\cdot) < 0$. We characterize dependence between the distribution of the latent failure times using Kendall's tau (τ_k) correlation coefficient. Genest and MacKay (1986) estabished the connection between Kendall's tau and the generator function ψ as:

$$\tau_k = 1 + 4\int_0^1 \frac{\psi^{-1}(u)}{\psi^{-1}(u)'}du = 1 - 4\int_0^\infty u(\psi(u))^2 du$$

with $\psi^{-1'}$ being the derivative of ψ^{-1}. While in theory, any copula and dependence measure may be used to link the marginal distributions of T_1 and T_2, in this paper, we focus on a popular Archimedean copula, indexed by a single dependence parameter θ having simple interpretations. The Gumbel copula is:

$$C(u_1, u_2) = \exp\left[-\{(-log(u_1))^\theta + (-log(u_2))^\theta\}^{\frac{1}{\theta}}\right] \tag{4.26}$$

with $\theta \in (1, +\infty)$. A special case of independence copula model: $C(u_1, u_2) = u_1 \cdot u_2$ is obtained when $\theta = 1$ for Gumbel copula with $\tau_k = 1 - \frac{1}{\theta}$.

4.3 Likelihood Estimation and Inference

In this section, we formulate our model using the disease-specific hazard while incorporating the covariate effects via the accelerated failure time model for any potentially dependent latent failure times T_1 and T_2. The all-cause survival function (observed survival $S_O(t)$), $T = \min(T_1, T_2)$ at time t without covariates is:

$$S_O(t) = S_T(t) \quad = \quad S_{T_1}(t) + S_{T_2}(t) - 1 + F_{T_1,T_2}(t, t) \tag{4.27}$$

The conventional observed additive hazard model defined in section 4.1 for the estimation of relative survival can be adapted to include covariates. Dickman et al. (2004) introduced the covariate effect through the hazard model using an exponential form: $\lambda_O(t, Z) = \lambda_E(t) + exp\{\beta Z\}$. In this study, we introduced covariates into our model through the Accelerated Failure Time (AFT) model format. Thus, $X = Z \cdot \beta + \epsilon$ where Z is the set of covariates such as (age, sex, calandar/period), β the parameters to be estimated and ϵ is the randomness associated with the covariates. The above model in equation 4.27 can be written to accommodate the stratifying covariate Z as:

$$S_O(t, Z) = S_T(t, Z) \quad = \quad S_{T_1}(t, Z) + S_{T_2}(t, Z) - 1 + F_{T_1,T_2}(t, Z, t, Z) \tag{4.28}$$

Under the independence of T_1 and T_2 the above equation degenerates to:

$$\begin{aligned} S_T(t, Z) &= S_{T_1}(t, Z) \cdot S_{T_2}(t, z) = exp\left\{\int_0^t \lambda_O(s, Z)\right\} ds \\ &= exp\left\{\int_0^t \lambda_{T_1}(s, Z)\right\} ds \cdot exp\left\{\int_0^t \lambda_{T_2}(s, Z)\right\} ds \end{aligned} \tag{4.29}$$

The additive hazard model is generally biologically more plausible for population-based cancer survival studies and provide a better estimate to the data than multiplicative models (Bolard, et al.,2001). The corresponding density function for T is:

$$f_O(t, Z) = f_T(t, Z) = f_{T_1}(t, Z) + f_{T_2}(t, Z) - f_{T_1,T_2}\,(t, Z, t, Z) \tag{4.30}$$

where $f_{T_j}(t, Z) = \frac{dF_{T_j}(t,Z)}{dt}$, and $f_{T_1,T_2}(t, Z) = \frac{dF_{T_1,T_2}(t,Z,t,Z)}{dt}$. If censoring of T by C is noninformative, then the likelihood contribution for an individual i is:

$$L_i = f_{X_i,Z_i,\Delta_i}(X_i, Z_i, \delta_i) = [f_T(X_i, Z_i)]^{\delta_i}[S_T(X_i, Z_i)]^{1-\delta_i} \tag{4.31}$$

In specifying parametric model for $F_{T_1}(t, Z)$ with parameter of interest $\eta = (\lambda, \alpha, \beta_1, \beta_2, \beta_3)$, and from equation (4.30), the full log-likelihood function based on n independent observations is:

$$\begin{aligned} l(\mathbf{X}, \mathbf{Z}, \Delta|\eta) &= \sum_{i=1}^{n} (\delta_i * \log f_T(X_i, Z_i) + (1-\delta_i) * \log S_T(X_i, Z_i)) \\ &= \sum_{i=1}^{n} \delta_i * \log\left[f_{T_1}(X_i, Z_i) + f_{T_2}(X_i, Z_i) - f_{T_1,T_2}(X_i, X_i|Z_i)\right] + \\ & \quad \sum_{i=1}^{n} (1-\delta_i) * \log\left[S_{T_1}(X_i, Z_i) + S_{T_2}(X_i, Z_i) - 1 + F_{T_1,T_2}(X_i, X_i|Z_i)\right] \end{aligned} \tag{4.32}$$

where $(\mathbf{X}, \mathbf{Z}, \Delta) = (X_i, Z_i, \Delta_i, i = 1, 2, 3, \cdots, n)$ with $S_{T_1}(t, Z|\eta) = \int_t^\infty f_{T_1}(s|\eta)ds$ and the distribution of T_2 is extracted from the reference population and assumed known with the usual assumption that disease-specific death is negligible in the reference population as illustrated in the French breast cancer data analysis in section 4.2.

For a pre-specified dependence structure, the copula distribution linking $F_{T_1}(t, Z)$ and $F_{T_2}(t, Z)$ may be specified using simple parametric copula models such as the Gumbel copula with Kendall's tau ($\tau_k = 1 - \frac{1}{\theta}$). The Gumbel copula exhibit tail behaviour that mimic the mortality trend observed in the cancer registry data. In the numerical illustrations, T_1 was assumed to follow a transformed Weibull distribution. The standard Weibull distribution with parameter $w_j = (\lambda_j, \alpha_j)$ has a probability density function $f_{T_j}(t|w_j) = \frac{\alpha_j}{\lambda_j}\left(\frac{t}{\lambda_j}\right)^{\alpha_j - 1} \exp\left\{-\left(\frac{t}{\lambda_j}\right)_j^{\alpha}\right\}$ and is versatile by accommodating a wide range of hazard shapes. The bivariate joint distribution and density functions for the Gumbel copula incoporating the covariate Z are:

$$\begin{aligned} F_{T_1,T_2}(t, t, Z|\eta) &= \exp\left\{-\left((-log\,(u_1))^\theta + (-log(u_2))^\theta\right)^{\frac{1}{\theta}}\right\} \\ f_{T_1,T_2}(t, t, Z|\eta) &= F_{T_1,T_2}(t, t, Z|\eta) \cdot \left(\left(-\log\,(u_1)^\theta\right) + \left(-log\,(u_2)^\theta\right)\right)^{\frac{1}{\theta}-1} \\ &\times \left(\left(-\log(u_1)^{\theta-1} \cdot \frac{f_{T_1}(t, Z|\eta)}{u_1}\right) + \left(-log\,(u_2)^{\theta-1} \cdot \frac{f_{T_2}(t, Z|\eta)}{u_2}\right)\right) \end{aligned} \tag{4.33}$$

with $u_1 = F_{T_1}(t, Z|\eta), u_2 = F_{T_2,Z}(t, Z)$. The maximum likelihood estimator (MLE) for η was implemented using Nelder-Mead algorithm (Nelder and Mead, 1965). There is evidence of instability for small sample sizes and larger censoring proportion (Adatorwovor et al., 2020). The usual regularity conditions for the MLE holds. The estimator converges in probability, in that $\hat{\eta} \xrightarrow{P} \eta$ and is asymptotically normal, $\widehat{\eta} \sim N\left(\eta, I_O(\eta)^{-1}\right)$ with variance estimated using the inverse of the observed information matrix $(I_O(\eta)^{-1})$ evaluated at the MLE, $\hat{\eta}$. The observed information matrix is:

$$\begin{aligned} I_O(\eta) &= \frac{\partial^2 l(\eta|\mathbf{X}, \Delta, \mathbf{Z})}{\partial\eta\partial\eta^T} \\ &= \sum_{i=1}^{n}\left\{\frac{\delta_i \cdot [f_T(X_i, Z_i)] \cdot \left\{\frac{\partial}{\partial\eta} f_T(X_i, Z_i)\right\}^T \left\{\frac{\partial}{\partial\eta}\left[f_T(X_i, Z_i)\right]\right\}}{[f_T(X_i, Z_i)]^T [f_T(X_i, Z_i)]}\right\} + \\ &\quad \sum_{i=1}^{n}\left\{\frac{(1-\delta_i) \cdot [S_T(X_i, Z_i)] \cdot \left\{\frac{\partial}{\partial\eta}[S_T(X_i, Z_i)]\right\}^T \left\{\frac{\partial}{\partial\eta}[S_T(X_i, Z_i)]\right\}}{[S_T(X_i, Z_i)]^T [S_T(X_i, Z_i)]}\right\} \end{aligned} \tag{4.34}$$

A sensitivity analysis is conducted across spectrum of assumed dependence structures to address the nonidentifiability and unverifiable constraints of the dependence structure observed in the registry data between the distributions of T_1 and T_2. Each level of dependence represent the varying levels of dependent competing mortality possibly observed in registry data. For each dependence structure, we estimate η with $\hat{\eta}$ and compute $F_{T_1}(t, Z|\hat{\eta})$ to estimate relative survival. The corresponding standard errors are obtained as the square root of the Delta method variance: $Var(\widehat{S_{T_1}(X, Z)}) = g(\widehat{S_{T_1}(X, Z)}) \cdot I_O(\hat{\eta})^{-1} \cdot g^T(\widehat{S_{T_1}(X, Z)})$ where $g(\eta)$ is the derivative of $S_{T_1}(t, Z|\eta)$ with respect to η. Due to the complex nature of the likelihood, numerical approximation is used to estimate the information matrix in the numerical illustrations in section 4.4.

4.4 Numerical Illustrations

4.4.1 Simulation Studies

We present simulation studies to illustrate covariate effects both under the independence (net survival) and dependence assumptions (crude probability). To incorporate covariates in the model using AFT, a transformation was required (see appendix) to facilitate interpretaion where a form of an extreme value distribution (EVD) was implemented. For example, when the shape parameter is zero, then the resulting distribution is the Gumbel extreme value distribution. We employ this technique to evaluate the performance of our covariate method based on simulated data that mimic the French breast cancer data set for sample sizes; 1000, 2500 and 5000 with 500 replications. The latent failure times were generated from the Weibull distribution ($T_j \sim Weibull(\alpha_j, \lambda_j)$) with probability density function defined above in section 4.3. The parameters for the transformed Weibull (Gumbel) distribution for T_1 were $\lambda_1 = 1.200$, $\alpha_1 = 3.000$, $\beta_1 = 2.000$, $\beta_2 = 3.000$, $\beta_3 = 4.000$ while those for T_2 were $\lambda_2 = 2.100$ and $\alpha_2 = 5.00$. We estimated $\lambda_1,\ \alpha_1,\ \beta_1,\ \beta_2,\ \beta_3$ for T_1, while $\lambda_2,\ \alpha_2$ are assumed known for T_2. Noninformative right censoring times were generated from a uniform distribution $(0, \gamma)$, where γ was chosen for 15% censoring. The dependence was chosen for the Gumbel copula with Kendall's tau, $\tau_k = 0,\ 0.25,\ 0.50,\ 0.80$, and 0.90. Whenever possible, initial starting parameter values were randomly chosen from uniform distributions, with multiple starting values wherever possible as described in section 4.3.

Tables 4.9, 4.10 and 4.11 show the results for estimation of the model for T_1 treating T_2 as a competing event. We observe small bias decreasing to zero for increasing sample size across each of the dependence levels for 15% censoring. The empirical variance and the model based variance were similar and the coverage is close to the nominal 0.95 level, particularly at larger sample sizes. The empirical variance decreases as the sample size increases at roughly the expected root n rate.

4.4.2 Application to French Breast Cancer Data

Data were obtained from Institue Curie breast cancer database in France for cancer registrations between 1980 and 2011. This database contains records from $24,458$ non metastatic breast cancer patients between the ages of 18 and 96 years and treated at the Institut Curie. Out of the $24,458$ breast cancer patients, $9,885$ (40.4%) died while $14,573$ were alive and administratively censored on December 31^{st} 2011. Five age group categories were considered for the estimation of relative survival. $3,970$ were between the ages of $15-44$, $6,895$ between the ages of $45-54$, $6,420$ between the ages of $55-64$, $4,675$ between the ages of $65-74$ and $2,498$ were in the $75-99$ age group category. We matched the observed death or censoring time in the disease cohort group with a corresponding time in the healthy reference population on age, sex, and year (date of diagnosis and the date of death or censored) for each participant and for each follow-up period. The background mortality data from the Human Mortality Database (https://www.mortality.org) was last modified on June 28, 2018. Within each follow-up year, we assumed that $\lambda_P(t)$ is piecewise constant (Dickman, et

Table 4.9: Estimated parameters of the model for T_1 across samples sizes (N), and dependence levels (τ_k) with 15% censoring treating T_2 as a competing event.

τ_k	N	$\widehat{\eta}$	Est.	Bias^a	Mod-B^a	EMP^a	CP
0.00	1000	$\widehat{\lambda}$	1.195	-5.338	1.479	1.469	0.933
		$\widehat{\alpha}$	3.000	-0.746	12.073	9.927	0.980
		$\widehat{\beta}_1$	2.001	1.310	22.131	19.218	0.961
		$\widehat{\beta}_2$	3.000	0.441	1.855	1.969	0.935
		$\widehat{\beta}_3$	4.000	-0.340	8.808	8.392	0.955
	2500	$\widehat{\lambda}$	1.197	-3.227	0.592	0.578	0.955
		$\widehat{\alpha}$	2.998	-2.047	4.820	5.011	0.941
		$\widehat{\beta}_1$	2.005	5.020	8.846	8.753	0.967
		$\widehat{\beta}_2$	3.000	-0.153	0.742	0.793	0.945
		$\widehat{\beta}_3$	3.998	-1.522	3.512	3.845	0.934
	5000	$\widehat{\lambda}$	1.198	-2.204	0.297	0.307	0.941
		$\widehat{\alpha}$	3.002	2.211	2.407	2.419	0.949
		$\widehat{\beta}_1$	1.998	-1.556	4.427	4.473	0.947
		$\widehat{\beta}_2$	3.001	1.110	0.369	0.350	0.953
		$\widehat{\beta}_3$	4.000	-0.349	1.756	1.976	0.934
0.25	1000	$\widehat{\lambda}$	1.194	-5.858	1.365	1.320	0.949
		$\widehat{\alpha}$	3.002	2.414	12.179	11.157	0.965
		$\widehat{\beta}_1$	2.001	1.225	22.468	20.507	0.959
		$\widehat{\beta}_2$	3.001	0.839	1.879	1.800	0.963
		$\widehat{\beta}_3$	3.997	-3.404	8.919	7.741	0.971
	2500	$\widehat{\lambda}$	1.198	-1.940	0.550	0.564	0.947
		$\widehat{\alpha}$	2.999	-0.633	4.888	4.233	0.961
		$\widehat{\beta}_2$	2.000	-0.354	9.031	8.405	0.955
		$\widehat{\beta}_2$	3.001	1.145	0.754	0.781	0.949
		$\widehat{\beta}_3$	4.003	2.842	3.579	3.581	0.955
	5000	$\widehat{\lambda}$	1.201	1.320	0.276	0.263	0.963
		$\widehat{\alpha}$	3.002	2.406	2.456	2.012	0.982
		$\widehat{\beta}_1$	2.000	0.071	4.526	4.178	0.965
		$\widehat{\beta}_2$	3.000	0.037	0.376	0.357	0.963
		$\widehat{\beta}_3$	3.997	-2.871	1.795	1.569	0.967

Est.: Estimated parameter value, Mod-B: Model-based variance, EMP: Empirical variance, CP: 95% Coverage, ^{a} : $\times 10^{-3}$

Table 4.10: Estimated parameters of the model for T_1 across samples sizes (N), and dependence levels (τ_k) with 15% censoring treating T_2 as a competing event.

τ_k	N	$\widehat{\eta}$	Est.	Bias^a	Mod-B^a	EMP^a	CP
0.50	1000	$\widehat{\lambda}$	1.195	-5.156	1.172	1.190	0.945
		$\widehat{\alpha}$	3.016	5.696	11.730	10.969	0.961
		$\widehat{\beta_1}$	1.998	-1.914	21.809	19.788	0.955
		$\widehat{\beta_2}$	2.999	-0.716	1.825	1.964	0.945
		$\widehat{\beta_3}$	3.996	-3.771	8.670	7.539	0.963
	2500	$\widehat{\lambda}$	1.198	-1.748	0.471	0.497	0.943
		$\widehat{\lambda}$	3.000	-0.144	4.697	4.272	0.961
		$\widehat{\beta_1}$	1.998	-1.818	8.759	8.511	0.953
		$\widehat{\beta_2}$	3.001	0.553	0.730	0.771	0.943
		$\widehat{\beta_3}$	4.003	3.065	3.469	3.566	0.951
	5000	$\widehat{\lambda}$	1.202	1.618	0.236	0.239	0.953
		$\widehat{\alpha}$	3.002	1.858	2.361	1.910	0.973
		$\widehat{\beta_1}$	2.002	1.838	0.004.387	3.965	0.959
		$\widehat{\beta_2}$	3.000	-0.162	0.365	0.348	0.961
		$\widehat{\beta_3}$	3.997	-3.346	1.740	1.545	0.955
0.75	1000	$\widehat{\lambda}$	1.190	-10.061	1.057	8.958	0.939
		$\widehat{\alpha}$	3.017	17.474	10.610	31.997	0.953
		$\widehat{\beta_1}$	2.017	17.248	19.743	88.790	0.939
		$\widehat{\beta_2}$	3.002	1.750	1.656	5.472	0.943
		$\widehat{\beta_3}$	4.001	1.357	7.852	46.490	0.959
	2500	$\widehat{\lambda}$	1.198	-1.571	0.425	0.446	0.949
		$\widehat{\lambda}$	2.999	-1.424	4.241	4.285	0.961
		$\widehat{\beta_1}$	2.000	-0.168	7.933	8.348	0.949
		$\widehat{\beta_2}$	3.000	0.121	0.662	0.710	0.943
		$\widehat{\beta_3}$	4.003	2.847	3.141	3.430	0.951
	5000	$\widehat{\lambda}$	1.194	-6.115	0.212	9.103	0.934
		$\widehat{\alpha}$	3.021	21.821	2.119	66.091	0.962
		$\widehat{\beta_1}$	2.006	6.022	3.949	18.767	0.951
		$\widehat{\beta_2}$	2.994	-5.800	0.328	12.658	0.944
		$\widehat{\beta_3}$	4.001	1.448	1.566	8.702	0.955

Est.: Estimated parameter value, Mod-B: Model-based variance, EMP: Empirical variance, CP: 95% Coverage, ^{a} : $\times 10^{-3}$

Table 4.11: Estimated parameters of the model for T_1 across samples sizes (N), and dependence levels (τ_k) with 15% censoring treating T_2 as a competing event.

τ_k	N	$\widehat{\eta}$	Est.	Bias^a	Mod-B^a	EMP^a	CP
0.80	1000	$\widehat{\lambda}$	1.194	-6.077	1.402	1.332	0.949
		$\widehat{\alpha}$	3.004	4.121	12.194	11.843	0.957
		$\widehat{\beta_1}$	2.000	-0.024	22.445	21.955	0.959
		$\widehat{\beta_2}$	3.001	1.221	1.877	1.756	0.967
		$\widehat{\beta_3}$	3.995	-4.870	8.914	7.954	0.969
	2500	$\widehat{\lambda}$	1.198	-1.946	0.564	0.559	0.949
		$\widehat{\alpha}$	2.999	-1.250	4.891	4.120	0.961
		$\widehat{\beta_3}$	2.000	-0.0917	9.021	8.304	0.955
		$\widehat{\beta_3}$	3.001	1.236	0.754	0.771	0.953
		$\widehat{\beta_3}$	4.003	3.013	3.575	3.522	0.957
	5000	$\widehat{\lambda}$	1.201	1.202	0.283	0.268	0.965
		$\widehat{\alpha}$	3.003	3.054	2.457	1.892	0.983
		$\widehat{\beta_1}$	1.999	-0.813	4.522	4.053	0.967
		$\widehat{\beta_2}$	3.000	0.017	0.376	0.361	0.958
		$\widehat{\beta_3}$	4.000	-3.017	1.793	1.567	0.963
0.90	1000	$\widehat{\lambda}$	1.195	-5.3751	1.459	1.337	0.955
		$\widehat{\alpha}$	3.003	2.612	12.146	9.395	0.993
		$\widehat{\beta_1}$	2.002	1.581	22.313	19.594	0.977
		$\widehat{\beta_2}$	3.001	1.914	1.867	1.668	0.966
		$\widehat{\beta_3}$	3.995	-5.226	8.860	7.394	0.960
	2500	$\widehat{\lambda}$	1.198	-1.812	0.587	0.555	0.946
		$\widehat{\alpha}$	2.999	-1.062	4.871	3.626	0.994
		$\widehat{\beta_3}$	2.000	-0.256	8.963	7.764	0.965
		$\widehat{\beta_3}$	3.001	1.122	0.748	0.768	0.948
		$\widehat{\beta_3}$	4.002	2.547	3.552	3.427	0.956
	5000	$\widehat{\lambda}$	1.201	1.089	0.294	0.277	0.956
		$\widehat{\alpha}$	3.002	2.291	2.444	1.696	0.992
		$\widehat{\beta_1}$	1.999	-0.598	4.493	3.788	0.979
		$\widehat{\beta_2}$	3.000	0.335	0.374	0.378	0.941
		$\widehat{\beta_3}$	3.997	-2.614	1.780	1.556	0.962

Est.: Estimated parameter value, Mod-B: Model-based variance, EMP: Empirical variance, CP: 95% Coverage, ^{a} : $\times 10^{-3}$

al., 2004) for each period up to time X. The cumulative hazard for each period based on $\lambda_P(t)$ is calculated from the background survival function at the beginning and end of the period. The cumulative hazard is then used to obtain $\lambda_P(t)$ under the piecewise constant assumption. The goal of matching in determining $\lambda_{T_2} = \lambda_P$ is to mitigate the impact of age and time since diagnosis (calendar year) on potentially dependent censoring by C (Perme et al., 2012). We estimate $2, 5, 10,$ and $15-$year relative survival assuming a Gumbel distribution for T_1 and a Gumbel copula model with differing levels of dependence to specify the joint distribution of T_1 and T_2. Unlike the standard independence competing methods, the results of our estimator compares favorably with the standard nonparametric estimator of Perme et al. (2012), which require independence of T_1 and T_2. Besides, our estimates also provide credible estimates (crude survival or crude probability of death) under the dependence of T_1 and T_2.

4.5 Discussion and Conclussion

We proposed dependence competing risk regression model to determine the effect of the risk factors on disease-specific survival analysis. Our model formulation was arbitrary but permitting but known copula function and any form of transformation required for incorporating covariates. We extracted the distribution of other cause mortality from external acturial reference data as is required by the underpinning of relative survival method. Due to the identifiability and unveriable nature of dependence between competing mortality, we proposed sensitivity analysis as a practical solution to this issue. The effect of the risk factors is useful not only for determining treatment regimen for elderly cancer patients but also for the physicians in determining plausible prognostic measures.

The dependent competing risk regression model for disease-specific mortality may in part be restrictive but flexible enough for applications where the disease-specific hazard is smooth over time, which is the case in cancer registry data. It also permit evaluation of the covariates in the model which lend to simple interpretation of disease-specific survival. The purpose of this paper is to evaluate covariate effect both under the independence and dependence of T_1 and T_2. We acknowledged that ignoring dependence in competing risk in survival processes may be regarded as modifying the research question to satisfy existing methods. The effect of covariates under independence assumption for competing risks is useful for practitioners who prefer net survival to crude survival probability of death which is achieved under the dependence assumption.

CHAPTER 5: Conclusion

5.1 Conclusion

This dissertation focused on the development of dependence models for disease-specific survival analysis using registry data. In such competing risk settings where interest lies in comparison of different cohort groups and or populations, our motivation stem from the sheer lack of appropriate disease-specific survival estimators for this type of registries. As a result, we proposed three novel estimators for disease-specific survival analysis without the need for cause of death information whether reliable or not. Our estimators were shown to be consistent and asymptotically normal in simulation studies satisfying all the usual regularity conditions. The estimators were applied to the French breast cancer data for estimating both the overall and age-specific survival under both independence and dependent assumptions. Our estimators across the levels of dependence may be interpreted as representing personalized prognostic measures.

First, we proposed a parametric dependence model for disease-specific survival in relative survival analysis via a copula function. Copula models capture scale invariance dependence between the unobservable failure times for disease-specific mortality and other cause mortality. A bivariate dependence competing risk model was formulated via copula taking as inputs the distribution of the minimum latent failure times where the distribution of the competing latent failure time was extracted from a healthy reference population. Likelihood-based estimation inference was proposed. We investigated theoretical properties such as Fisher consistency for the dependent parametric model under both Gumbel and Clayton copulas.

We relaxed the parametric assumption for relative survival analysis where we proposed a nonparametric estimator for disease-specific survival. We modelled a function of the standard Kaplan-Meier estimator in a form of estimation equation where the inversion of the nonlinear function was required for disease-specific survival estimation. Nonparametric bootstrap procedure was implemented for variance estimation under the usual regularity conditions.

Whereas for the last estimator, we assessed the effects of disease-specific risk factors via a dependence competing regression model. The covariates were incorporated in the regression model using the standard Accelerated Failure Time procedure.

The dependence between disease-specific death and other cause mortality is nonidentifiable and unverifiable from the observed registry data. As such, sensitivity analysis was proposed where disease-specific mortality is estimated across a range of rich dependence levels. Our methods performed well in both simulation studies and real-world data.

In conclusion, it is important to warn readers that prognostic measures (net survival, crude survival or crude probability of death) for disease-specific survival are lagging resulting in unrealistic clinician and or patient prognostic expectations that may lead to inappropriate therapeutic goals. Net survival is valid for use in a hypothetical world where the disease is the only cause of death and use for comparison of prognosis among different groups and or populations. The crude survival or crude probability of death is a valid measure in the real world where competing mortality exists simultaneously with the disease-specific under study and can be regarded as the analog for competing risk estimator when cause of death is known. Some practitioners prefer one prognostic measure over the other. The issue of predictive prognostic measures and time-dependent covariate models for relative survival analysis has being deferred for later.

APPENDIX 1: Results from Clayton Copula

Distribution of Death and Competing Causes of Death

The following figures: A.6, A.7 and A.8 show the distributions of time to disease-specific death and death due to competing mortality possible in breast cancer registry data. The simulation reveal the different levels of dependency that is typically observed from time since diagnosis.

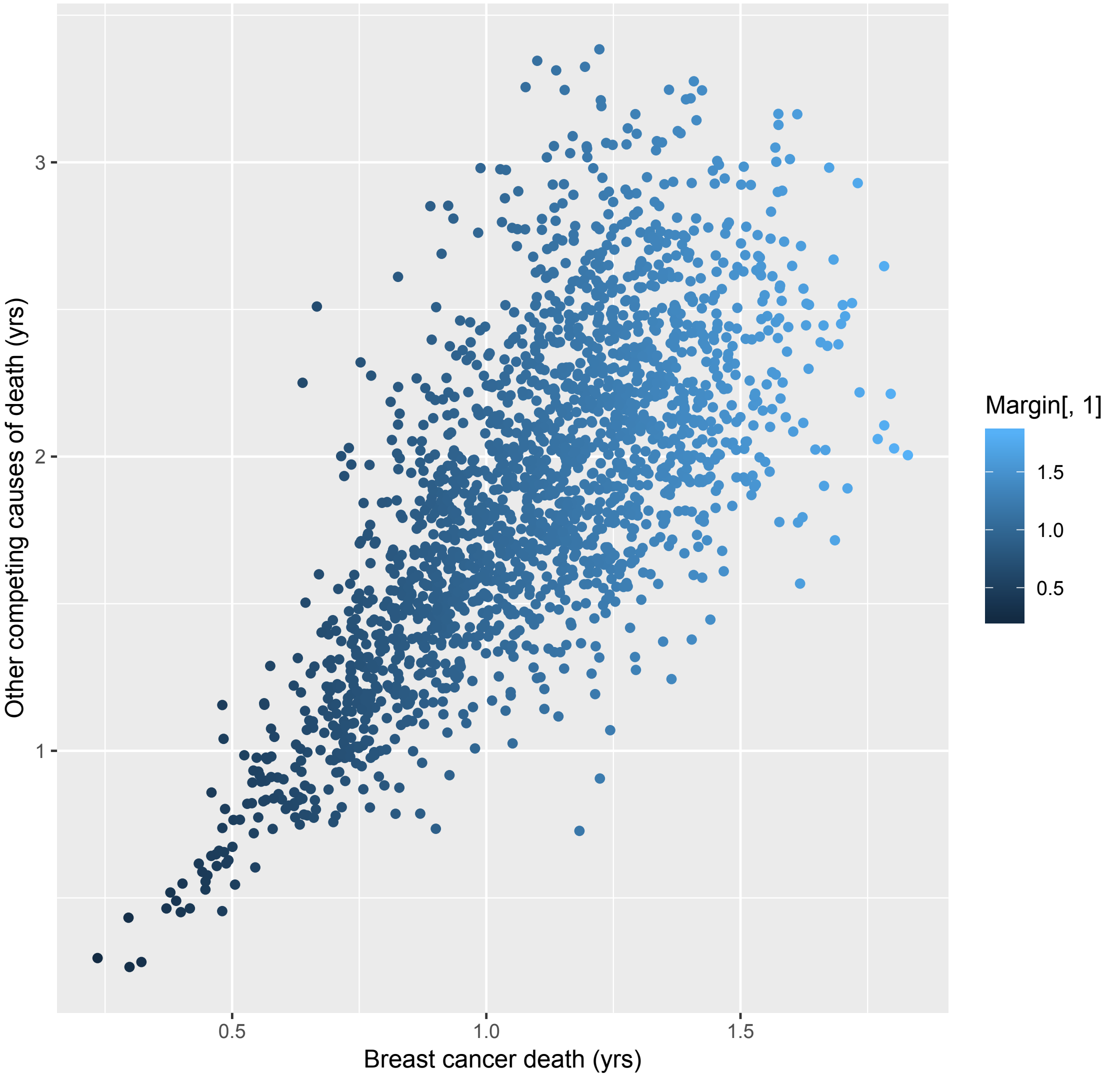

Figure A.6: A simulated breast cancer data with moderate dependence (50%) through the use of Clayton copula.

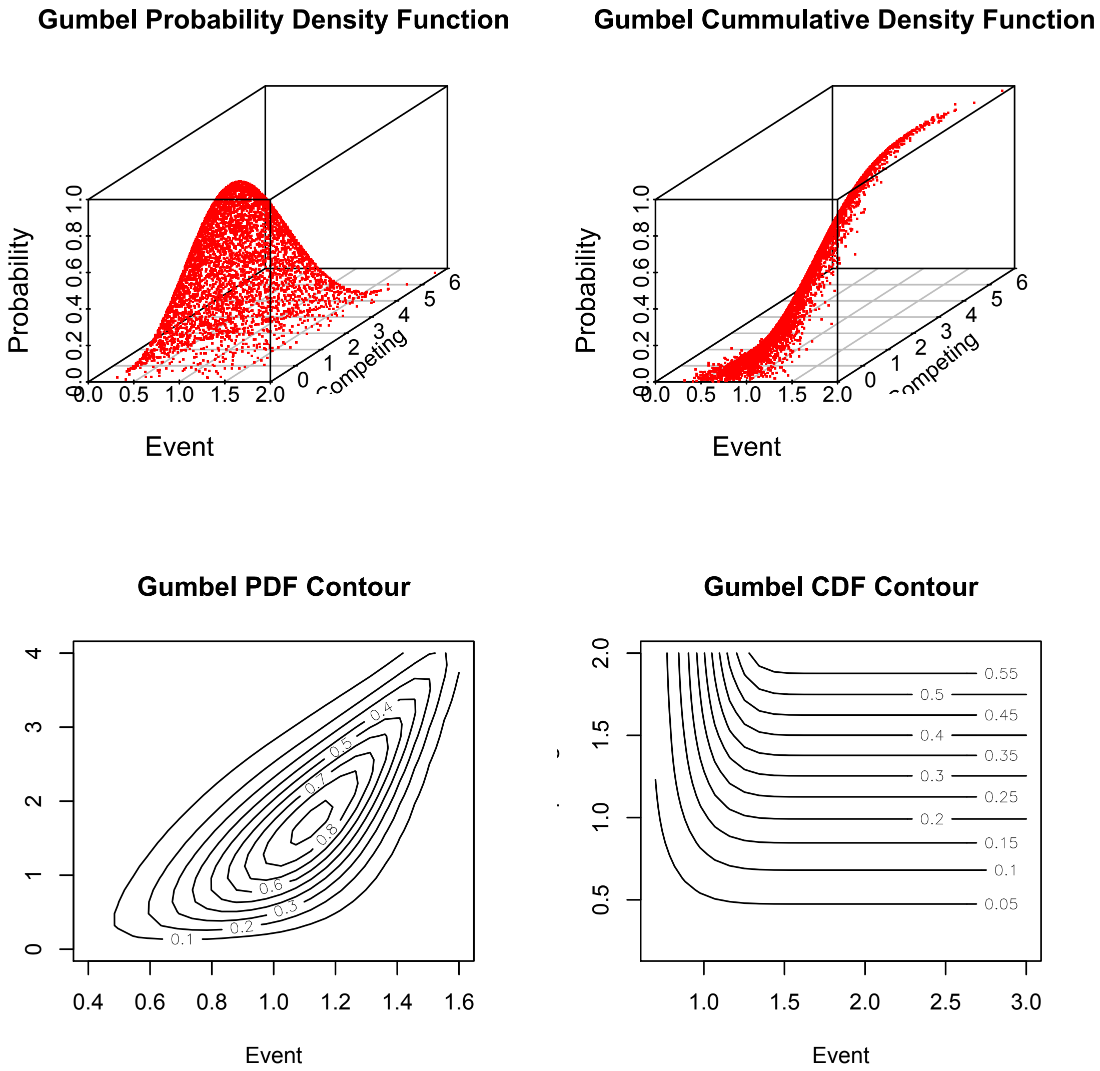

Figure A.7: A plot showing the PDF, CDF and contour plots for the Gumbel copula which exhibit upper tail dependency.

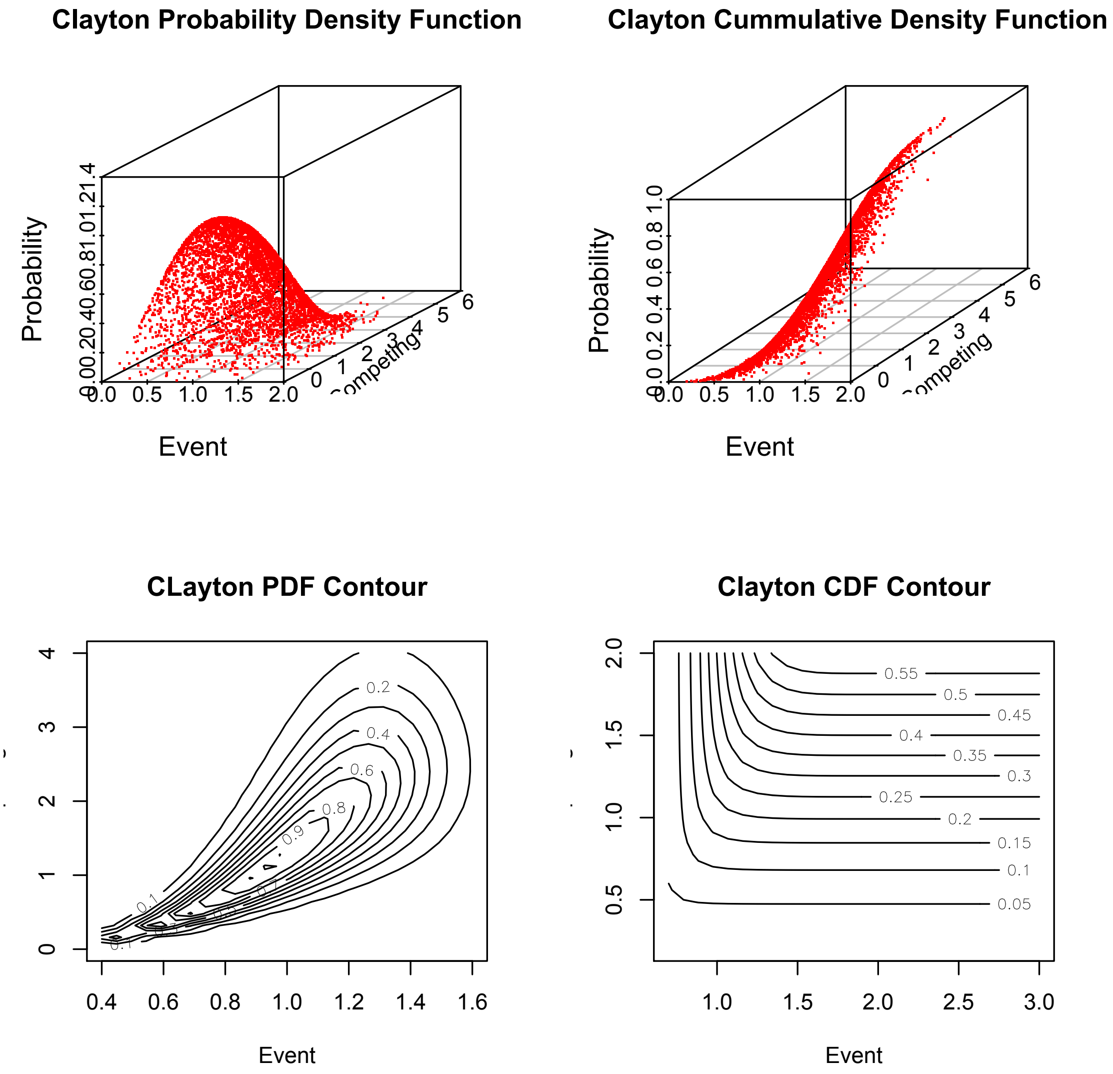

Figure A.8: A plot showing the PDF, CDF and contour plots for the Clayton copula which exhibit lower tail dependency.

Parameter Estimates for the Clayton Copula

We simulated data to mimic the French breast cancer data set for sample sizes; $1000, 2500$ and 5000 with 500 replications. The latent failure times for $T_j \sim Weibull(\alpha_j, \lambda_j)$ with probability density function defined in section 2.3. The parameters for the Weibull distribution for T_1 were $\lambda_1 = 0.182$ and $\alpha_1 = 1.609$, while those for T_2 were $\lambda_2 = 0.742$ and $\alpha_2 = 0.693$. In the estimation of $\lambda_1,\ \alpha_1$ for T_1, $\lambda_2,\ \alpha_2$ are assumed known for T_2. Noninformative censoring times were generated from a uniform distribution $(0, \gamma)$, where γ was chosen for $10,\ 30$ and 50% censoring. We consider the Clayton copula with Kendall's tau, $\tau_k = \frac{\theta}{\theta+2} = 0,\ 0.25,\ 0.50,\ 0.75$. Initial parameter values were randomly chosen from uniform distributions, with multiple starting values wherever possible as described in section 2.3. The simulation results based on the Clayton copula is presented in the table below:

Table A.12: Estimated parameters of the model for T_1 across samples sizes (N), dependence levels (τ_k) and levels of censoring (C) treating T_2 as a competing event and vice versa.

	C		0.10					0.30					0.50				
τ_k	N	$\hat{\eta}$	Mean	Bias^a	ModB^a	EMP ^{a}	CP	Mean	Bias^a	ModB ^{a}	EMP ^{a}	CP	Mean	Bias^a	ModB^a	EMP^a	CP
0.00	1000	$\hat{\lambda}_1$	0.182	0.000	0.080	0.090	0.948	0.182	-0.240	0.110	0.110	0.940	0.182	-0.340	0.140	0.160	0.930
		$\hat{\alpha}_1$	1.610	0.710	1.390	1.520	0.942	1.611	1.200	1.770	1.930	0.936	1.611	1.820	2.490	2.540	0.948
	2500	$\hat{\lambda}_1$	0.182	-0.280	0.030	0.030	0.962	0.182	-0.280	0.040	0.040	0.950	0.182	-0.050	0.060	0.050	0.958
		$\hat{\alpha}_1$	1.609	-0.470	0.550	0.570	0.938	1.610	0.090	0.710	0.720	0.954	1.609	-0.720	0.990	1.040	0.944
	5000	$\hat{\lambda}_1$	0.182	-0.020	0.020	0.020	0.942	0.182	0.140	0.020	0.020	0.944	0.182	0.060	0.030	0.030	0.946
		$\hat{\alpha}_1$	1.610	0.520	0.280	0.280	0.956	1.610	0.610	0.350	0.350	0.960	1.610	0.710	0.500	0.510	0.952
0.25	1000	$\hat{\lambda}_1$	0.182	-0.010	0.070	0.080	0.952	0.182	-0.510	0.090	0.080	0.964	0.182	-0.670	0.130	0.110	0.964
		$\hat{\alpha}_1$	1.610	0.560	1.280	1.410	0.930	1.609	0.020	1.660	1.740	0.936	1.611	1.870	2.390	0.220	0.952
	2500	$\hat{\lambda}_1$	0.182	-0.130	0.030	0.020	0.962	0.182	-0.100	0.030	0.030	0.962	0.182	0.150	0.040	0.040	0.958
		$\hat{\alpha}_1$	1.609	-0.250	0.450	0.460	0.944	1.610	0.660	0.590	0.600	0.950	1.609	-0.240	0.870	0.840	0.960
	5000	$\hat{\lambda}_1$	0.182	-0.180	0.010	0.010	0.946	0.182	-0.020	0.020	0.020	0.952	0.182	-0.120	0.020	0.020	0.940
		$\hat{\alpha}_1$	1.609	-0.310	0.230	0.230	0.950	1.610	0.150	0.300	0.300	0.956	1.609	-0.030	0.430	0.450	0.958
0.50	1000	$\hat{\lambda}_1$	0.182	-0.090	0.060	0.050	0.956	0.182	0.000	0.070	0.070	0.956	0.182	-0.440	0.100	0.100	0.964
		$\hat{\alpha}_1$	1.608	-0.980	1.030	1.090	0.952	1.610	0.710	1.360	1.420	0.924	1.612	2.580	2.020	2.060	0.968
	2500	$\hat{\lambda}_1$	0.182	-0.110	0.020	0.020	0.958	0.182	-0.110	0.030	0.030	0.950	0.183	0.230	0.040	0.040	0.958
		$\hat{\alpha}_1$	1.609	-0.330	0.410	0.440	0.942	1.610	0.610	0.540	0.570	0.934	1.609	-0.130	0.800	0.800	0.950
	5000	$\hat{\lambda}_1$	0.182	-0.180	0.010	0.010	0.936	0.182	-0.010	0.010	0.010	0.940	0.182	-0.180	0.020	0.020	0.940
		$\hat{\alpha}_1$	1.609	-0.190	0.200	0.210	0.952	1.610	0.540	0.270	0.290	0.950	1.610	0.240	0.400	0.410	0.948
0.75	1000	$\hat{\lambda}_1$	0.182	0.130	0.040	0.040	0.944	0.182	0.110	0.050	0.050	0.952	0.182	-0.050	0.080	0.080	0.937
		$\hat{\alpha}_1$	1.609	0.030	7e-04	0.860	0.924	1.611	1.330	0.930	1.110	0.924	1.610	1.120	1.410	1.450	0.947
	2500	$\hat{\lambda}_1$	0.182	-0.110	0.020	0.020	0.958	0.182	-0.100	0.020	0.020	0.960	0.183	0.310	0.030	0.030	0.962
		$\hat{\alpha}_1$	1.609	-0.410	0.280	0.320	0.940	1.610	0.520	0.370	0.410	0.940	1.609	0.000	0.550	0.560	0.948
	5000	$\hat{\lambda}_1$	0.182	-0.120	0.010	0.010	0.940	0.182	0.070	0.010	0.010	0.950	0.182	-0.040	0.020	0.020	0.948
		$\hat{\alpha}_1$	1.609	0.040	0.140	0.160	0.936	1.610	0.710	0.180	0.210	0.930	1.610	0.780	0.270	0.320	0.926

$\hat{\eta}$: estimated parameters, ModB: model-based variance, EMP: empirical variance, CP: 95% coverage probability. ^{a} : $\times 10^{-3}$

APPENDIX 2: Estimating Equation

Relative survival is used extensively in population-based cancer studies to measure patient survival correcting for causes of death not related to the disease of interest. For many years, the gold standard for nonparametric estimation of survival curves has been the Hakulinen estimator Hakulinen and Tenkanen (1987), but recently, Pokhrel and Hakulinen (2009) and Hakulinen et al. (2011) had shown that this estimator does not have the expected properties. In their work, they employ restricted cubic splines for the baseline cumulative excess hazard and for any time-dependent effects. All the attempts to correct for these bias still require correct classification of cause of death, and an assumption of independence between event and competing events.

Variance Estimation for $S_{T_1}(X)$

The estimation of the variance of the model was achieved by using the first order Delta method which is in the broad spectrum of the usual estimating equations. The Greenwood formula for estimating the variance of all-causes survival function is:

$$Var\left(\widehat{S_T(X,X)}\right) = \widehat{S_T(X,X)}^2 \sum_{X_i \leq x} \frac{d_i}{n_i(n_i - d_i)} \tag{H.35}$$

The variance estimation follows from Delta method or the application of the first Taylor series expansion.

Estimating Equation

The variance estimation for the nonlinear function is achieved by the use of generalized estimating equation. We define the estimating equation for 3.20 for a guumbel copula as;

$$g(S_{T_1}(X)) = \widehat{S_T(X,X)} - S_{T_1}(X) - S_{T_2}(X) + 1 + \tag{H.36}$$

$$\exp\left\{ - \left((-\log(1 - S_{T_1}(X))^\theta + (-\log(1 - S_{T_2}(X))^\theta \right)^{\frac{1}{\theta}} \right\} = 0 \tag{H.37}$$

From the Taylor series expansion around the $S_{T_1}(X)$,

$$\begin{aligned} g\left(\widehat{S_{T_1}(X)}\right) &= g\left(S_{T_1}(X)\right) + g'\left(S_{T_1}(X)\right)\cdot\left(\widehat{S_{T_1}(X)} - S_{T_1}(X)\right) \\ &\quad + \frac{g''\left(S_{T_1}(X)\right)}{2}\cdot\left(\widehat{S_{T_1}(X)} - S_{T_1}(X)\right)^2 + R \end{aligned} \tag{H.38}$$

So the variance of $\widehat{S_{T_1}(X)}$ is:

$$Var(\widehat{S_{T_1}(X)}) = \left[g'\left(S_{T_1}(X)\right)\right]^T \cdot Var\left(g\left(\widehat{S_{T_1}(X)}\right)\right)\cdot\left[g'\left(S_{T_1}(X)\right)\right] \tag{H.39}$$

$$g\left(\widehat{S_{T_1}(X)}\right) = -\widehat{S_{T_1}(X)} + \exp\left\{-\left[(-log(1-\widehat{S_{T_1}(X)}))^\theta + (-log(1-S_{T_2}(X)))^\theta\right]^{\frac{1}{\theta}}\right\}$$

$$g(\,\cdot\,) + \frac{\partial}{\partial S_{T_1}(X)}g(\,\cdot\,)\cdot(\widehat{S_{T_1}(X)} - S_{T_1}(X)) + \frac{\partial^2}{\partial S_{T_1}(X)^2}g(\,\cdot\,)\cdot(\widehat{S_{T_1}(X)} - S_{T_1}(X))^2 + \cdots = 0$$

and

$$\begin{aligned} \widehat{S_{T_1}(X)} - S_{T_1}(X) &= -\frac{g(\,\cdot\,)}{g'(\,\cdot\,)} + \epsilon \\ \Longrightarrow \sqrt{n}(\widehat{S_{T_1}(X)} - S_{T_1}(X)) &\to -\sqrt{n}\frac{g(\,\cdot\,)}{g'(\,\cdot\,)} + \epsilon \\ \Longrightarrow \sqrt{n}(\widehat{S_{T_1}(X)} - S_{T_1}(X)) &\backsim N\left(0, var\left(\sqrt{n}\frac{g(\,\cdot\,)}{g'(\,\cdot\,)}\right)\right) \end{aligned}$$

Using Delta method, we can compute the variance as

$$var\left(\sqrt{n}\ \frac{g(\,\cdot\,)}{g'(\,\cdot\,)}\right) = \frac{1}{g'(\,\cdot\,)^2}\cdot var\left(\sqrt{n}g(\,\cdot\,)\right)$$

Here, per the definition of the $g(\,\cdot\,)$, the first derivative is given by

$$I(\eta) = \begin{pmatrix} \frac{\partial^2 l(\eta|\mathbf{X})}{\partial\alpha^2} & \frac{\partial^2 l(\eta|\mathbf{X})}{\partial\alpha\partial\lambda} \\ \frac{\partial^2 l(\eta|\mathbf{X})}{\partial\lambda\partial\alpha} & \frac{\partial^2 l(\eta|\mathbf{X})}{\partial\lambda^2} \end{pmatrix} \tag{A.6}$$

Our model was too complicated for the calculation of the derivatives. We approximated the Hessian matrix by the use of numerical methods. The negative of the Hessian matrix give the covariance. The diagonal of the covariance matrix is the variance.

The variance estimation follows from Delta method or the application of the first Taylor series expansion. Thus, we suppose $g(t) = \psi\left(\psi^{-1}(S(t)) - \psi^{-1}(S_2(t))\right)$ The first order moment is given by $g(T) \approx g(\theta) + \sum_{i=1}^{k} g_i(\theta)(T_i - \theta_i) + Remainder$. The expectation is given by $\mathbb{E}g(T) \approx g(\theta) + \sum_{i=1}^{k} g_i'(\theta)\mathbb{E}(T_i - \theta_i) = g(\theta)$. The variance can be derived as

$$\begin{aligned} Var(g(T)) &= \mathbb{E}\left(g(T_i) - g(\theta_i)\right)^2 \\ &= \mathbb{E}\left(g(\theta) + \sum_{i=1}^{k} g_i'(\theta)(T_i - \theta_i) - g(\theta)\right)^2 = \mathbb{E}\left(\sum_{i=1}^{k} g_i'(\theta)(T_i - \theta_i)\right)^2 \end{aligned}$$

$$= \sum_{i=1}^{k} g_i'(\theta)^2 Var(T_i) + 2\sum_{i>j}^{k} g_i'(\theta)g_j'(\theta)cov(T_i, T_j)$$

(A.7)

The Delta method for estimates the variance of $\hat{S}_1(x)$ is given by;

$$\widehat{Var}\left(\widehat{S_1(x)}\right) = g'\left(\widehat{S(x,x)}, S_2(x), \theta\right) \cdot Var\left(\widehat{S(x,x)}\right) \cdot g'\left(\widehat{S(x,x)}, S_2(x), \theta\right)^T \quad \text{(H.40)}$$

$$= g'\left(\widehat{S(x,x)}, S_2(x), \theta\right)^2 \cdot Var\left(\widehat{S(x,x)}\right) \quad \text{(A.8)}$$

where $g(.)$ is a monotone transformation function of $S_1(x)$. The derivative of $g(.)$ is given in the next appendix.

$$g\left(S(x,x), S_2(x), \theta\right) = \left(S(x,x)^{1-\theta} - S_2(x)^{1-\theta} + 1\right)^{\frac{1}{1-\theta}}$$

$$\begin{aligned} g'(\cdot) &= \frac{\partial g(\cdot)}{\partial S(x,x)} = \Big(S(x,x)^{1-\theta} - S_2(x)^{1-\theta} + 1\Big)^{\frac{\theta}{1-\theta}} \cdot S(x,x)^{-\theta} \\ &= \{g\left(S(x,x), S_2(x), \theta\right)\}^{\theta} \cdot S(x,x)^{-\theta} \qquad \text{(H.41)} \end{aligned}$$

APPENDIX 3: Covariate Model Derivation

AFT Model formulation

Emphasis is placed on the estimates of the regression coefficients. If $T_j \backsim Weibull(t_j : \mu_j, \sigma_j),\ \forall\ j \in (1,2)$ is a random variable, then $\log(T_j)$ follows an extreme value distribution, $W_j \backsim Gumbel(w_j : \alpha_j, \lambda_j)$ with α the mode (location) and λ the scale parameters. The Gumbel distribution is generally used for modelling the exceedance above a threshold such as time to death. The survival probability for people living beyond this threshold time is important for clinicians. A model without covariate is represented as $\log(T_j) = \alpha_j + \frac{W_j}{\lambda_j}$ while a model with covariates Z is given by:

$$X_j = \log(T_j) = Z\beta + \frac{W_j}{\lambda_j} \tag{I.42}$$

where β is the parameters of the model associated with the location. The random variable T_j is non-negative and is expressed as $T = \exp\left\{Z\beta + \frac{W}{\lambda}\right\}$. It is clear from appendix $A.3$ that $X_j \sim Gumbel(x_j; \alpha_j, \lambda_j)$. The domain of the parameters is $(\alpha, \lambda) \rightarrow (\mathbb{R} \times (0, \infty))$. Extensions to Gumbel valued functions over continuous spaces have being extensively explored in random choice theory, (Malmberg, 2013).

$$\hat{\beta} = Z(Z'Z)^{-1}X \tag{I.43}$$

where X is the dependent(response) variable representing $X = \min(T, C)$ where $T = \min(T_1, T_2)$ and C is the right censoring time, Z is the set of covariates, and β is the regression coefficients for the risk factors. The cumulative distribution function is given by

$$F_{W_j}(w_j|\alpha_j, \lambda_j) = \exp\left\{-\exp\left\{-\frac{w_j - \alpha_j}{\lambda_j}\right\}\right\} \tag{I.44}$$

and its probability density function is given by

$$f_{W_j}(w_j) = \frac{1}{\lambda_j} * \exp\left\{-\frac{w_j - \alpha_j}{\lambda_j} - \exp\left\{-\frac{w_j - \alpha_j}{\lambda_j}\right\}\right\}$$

$$f_{W_j}(w_j) = \frac{1}{\lambda_j} * \exp\left\{-\frac{w_j - \alpha_j}{\lambda_j}\right\} \cdot \exp\left\{-\exp\left\{-\frac{w_j - \alpha_j}{\lambda_j}\right\}\right\}$$

$$f_{W_j}(w_j) = \frac{1}{\lambda_j} * \exp\left\{-\frac{w_j - \alpha_j}{\lambda_j}\right\} \cdot F_{W_j}(w_j|\alpha_j, \lambda_j)$$

The corresponding survival function is given by

$$S_{W_j}(w_j|\alpha_j, \lambda_j) = 1 - F_{W_j}(w_j|\alpha_j, \lambda_j)$$

REFERENCES

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

Adatorwovor, R., Latouche, A., and Fine, J. P. (2020). Relaxing the independence assumption in relative survival analysis. *Statistics in medicine*.

Austin, P. C. and Fine, J. P. (2017). Accounting for competing risks in randomized controlled trials: a review and recommendations for improvement. *Statistics in medicine*, 36(8):1203–1209.

Austin, P. C., Latouche, A., and Fine, J. P. (2020). A review of the use of time-varying covariates in the fine-gray subdistribution hazard competing risk regression model. *Statistics in medicine*, 39(2):103–113.

Austin, P. C., Lee, D. S., and Fine, J. P. (2016). Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133(6):601–609.

Bäuerle, N. and Müller, A. (1998). Modeling and comparing dependencies in multivariate risk portfolios. *ASTIN Bulletin: The Journal of the IAA*, 28(1):59–76.

Begg, C. B. and Schrag, D. (2002). Attribution of deaths following cancer treatment.

Berkson, J. and Gage, R. P. (1950). Calculation of survival rates for cancer. In *Proceedings of the staff meetings. Mayo Clinic*, volume 25, page 270.

Bernstein, S. et al. (1929). Sur les fonctions absolument monotones. *Acta Mathematica*, 52:1–66.

Bland, J. M. and Altman, D. G. (2004). The logrank test. *Bmj*, 328(7447):1073.

Brinkhof, M. W., Spycher, B. D., Yiannoutsos, C., Weigel, R., Wood, R., Messou, E., Boulle, A., Egger, M., Sterne, J. A., epidemiological Database to Evaluate AIDS (IeDEA, I., et al. (2010). Adjusting mortality for loss to follow-up: analysis of five art programmes in sub-saharan africa. *PloS one*, 5(11):e14149.

Caplan, R. J., Pajak, T. F., and Cox, J. D. (1994). Analysis of the probability and risk of cause-specific failure. *International Journal of Radiation Oncology* Biology* Physics*, 29(5):1183–1186.

Chappell, R. (2012). Competing risk analyses: how are they different and why should you care? *Clinical Cancer Research*, 18(8):2127–2129.

Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula methods in finance*. John Wiley & Sons.

Clarke, B., Fokoue, E., and Zhang, H. H. (2009). *Principles and theory for data mining and machine learning*. Springer Science & Business Media.

Corazziari, I., Quinn, M., and Capocaccia, R. (2004). Standard cancer patient population for age standardising survival ratios. *European Journal of Cancer*, 40(15):2307–2316.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Cronin, K. A. and Feuer, E. J. (2000). Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival. *Statistics in medicine*, 19(13):1729–1740.

D'Amico, G., Morabito, A., D'Amico, M., Pasta, L., Malizia, G., Rebora, P., and Valsecchi, M. G. (2018). Clinical states of cirrhosis and competing risks. *Journal of hepatology*, 68(3):563–576.

de Lacerda, G. F., Howlader, N., and Mariotto, A. B. (2019). Differences in cancer survival with relative versus cause-specific approaches: an update using more accurate life tables. *Cancer Epidemiology and Prevention Biomarkers*, 28(9):1544–1551.

Deheuvels, P. (1978). Caractérisation complète des lois extrêmes multivariées et de la convergence des types extrêmes. *Publ. Inst. Statist. Univ. Paris*, 23(3):1–36.

Denham, J., Hamilton, C., and O'Brien, P. (1996). Regarding actuarial late effect analyses: Bentzen et al., ijrobp 32: 1531–1534; 1995 and caplan et al., ijrobp 32: 1547; 1995. *International Journal of Radiation Oncology? Biology? Physics*, 35(1):197.

Denuit, M., Genest, C., and Marceau, É. (1999). Stochastic bounds on sums of dependent risks. *Insurance: Mathematics and Economics*, 25(1):85–104.

Dickman, P. W., Sloggett, A., Hills, M., and Hakulinen, T. (2004). Regression models for relative survival. *Statistics in medicine*, 23(1):51–64.

Dignam, J. J. and Kocherginsky, M. N. (2008). Choice and interpretation of statistical tests used when competing risks are present. *Journal of Clinical Oncology*, 26(24):4027.

Dignam, J. J., Zhang, Q., and Kocherginsky, M. (2012). The use and interpretation of competing risks regression models. *Clinical Cancer Research*, 18(8):2301–2308.

Ederer, F. (1961). The relative survival rate: a statistical methodology. *Cancer: end results and mortality trends.*

Ederer, F., Geisser, M. S., Mongin, S. J., Church, T. R., and Mandel, J. S. (1999). Colorectal cancer deaths as determined by expert committee and from death certificate: a comparison. the minnesota study. *Journal of clinical epidemiology*, 52(5):447–452.

Ederer, F. and Heise, H. (1959). Instructions to ibm 650 programmers in processing survival computations. Technical report, Methodological note.

Fermanian, J.-D. (2003). Nonparametric estimation of competing risks models with covariates. *Journal of Multivariate Analysis*, 85(1):156–191.

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

Fine, J. P., Jiang, H., and Chappell, R. (2001). On semi-competing risks data. *Biometrika*, 88(4):907–919.

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon, 3ˆ e serie, Sciences, Sect. A*, 14:53–77.

Genest, C. and MacKay, J. (1986). The joy of copulas: Bivariate distributions with uniform marginals. *The American Statistician*, 40(4):280–283.

German, R. R., Fink, A. K., Heron, M., Stewart, S. L., Johnson, C. J., Finch, J. L., Yin, D., of Cancer Mortality Study Group, A., et al. (2011). The accuracy of cancer mortality statistics based on death certificates in the united states. *Cancer epidemiology*, 35(2):126–131.

Gichangi, A. and Vach, W. (2005). The analysis of competing risks data: A guided tour. *Statistics in Medicine*, 132(4):1–41.

Gooley, T. A., Leisenring, W., Crowley, J., and Storer, B. E. (1999). Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in medicine*, 18(6):695–706.

Gray, R. J. et al. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of statistics*, 16(3):1141–1154.

Hakulinen, T. (1982). Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics*, pages 933–942.

Hakulinen, T., Seppä, K., and Lambert, P. C. (2011). Choosing the relative survival method for cancer survival estimation. *European Journal of Cancer*, 47(14):2202–2210.

Hakulinen, T. and Tenkanen, L. (1987). Regression analysis of relative survival rates. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(3):309–317.

Hasselman, B. (2009). nleqslv: Solve systems of non linear equations. *R package version*, 1.

Heckman, J. J. and Honoré, B. E. (1989). The identifiability of the competing risks model. *Biometrika*, 76(2):325–330.

Hoeffding, W. (1940). Masstabinvariante korrelationstheorie. *Schriften des Mathematischen Instituts und Instituts fur Angewandte Mathematik der Universitat Berlin*, 5:181–233.

Hoel, D. G., Ron, E., Carter, R., and Mabuchi, K. (1993). Influence of death certificate errors on cancer mortality trends. *JNCI: Journal of the National Cancer Institute*, 85(13):1063–1068.

James, D. and Bull, A. (1996). Information on death certificates: cause for concern? *Journal of clinical pathology*, 49(3):213–216.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts.* CRC Press.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Lambert, P., Dickman, P., Nelson, C., and Royston, P. (2010). Estimating the crude probability of death due to cancer and other causes using relative survival models. *Statistics in medicine*, 29(7-8):885–895.

Lau, B., Cole, S. R., and Gange, S. J. (2009). Competing risk regression models for epidemiologic data. *American journal of epidemiology*, 170(2):244–256.

Louzada, F., Cancho, V. G., and Yiqi, B. (2015). The log-weibull-negative-binomial regression model under latent failure causes and presence of randomized activation schemes. *Statistics*, 49(4):930–949.

Makkar, N., Ostrom, Q. T., Kruchko, C., and Barnholtz-Sloan, J. S. (2018). A comparison of relative survival and cause-specific survival methods to measure net survival in cancer populations. *Cancer medicine*, 7(9):4773–4780.

Malmberg, H. (2013). *Random Choice over a Continuous Set of Options*. PhD thesis, Department of Mathematics, Stockholm University.

Maudsley, G. and Williams, E. (1996). ?inaccuracy?in death certification–where are we now? *Journal of Public Health*, 18(1):59–66.

McNeil, A. J., Nešlehová, J., et al. (2009). Multivariate archimedean copulas, d-monotone functions and ?1-norm symmetric distributions. *The Annals of Statistics*, 37(5B):3059–3097.

Mieno, M. N., Tanaka, N., Arai, T., Kawahara, T., Kuchiba, A., Ishikawa, S., and Sawabe, M. (2016). Accuracy of death certificates and assessment of factors for misclassification of underlying cause of death. *Journal of epidemiology*, 26(4):191–198.

Müller, A. (1996). Orderings of risks: A comparative study via stop-loss transforms. *Insurance: Mathematics and Economics*, 17(3):215–222.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The computer journal*, 7(4):308–313.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966.

Nixon, A. J., Neuberg, D., Hayes, D. F., Gelman, R., Connolly, J. L., Schnitt, S., Abner, A., Recht, A., Vicini, F., and Harris, J. R. (1994). Relationship of patient age to pathologic features of the tumor and prognosis for patients with stage i or ii breast cancer. *Journal of Clinical Oncology*, 12(5):888–894.

Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 414–422.

Organization, W. H. et al. (1977). *Manual of the international statistical classification of diseases, injuries, and causes of death: based on the recommendations of the ninth revision conference, 1975, and adopted by the Twenty-ninth World Health Assembly.* Geneva: World Health Organization.

Percy, C. (1989). International comparability of coding cancer data: present state and possible improvement by icd-10. In *Cancer Mapping*, pages 240–252. Springer.

Percy, C., Stanek 3rd, E., and Gloeckler, L. (1981). Accuracy of cancer death certificates and its effect on cancer mortality statistics. *American journal of public health*, 71(3):242–250.

Perme, M. P., Stare, J., and Estève, J. (2012). On estimation in relative survival. *Biometrics*, 68(1):113–120.

Platell, C. F. and Semmens, J. B. (2004). Review of survival curves for colorectal cancer. *Diseases of the colon & rectum*, 47(12):2070–2075.

Pokhrel, A. and Hakulinen, T. (2009). Age-standardisation of relative survival ratios of cancer patients in a comparison between countries, genders and time periods. *European journal of cancer*, 45(4):642–647.

Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, pages 541–554.

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

Reason, J. (1990). The contribution of latent human failures to the breakdown of complex systems. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 327(1241):475–484.

Rebolj Kodre, A. and Pohar Perme, M. (2013). Informative censoring in relative survival. *Statistics in medicine*, 32(27):4791–4802.

Sarfati, D., Blakely, T., and Pearce, N. (2010). Measuring cancer survival in populations: relative survival vs cancer-specific survival. *International journal of epidemiology*, 39(2):598–610.

Sasieni, P. and Brentnall, A. R. (2017). On standardized relative survival. *Biometrics*, 73(2):473–482.

Slud, E. V., Byar, D. P., Schatzkin, A., Prentice, R., and Kalbfleisch, J. (1988). Dependent competing risks and the latent-failure model.

Sturgeon, K. M., Deng, L., Bluethmann, S. M., Zhou, S., Trifiletti, D. M., Jiang, C., Kelly, S. P., and Zaorsky, N. G. (2019). A population-based study of cardiovascular disease mortality risk in us cancer patients. *European heart journal*, 40(48):3889–3897.

Suissa, S. (1999). Relative excess risk: an alternative measure of comparative risk. *American journal of epidemiology*, 150(3):279–282.

Tan, K. S., Eguchi, T., and Adusumilli, P. S. (2019). Reporting net survival in populations: a sensitivity analysis in lung cancer demonstrates the differential implications of reporting relative survival and cause-specific survival. *Clinical Epidemiology*, 11:781.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22.

Venter, G. G. (2002). Tails of copulas. In *Proceedings of the Casualty Actuarial Society*, volume 89, pages 68–113.

Welch, H. G. and Black, W. C. (2002). Are deaths within 1 month of cancer-directed surgery attributed to cancer? *Journal of the National Cancer Institute*, 94(14):1066–1070.

Williamson, P., Kolamunnage-Dona, R., and Tudur Smith, C. (2007). The influence of competing-risks setting on the choice of hypothesis test for treatment effect. *Biostatistics*, 8(4):689–694.