Copula-Based Survival Models for Dependent Competing Risks

by

Ali Hossein Gharari Foomani

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

A survival dataset describes a collection of instances, such as patients, and associates each instance with either the time until an event (such as death), or the censoring time (eg, when the instance is lost to follow-up), which is a lower bound on the time until the event. While there are several approaches to survival prediction, this thesis focuses on models that produce an individual survival curve (providing $P(death \geq t|x)$ for each $t > 0$) for each individual patient, x – here based on a "Deep Weibull" model. Most survival prediction methods assume that the event and censoring distributions are independent given the instance's covariates. This assumption is challenging to verify since we only observe a single outcome (event xor censor time) for each instance. Moreover, models that assume this independence can be substantially biased when this independence does not hold. Moreover, the standard methods to evaluate survival models do not provide meaningful values here.

In this study, we present a way to relax the assumption of conditional independence, using a parametric model of survival that incorporates Archimedean copulas to address residual dependency that cannot be explained by the covariates in the dataset. Additionally, we show how to extend this to a broader range of dependencies by using a convex combination of members from the Archimedean copula family, rather than relying on a specific member. Our empirical studies, conducted on synthetic and semi-synthetic data, demonstrated that our approach significantly improves the estimation of survival distributions in terms of log-likelihood (which is a proper scoring for the sur-

vival analysis task) and $L1$ survival distance (which we proposed), compared to the standard approach that assumes conditional independence.

Preface

This thesis is an extension of work submitted to the Uncertainty in Artificial Intelligence (UAI) under the title "Copula-Based Deep Survival Models for Dependent Censoring". The submission was a collaborative effort led by Professor R. Greiner and included Michael Cooper and Rahul G. Krishnan in addition to myself. All of the methods presented in this thesis are my original work.

To my family

For constantly having faith in me, even during the periods when I lacked it myself.

All models are wrong, some are useful.

– George E.P. Box, 1976.

Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Russell Greiner for his support and insights throughout my MSc career. I would also like to thank Micheal Cooper for his help and support, Rahul G. Krishnan for his constructive comments, and Amii for the funding that allowed me to complete this research. I would like to acknowledge my committee members, Martha White, and Ivor Cribben, for taking time out of their busy schedules and reading my thesis.

Many thanks to my dear friends, Mohammad, Ramin, Amir, Afshin, Reza, Emad, Farzad, Parastoo, Keyvan, Marjan, and Mehrnoosh for their friendship and support.

Last but not least, I wanted to thank three people for always being there for me: My mother, who has been my first teacher, my father, for his unconditional support in all stages of my life, my brother, who is also my best friend and I could not imagine my life without him.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Clinical and epidemiological investigations often want to predict the time until the onset of an event of interest. As examples, (1) a clinical trial of a therapeutic cancer regimen may want to compare the time-to-mortality in patients who received experimental therapy, against that of the patients in the control arm who did not [16], and (2) a study developing a clinical risk score may want to regress the time until patient mortality onto certain covariates of interest, with the aim of leveraging the learned covariates as parameters in a predictive risk algorithm [28].

In such time-to-event prediction tasks, it is common to only have a bound of the time-to-event for some instances in the study cohort. Here, we focus on *right censored* instances – *e.g.* patients who left the study prior to observing their time of death (loss to follow up), or patients who did not die prior to the conclusion of the study (administrative censoring) [43], [44].

Survival prediction refers to the development of statistical models that support time-to-event prediction typically from data that includes censored instances. Rather than discarding such censored instances, methods in survival prediction leverage the censoring time as a *lower bound* on that individual's time-of-event [29]. Let $X^{(i)} \in \mathcal{X}$ refer to a patient's covariates, and let $T^{(i)}_{\text{obs}} \in \mathbb{R}_+$ refer to their time of last observation, taken to be the minimum of the (potentially unobserved) event time $T^{(i)}_E \in \mathbb{R}_+$ and censorship time $T^{(i)}_C \in \mathbb{R}_+$. A common assumption in survival analysis is *conditionally independent*

censoring [29] or censoring at random (CAR). This CAR assumption,

$$T_E \perp T_C \mid X. \tag{1.1}$$

assumes that, once X is known, knowing either the event or censoring time does not provide any additional information about the other quantity. Such an assumption is often unrealistically strong since we never get to know if we have included all of the covariates affecting the desired outcome in our vector of covariates X; it is also difficult to verify in practice because we only observe one outcome (either event or censorship) per instance, but never both. Figures 1.1(b) and 1.1(c) show that this relationship is violated when the event time affects the censoring time, or in the presence of unobserved confounding variables. When the conditional independence assumption of Equation 1.1 does not hold, we say that the data features *dependent censorship* or *censoring not at random* (CNAR), which is a common characteristic of survival data that is often unaccounted for in modern methods of survival prediction.

This is not just a theoretical concern: consider the use of risk scores in organ transplantation. Deceased-donor transplant organs are a scarce resource that represents a life-saving intervention for patients suffering from end-stage liver disease [38]. Under the Final Rule, the American federal policy of organ allocation, patients are prioritized according to their degree of hepatic dysfunction [15]. Implicit in the implementation of the Final Rule is the notion of a risk score, a function $R : \mathcal{X} \to \mathbb{R}_+$ that estimates the urgency of a patient's need for a deceased-donor transplant organ. This problem can be viewed as a survival analysis where the event of interest is patient death pre-transplant, and censoring is patient removal from the waitlist (for among other reasons, receiving a transplant). The current implementation of the Final Rule in the United States leverages the MELD-Na [2], a risk measure derived using survival analysis based on four biomarkers: serum creatinine, serum sodium, serum bilirubin, and INR (internal normalized ratio).

In reality, there are patient covariates not included in the MELD-Na that may affect both T_E and T_C, such that Equation 1.1 does not hold (*e.g.* patient sex). Specifically, serum creatinine, a biomarker that tends to be lower in

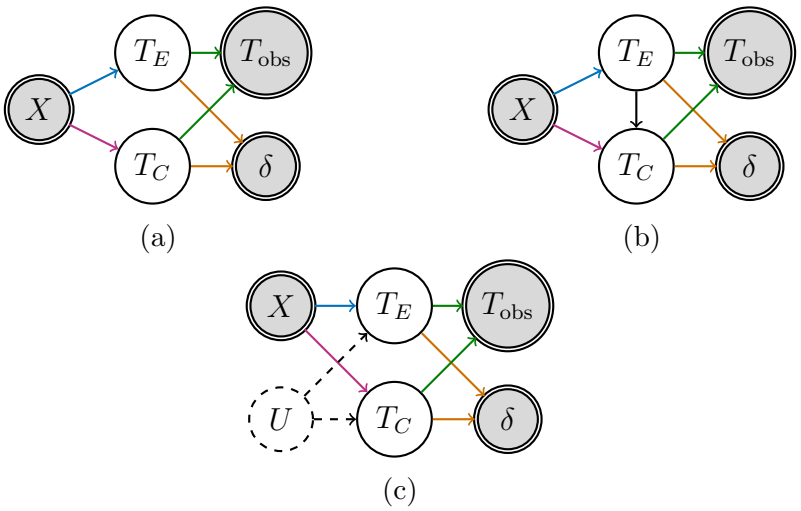

Figure 1.1: Graphical models of survival analysis, showing three different dependencies between covariates X, event/censorship times T_E/T_C, time of last observation T_{obs} and event indicator δ. Shaded nodes represent variables whose values we can observe. Blue and magenta arrows represent the event and censoring functions $f_E : \mathcal{X} \to \mathbb{R}_+$, $f_C : \mathcal{X} \to \mathbb{R}_+$, respectively, of arbitrary functional form. Green arrows into the T_{obs} node represent the function $\mathbb{R}_+^2 \to \mathbb{R}_+$ defined by $\min(t_e, t_c)$. Orange arrows into the δ node represent the indicator function $\mathbb{R}_+^2 \to \{0, 1\}$ defined by $\mathbb{1}[t_e < t_c]$. Graph (a) demonstrates the case of conditionally independent censoring (CAR) because conditioning on X d-separates [22] T_E and T_C – see 1.1. Graphs (b) and (c) represent cases in which the censoring and event times may be conditionally dependent (CNAR): in graph (b), this is through a direct dependency between T_E and T_C, while in graph (c), this is via the unobserved confounding node, U, that affects both T_E and T_C.

women than in men, plays a key role in the calculation of the MELD-Na score; consequently, Cholongitas *et al.* [8] argues that the MELD-Na systematically underestimates women's degree of medical urgency. Therefore, patient sex may influence T_E in a way that is not adequately captured by X, the MELD-Na covariates. Additionally, biological factors that make the female body habitus comparatively smaller reduce the pool of transplant organs available to women [39], [47]. By influencing the time at which a viable deceased-donor transplant organ can be found, patient sex may influence the censoring time, T_C. These factors suggest that the conditional independence assumption is dubious if the covariates X under consideration are only those of the MELD-Na.

Recent years have seen the emergence of a burgeoning subfield of survival analysis focused on relaxing the conditional independence assumption of Equation 1.1. However, existing approaches either do not permit the incorporation of covariates (*e.g.* [67], [54], [62]), or make strict assumptions over the form of the marginal distributions of f_{T_E} and f_{T_C} (*e.g.* [19]). These limitations do not admit the easy or direct application of these ideas to survival times modeled via nonlinear functions (such as neural networks) that are increasingly being used. In this vein, our work makes the following contributions:

1. We show how to leverage copulas (defined in Sec 2.2) to correct for dependent censorship in survival models. We present a parametric proportional hazards model that leverages neural networks to relax assumptions on the form of the risk function and employs copulas to model dependence in censoring. To our knowledge, this work represents the first neural network-based model of survival analysis to account for dependent censoring. Note that it is still a question to answer if using a neural network to estimate marginals could impact the identifiability of the problem.

2. We devise a method to learn both the model and dependence parameters from data.

3. We study the challenges associated with evaluating the performance of survival models under dependent censoring. We comment on the proprietary of

common evaluation metrics under dependent censoring and prove that the time-dependent concordance index and integrated Brier score are not capable of identifying the bias in survival models in the presence of dependent censoring.

1.1 Related Work

Note: Unless stated otherwise, these authors all evaluate the quality of a survival model based on the C-index.

Deep Learning in Survival Analysis: Linear models of survival analysis make the (often unrealistic) assumption that an individual's time-to-event is determined by a linear function of the covariates. Faraggi and Simon [20] presented the first neural-network-based model of survival, by incorporating a neural network into a Cox Proportional Hazards (CoxPH) model [10]. Although subsequent experimentation found the Farragi-Simon model unable to outperform its linear CoxPH counterpart [46], [65], DeepSurv [31] leveraged modern tools from deep learning such as SELU units [35] and the Adam optimizer [34] to learn a practical neural network-based CoxPH model that reliably outperformed the linear CoxPH on nonlinear outcome data. Since then, variations of neural network-based models of survival, such as DeepHit [41] (and its extension to time-varying data, Dynamic-DeepHit [40]), Deep Survival Machines [49], SuMo-net [53], Transformer-based survival models [26], [64], and methods based on Neural ODEs [59] have been introduced to model survival outcomes. Though these models successfully relax assumptions around the functional form of marginal risk, they do not jointly model the event and censoring times, a limitation that does not allow them to appropriately account for dependent censorship.

DeepSurv has enjoyed enduring success in part due to its broad applicability and strong performance on clinical data (*e.g.* [27], [33], [57]). Therefore, our investigation will focus on relaxing the conditional independence assumption in a parametric proportional hazards model. We decided to focus on parametric proportional hazard models for simplicity and we leave to future work

the relaxation of the conditional independence assumption in other classes of survival models.

Missing/Censored-Not-At-Random Data and Identification: Since we do not simultaneously observe T_E and T_C, we can treat the problem of survival analysis as a missing data problem. The standard taxonomy of missing data [55], [61] partitions variables into one of three classes: *missing completely at random (MCAR)* where the missingness process is determined only by randomness, *missing at random (MAR)* where the missingness process is determined by randomness and/or observed covariates, and *missing not at random (MNAR)* where the missingness process may depend on unobserved variables (such as unobserved confounding or self-masking). Similarly, censorship in survival analysis can take place *completely at random (CCAR)*, *at random (CAR)*, or *not at random (CNAR)* [44], [45]. The conditional independence assumption of Equation 1.1 is equivalent to asserting either CCAR or CAR in the data.

MNAR data, in the general case, is non-identifiable which means that there is no unique answer to the problem we are solving [48]; but survival analysis imposes stronger assumptions on the data than general models of missing data, since observed event time acts as a lower bound for unobserved event time (in the case of censored data). Therefore, prior work has focused on investigating the scenarios in which model parameters of survival data can be uniquely identified. Tsiatis [60] established that, in the general case, the joint distribution over M variables, $\Pr(T_1, ..., T_M)$ is not generally identifiable from observations of the random variable $T = \min(T_1, ..., T_M)$; although if the joint distribution is defined in terms of a known copula C, which is a mathematical tool to define dependence structure between random variables, and the marginals are continuous, then identifiability holds [6], [68]. Crowder [12] extended the work of Tsiatis, showing that even if all the marginal distributions $f_1, ..., f_M$ are known, the joint distribution remains non-identifiable. Research has since defined tuples of marginals and copulas for which the joint distribution is identifiable. Notably, Schwarz *et al.*[56] focus on the bivariate case, and prove that if the marginals f_E and f_C are known, several sub-classes of

Archimedean copulas are identifiable. Zheng and Klein [68], Carrière [6] highlight conditions for identifiability when the form and parameter of the copula are known *a priori*. Schwarz *et al.* [56] categorize copulas into sub-classes wherein the ground-truth copula, C_{θ^*}, is identifiable. Our current analysis does not touch upon the identifiability of the joint distribution in the context of neural network based models of survival outcomes though the success of our method does suggest this as an important area for future study. Many machine learning models remain non-identified [3] while remaining useful as predictive and descriptive models. Our method is similar in this respect.

Copula-Based Models of Dependent Censoring: Prior literature has leveraged copulas to model the relationship between the event and censoring distributions in order to account for the effect of dependent censoring [18]. To our knowledge, the first such work was that of Zheng and Klein [67] and Rivest and Wells [54], whose development of the nonparametric Copula-Graphic Estimator extended the Kaplan-Meier Estimator [30] to cases where the dependence between T_E and T_C takes the form of an assumed copula (both form (C) and parameter of the copula (θ) assumed to be known). Though parametric estimators for this problem have been proposed in prior literature, they tend to make strict assumptions over the distributional form of $f_{T|X}$ (*e.g.* that it is a linear-Weibull function [19][1]). Proposed semi-parametric estimators [7], [13], [17] suffer from the same problem, as both of these approaches assume that the hazard is a linear function of the instance covariates. To our knowledge, no such copula-based model exists to accommodate more complex relationships between covariates and risk while also accounting for dependent censoring. This is the gap our research aims to fill.

[1] Although Escarela does not directly model dependent censoring, but rather dependent competing events, the approach can be directly extended to this domain.

Chapter 2

Background

In this chapter, we present the basics of survival analysis and introduce the concept of copula as a tool to account for dependency between marginal distributions.

2.1 Survival Analysis

2.1.1 Survival Time and Censoring

Survival time, in the context of survival analysis, is defined as the duration between the initial time point of observation and the occurrence of the event of interest. In medical studies, the initial time point can correspond to the initiation of a treatment regimen, the date of surgical intervention, or the date of admission to a hospital, while the event of interest may involve observing an improvement after treatment, death following surgery, or discharge from the hospital.

In the course of studying a survival problem, it is possible that the occurrence of the event of interest is not observed in certain samples. This can arise due to a range of factors, including premature study termination, loss of contact with participants who are no longer interested in participating in the study, or the occurrence of other events preceding the event of interest. Such a phenomenon is referred to as censoring [36]. Censoring is typically classified into three principal categories [42]: right censoring, left censoring, and interval censoring.

- Right censoring occurs when the observed survival time is either equal

to or less than the actual survival time.

- Left censoring refers to a situation where the observed survival time is either equal to or greater than the actual survival time.

- Interval censoring is a circumstance where the precise timing of the occurrence of the event of interest is not known, and it is only known that the event occurred at some point between two distinct known time points.

Of the different censoring scenarios described, right censoring is the most frequently encountered in real-world settings. This thesis will analyze the survival data when some instances leave the study due to right censoring.
In a survival analysis problem, the occurrence time of an event of interest and the time of censoring for the i-th subject are denoted by T_i and C_i, respectively, with x_i representing the corresponding covariate vector. In the presence of censoring, we observe exactly one of T_i or C_i. Specifically, if the event of interest occurs before censoring, then only T_i is observed ($T_i \leq C_i$). Conversely, if censoring happens prior to the event of interest, then only C_i is observed ($T_i > C_i$). In the latter situation, the sole information available about T_i is that it is greater than C_i. Thus, a survival dataset is typically represented as (x_i, y_i, δ_i), where:

- x_i: vector of covariates

- y_i: The time of occurrence of the event of interest or the time of censoring, whichever comes first.

- δ_i: Censoring indicator ($\delta_i = 1$ if $y_i = T_i$, $\delta_i = 0$ if $y_i = C_i$)

Survival models that operate under the assumption of *independent censoring* postulate that the random variables T_i and C_i are statistically independent, given the covariate vector x_i. In mathematical terms, this assumption is expressed as $Pr(T_i, C_i|x_i) = Pr(T_i|x_i)\ Pr(C_i|x_i)$.

Figure 2.1 provides an illustrative example for understanding survival data and the associated challenges. The study depicted in Figure 2.1 spans over

12 months and involves 6 participants. As shown in the figure, only two participants (S_4 and S_6) experienced the event of interest (denoted by X), while the remaining participants were censored (denoted by red dots). Specifically, subjects S_2 and S_6 are censored due to the end of the study, whereas subjects S_1 and S_3 are censored due to withdrawal or loss of follow-up.

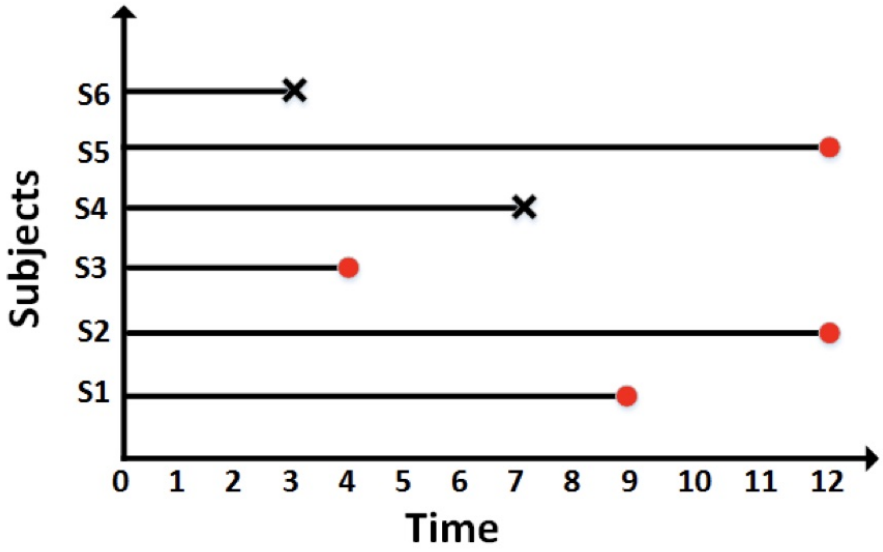

Figure 2.1: An illustration of survival data and problem

2.1.2 Event Time Distributions

There are several ways to specify the probability distribution over the non-negative continuous random variable T, which represents the event time. In this context, we will discuss three such methods that are commonly utilized in survival analysis.

The survival function [36], [42] can be defined as the probability that the event time exceeds a certain value t, given by the expression:

$$S(t) = Pr(T > t), \quad 0 \leq t < \infty \ . \tag{2.1}$$

Note the survival function can be expressed as:

$$S(t) = \ Pr(T \geq t) \ = \ 1 - Pr(t < T) \ = \ 1 - CDF(t) \ . \tag{2.2}$$

The probability density function (PDF) can be obtained using Equation 2.2 as:

$$f(t) = -\frac{dS(t)}{dt} \ . \tag{2.3}$$

Another concept in survival analysis is the *hazard function* [14], which represents the instantaneous rate of occurrence of the event of interest for an instance that has survived up to time t.

$$\lambda(t) = \lim_{h \to 0} \frac{Pr(t \leq T < t + h \mid T \geq t)}{h} \ . \tag{2.4}$$

By using the definition of PDF given in Equation 2.3, we can obtain the following expression for small intervals (h):

$$f(t)h \ \simeq \ Pr(t \leq T < t + h) \ = \ S(t) - S(t + h) \ . \tag{2.5}$$

Based on Equation 2.5, the hazard function can be defined as:

$$\lambda(t) = f(t)/S(t) = -\frac{d \log(S(t))}{dt} \ . \tag{2.6}$$

By integrating both sides of Equation 2.6 with respect to t and considering $S(0) = 1$, we obtain:

$$\begin{aligned} S(t) =& exp(-\int_0^t \lambda(s)ds) \\ =& \ exp(-\Lambda(t)) \end{aligned} \ . \tag{2.7}$$

$\Lambda(t) = \int_0^t \lambda(s)ds$, is known as the cumulative hazard function [42].

2.1.3 Survival Analysis Methods

Survival analysis methods can be broadly classified into three categories: Non-parametric, Semi-parametric, and Parametric methods.

1. Non-parametric methods are employed when there is no prior assumption regarding the underlying distribution of event times or when covariates are unavailable. Two of the simplest non-parametric methods are the Kaplan-Meier estimator (KM) [30] and the Nelson-Aalen estimator (NA) [1], [51], [52]. These methods work on a population level

and do not consider the covariates of each sample. There are also non-parametric methods that can work on an individual level by including patients covariates such as DeepHit [41] and MTLR [66].

2. The semi-parametric category of survival analysis methods is led by the widely used Cox Regression method [10], which is built on the proportional hazard assumption [10] and partial likelihood [11]. This method involves using a non-parametric approach to determine the baseline cumulative hazard function and a regression method to estimate the *risk* associated with each instance.

3. Parametric methods are employed when there is an assumption about the underlying distribution of event times. These methods assume that event times originate from a fully parametric known distribution, such as the Exponential or Weibull distribution; user can then estimate the parameters of these distributions based on the covariates [42].

In the remainder of this section, we provide an example from each of the three categories mentioned above:

Kaplan-Miere

The Kaplan-Meier estimator is one of the simplest methods used in survival analysis, which belongs to the non-parametric family of methods and assumes independent censorship. The derivation of the Kaplan-Meier estimator is presented below:

$$
\begin{aligned}
\hat{S}(t) &= Pr(T > t) \\
&= \prod_{t_i \leq t, \delta_i = 1} (1 - \frac{Pr(T = t_i)}{Pr(T \geq t_i)}) \\
&= \prod_{t_i \leq t, \delta_i = 1} (1 - \frac{Pr(T = t_i, C \geq t_i)}{Pr(T \geq t_i, C \geq t_i)}) \\
&= \prod_{t_i \leq t, \delta_i = 1} (1 - \frac{\sum_{l=1}^{n} \mathbf{1}(t_l = t, \delta_l = 1)/n}{\sum_{l=1}^{n} \mathbf{1}(t_l \geq t)/n}) \\
&= \prod_{t_i \leq t, \delta_i = 1} (1 - \frac{d_i}{n_i})
\end{aligned}
\quad . \tag{2.8}
$$

The estimator is expressed as a step function with jumps at the times an event occurs, given by the formula $\hat{S}(t) = \prod_{t_i \leq t, \delta_i = 1} (\frac{n_i - d_i}{n_i})$, where n_i is the number of instances at risk at time t_i (instances who have not experienced the event or censoring yet) and d_i is the number of events that occur at time t_i. The KM estimator is capable of handling both uncensored and (right) censored data. As we mentioned before, this method does not include covariates and provides a curve on a population level.

Figure 2.2 depicts two different scenarios using the KM estimator, where for the blue curve, all of the samples have experienced the event of interest while for the black curve, four instances have experienced censoring at time points indicated with (+). Tables 2.1 and 2.2 provide datasets used to generate the blue and black curves respectively.

Cox-Regression

Cox-regression, belonging to the second family of methods mentioned earlier, incorporates the vector of covariates for each patient to predict their corresponding survival rate.

To integrate covariates in survival analysis, patient-specific hazard function can be defined as $\lambda(t|x) = \lim_{h \to 0} \Pr(t \leq T < t + h | T \geq t, x)/h$, where x represents the vector of covariates of each patient used to predict the corresponding survival rate. The Cox proportional hazard model is one of the methods used in this context, which assumes that the hazard function of an individual is proportional to the hazard function of the population at all times.

Table 2.1: Dataset for KM curve without censoring

Patient's ID	time	δ
1	1	1
2	2	1
3	2	1
4	2	1
5	4	1
6	5	1
7	6	1
8	6	1
9	8	1
10	9	1

Table 2.2: Dataset for KM curve with censoring

Patient's ID	time	δ
1	1	0
2	2	1
3	2	1
4	2	1
5	4	0
6	5	0
7	6	1
8	6	1
9	8	1
10	9	0

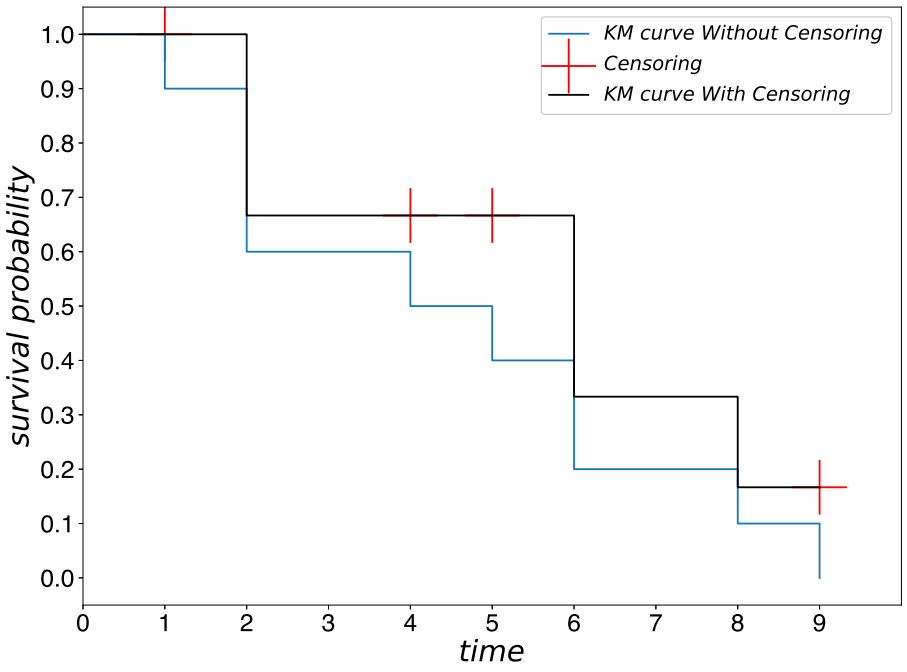

Figure 2.2: KM curve without censoring

Mathematically, the model is expressed as below in terms of hazard function:

$$\lambda(t|x_i) = \lambda_0(t)\exp(\beta^T x_i) \tag{2.9}$$

The non-parametric approach is used to obtain the baseline cumulative hazard

function Λ_0, whereas the vector of coefficients β can be determined independently of Λ_0 by maximizing the partial likelihood function. Specifically, the partial likelihood function can be expressed as [11]:

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta^T x_i)}{\sum_{l \in R_i} \exp(\beta^T x_l)}\right)^{\delta_i} . \tag{2.10}$$

The Breslow estimator [4], which is defined as:

$$\Lambda(t) = \sum_{i:t_i \leq t} \left(\frac{d_i}{\sum_{l \in R_i} \exp(\beta^T x_l)}\right) . \tag{2.11}$$

can be used to obtain the baseline cumulative hazard function $\Lambda_0(t)$, where R_i denotes the set of patients at risk at time t_i, and d_i is the number of events at time t_i. As we can see, in comparison with KM method, Cox-Regression includes patients covariates in the computation and provides and provides survival curves on an individual level.

Parametric Models

As previously stated, a parametric model, assumes that the distribution of event times comes from a specific family. Our objective is to determine the parameters of the distribution based on the covariates. Two of the most commonly utilized distributions in survival analysis are the Exponential and Weibull distributions. After introducing these distributions we will explain how we can include covariates in a parametric model.

- **Exponential Distribution** [42]: The hazard function for this distribution is invariant with respect to time.

$$\begin{aligned} \lambda(t) &= \lambda \quad \forall t > 0 \\ S(t) &= \exp(-\lambda t) \end{aligned} . \tag{2.12}$$

- **Weibull Distribution** [42]: This is an extension of the Exponential distribution that permits the hazard function to be dependent on time. It is a two-parameter distribution with the hazard and survival function defined as:

$$\begin{aligned} \lambda(t) &= \left(\frac{\gamma}{\rho}\right)\left(\frac{t}{\rho}\right)^{\gamma-1} \\ S(t) &= \exp\left(-\left(\frac{t}{\rho}\right)^{\gamma}\right) \end{aligned} . \tag{2.13}$$

In the case of parametric models, assuming proportional hazard is a common approach. Consequently, the hazard function will be in the form of:

$$\lambda(t|x) = \lambda_0(t) \exp\left(f\left(x\right)\right) \ . \tag{2.14}$$

In this context, $f(x)$ is commonly referred to as the risk function. As a result, the cumulative hazard function will be expressed as:

$$\Lambda(t|x) = \Lambda_0(t) \exp\left(f\left(x\right)\right) \ . \tag{2.15}$$

The parameters of the model may be derived by maximizing the log-likelihood function, which will be discussed in Chapter 3. For this thesis, we will utilize a parametric model featuring a Weibull distribution as the baseline distribution, assuming proportional hazards.

2.2 Copula

A copula refers to a function that connects two random variables by defining their dependence structure. It provides a way to separate the marginal distribution of each random variable from the dependence structure between them. The word "copula" is derived from the Latin term "copulare," meaning "to join together" [50]. Abe Sklar introduced the term "copula" in his research on probabilistic metric space, where he presented a mathematical definition of copulas and established the Sklar theorem, the most fundamental theorem concerning copulas [58]. Copulas can handle any number of variables, but in this thesis, our focus is on bivariate copulas.

To illustrate how copulas relate to survival analysis, we will start with a simple example. Suppose all patients leave the study one day before the event of interest (death) occurs. In this scenario, the time of censoring is strongly dependent on the event time and can be modeled using a copula. However, this is an extreme case. In reality, there may be some variability in both the time of censoring and the event time, which decreases as the study duration increases. This type of dependence can be modeled using a Clayton copula, which we will introduce later.

Bivariate Copula

A bivariate copula is a distribution function that involves two variables and has uniform marginals ranging from 0 to 1. Let $C_\theta : [0,1]^2 \to [0,1]$ be a copula with a dependence parameter θ. By definition [50], any copula must satisfy the following conditions:

- $C_\theta(0,v) = C_\theta(u,0) = 0$ for $0 \leq u \leq 1$ and $0 \leq v \leq 1$
- $C_\theta(1,v) = v$ and $C_\theta(u,1) = u$ for $0 \leq u \leq 1$ and $0 \leq v \leq 1$
- $C_\theta(u_2,v_2) - C_\theta(u_2,v_1) - C_\theta(u_1,v_2) + C_\theta(u_1,v_1) \geq 0$ for $0 \leq u1 \leq u2 \leq 1$ and $0 \leq v1 \leq v2 \leq 1$

The initial two conditions demand that the marginals are uniformly distributed, while the final condition states that the probability mass on any rectangular region $[u1,u2] \times [v1,v2]$ is non-negative.
Additionally, we define the partial derivatives of the copula function as:

- $C_\theta^{[1,0]}(u,v) = \frac{\partial}{\partial u} C_\theta(u,v)$
- $C_\theta^{[0,1]}(u,v) = \frac{\partial}{\partial v} C_\theta(u,v)$
- $C_\theta^{[1,1]}(u,v) = \frac{\partial^2}{\partial u \partial v} C_\theta(u,v)$

Archimedean Copulas

Archimedean copulas are a class of copulas that are defined by [50]:

$$C_\theta(u,v) = \phi^{-1}\left(\phi(u) + \phi(v)\right) \tag{2.16}$$

for some strictly decreasing function $\phi : [0,1] \to [0,\infty]$ which is referred to as the *generator of the copula*. The Clayton [9], Frank [21], and independent copulas are a few examples of this family of copulas.

Kendall's tau(τ)

Kendall's tau (τ) [32] is a well-known measure used to assess the dependence between two random variables U and V:

$$\tau = \Pr\left\{(U_2 - U_1)(V_2 - V_1) > 0\right\} - \Pr\left\{(U_2 - U_1)(V_2 - V_1) < 0\right\}. \tag{2.17}$$

Table 2.3: Examples of Archimedean copulas

Copula	**Closed Form:** $C_\theta(u,v)$	**Generator:** $\phi(t)$	**Parmeter**
Independent	uv	$ln(t)$	NA
Clayton	$(u^{-\theta}+v^{-\theta}-1)^{-1/\theta}$	$(t^\theta-1)/\theta$	$\theta>0$
Frank	$-\frac{1}{\theta}\log\{1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\}$	$-\log(\frac{e^{-\theta t}-1}{e^{-\theta}-1})$	$\theta\neq 0$

Note this is independent of the marginals from which U and V are sampled. For an Archimedean copula, Kendall's tau is equal to [50]:

$$\tau(C_\theta(u,v)) = 4\int_0^1\int_0^1 C_\theta(u,v)\frac{\partial^2}{\partial u\partial v}C_\theta(u,v)\,du\,dv - 1\ . \tag{2.18}$$

To provide a better understanding of Kendall's τ, let's revisit our previous example. In this case, Kendall's τ measures the probability of sample i censoring earlier than sample j, given that the event occurred earlier for sample i. In other words, it quantifies the likelihood of the two samples being ordered similarly with respect to the time of censoring, based on their respective event times.

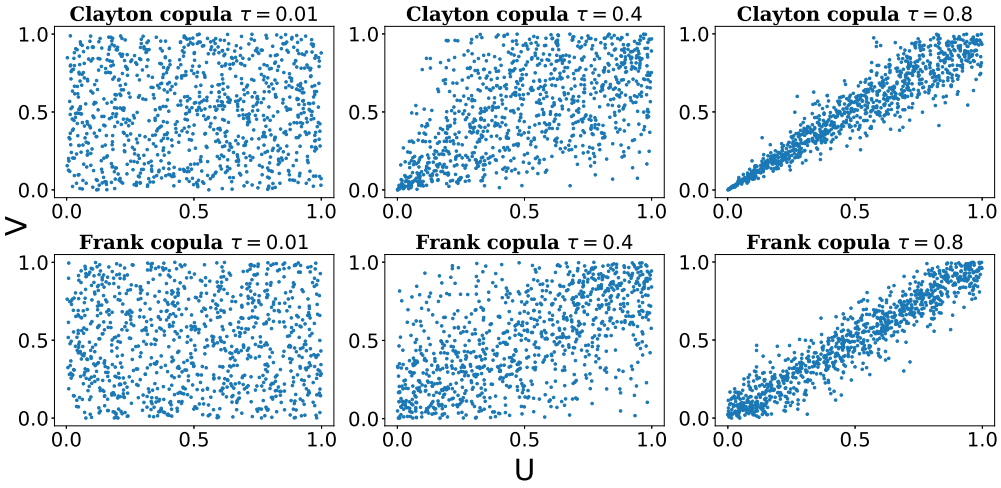

Figure 2.3: Scatter plot of 1000 pairs of (u,v) generated from Clayton and Frank Copula for $\tau = 0.01$, $\tau = 0.4$, and $\tau = 0.8$

Note that the parameter of a copula is denoted as θ, while the strength of dependence is measured using Kendall's τ. For a given copula, each value of θ corresponds to a specific value of τ. For instance, in the case of the Clayton

copula, we have $\tau = \theta/(\theta+2)$. However, the same θ value in a different family of copulas may not represent the same level of dependence. Therefore, in this thesis, we adopt Kendall's τ as a standard metric to evaluate the degree of dependence in a dataset. The structure of the dependency between pairs sampled from a copula can be influenced by the type of copula and the strength of the dependency(τ), as shown in Figure 2.3.

Dependent Censoring Example

We conclude this chapter with an example that illustrates the distinction between dependent and independent censoring. Returning to our initial scenario where all patients exit the study the day before experiencing the event if we assume independent censoring, our censoring model will be accurate, but the model for the event will indicate a constant survival probability of 1 at all time points. However, if we acknowledge the strong dependence between censoring and the event, we can construct a model for the event that is nearly identical to the censoring model. Both event models will perform similarly under standard survival metrics, such as the IBS [5] and C-index [23], [25], [63] (introduced in Chapter 3). Similar performance in terms of survival metrics indicates the shortcoming of these metrics to identify the bias in survival curves we are trying to address in this thesis. This example is an extreme case that serves to clarify the problem addressed in this thesis.

Chapter 3

Methodology

This thesis will explore ways to incorporate dependent censorship into any given survival model. We define our task as learning individual survival curves, for both event and censoring, given a survival dataset in which we do not know if independent censorship is a valid assumption. Our framework is capable of estimating the true parameters of underlying distributions based on covariates for both event of interest and censoring in addition to recovering the strength of dependency in the dataset. In the context of this thesis, we have made the following assumptions:

Assumption 1 (Known Form of the Copula)**.** *We assume prior knowledge of the functional form of the copula (e.g. that C_{θ^*}, the copula associated with the data-generating process, is a Clayton copula).*

Assumption 2 (Proportional Hazards [10])**.** *The hazard function of each outcome, i.e., event or censorship, can be expressed as a composite of a base-line hazard, denoted by $\lambda_0(t)$, which is solely dependent on time, and a risk function, denoted by $g(x)$, which is dependent only on the covariates X. More specifically, there exist appropriate choices of λ_0 and g such that the hazard function conditional on X, denoted by $h_{T|X}(t|x)$, can be represented as $h_{T|X}(t|x) = \lambda_0(t)\exp(g(x))$.*

3.1 Models

Our methodology involves modeling each possible outcome, namely event, and censorship, separately using a modified version of a model from the Cox Proportional Hazards family, and linking them through a copula in the likelihood function during the training process. With regard to Assumption 2, we have developed two distinct models:

- **Parametric Weibull:** We assume that the underlying distribution follows a Weibull distribution and that the risk score for each instance can be described by a linear model over the covariates.
- **Non-Parametric Weibull:** Our assumption is that the underlying distribution is Weibull and that the risk score lacks a known general form.

Note that the Exponential distribution is a particular instance of the more general Weibull distribution. Thus, our methodology encompasses both parametric and non-parametric versions of the Exponential distribution.

Let $\lambda_0(t)$ denote the baseline hazard of the model, and let $g_\psi : \mathcal{X} \to \mathbb{R}$ represent a function (such as a neural network or a parametric model) with parameters ψ that maps the covariate space $\mathcal{X}$ to the real line. Exploiting the proportional hazards assumption, we construct our model in terms of its hazard function:

$$\hat{h}_{T|X}(t|x;\psi) = \hat{\lambda}_0(t) \exp\left(g_\psi\left(x\right)\right) \ . \tag{3.1}$$

Through rearranging Equation 3.1, this category of models naturally yields estimates of the survival function, denoted by $\hat{S}(T|X)$, and the corresponding probability mass function, denoted by $\hat{f}(T|X)$, for each outcome. These estimates are beneficial in performing maximum likelihood estimation.

$$\begin{aligned} \hat{S}_{T|X}\left(t|x;\psi\right) = & \ \exp\left(-\hat{\Lambda}\left(t|x;\psi\right)\right) \\ & = \exp\left(-\hat{\Lambda}_0\left(t\right)\exp\left(g_\psi\left(x\right)\right)\right) \end{aligned} \tag{3.2}$$

$$\begin{aligned} \hat{f}_{T|X}\left(t|x;\psi\right) &= \hat{S}_{T|X}\left(t|x;\psi\right) \ \hat{h}_{T|X}\left(t|x;\psi\right) \\ &= \exp\left(-\hat{\Lambda}_0\left(t\right)\exp\left(g_\psi\left(x\right)\right)\right)\hat{\lambda}_0\left(t\right)\exp\left(g_\psi\left(x\right)\right) \end{aligned} \tag{3.3}$$

By substituting the general form of the hazard function for the Weibull distribution in place of $\lambda_0(t)$, we can derive the equations presented above for the Weibull distribution.

$$\hat{h}_{T|X}\left(t|x;\rho,\gamma,\psi\right) = \left(\frac{\gamma}{\rho}\right)\left(\frac{t}{\rho}\right)^{\gamma-1}\exp\left(g_\psi\left(x\right)\right) \tag{3.4}$$

$$\hat{S}_{T|X}\left(t|x;\rho,\gamma,\psi\right) = \exp\left(-\left(\frac{t}{\rho}\right)^{\gamma}\exp\left(g_\psi\left(x\right)\right)\right) \tag{3.5}$$

3.2 Maximum Likelihood

Consider a dataset $\left\{\left(X^{(i)}, T_{\text{obs}}^{(i)}, \delta^{(i)}\right)\right\}_{i=1}^{N}$ comprising N independent and identically distributed (i.i.d.) draws from a data-generating distribution $\mathcal{D}$. Each draw consists of a set of baseline covariates $X^{(i)} \in \mathcal{X}$, the time of last observation $T_{\text{obs}}^{(i)} \in \mathbb{R}^{\geq 0}$, and an event indicator $\delta^{(i)} \in \{0,1\}$, which takes the value $\mathbf{1}\{T_E^{(i)} < T_C^{(i)}\}$, where $T_E^{(i)} \in \mathbb{R}^{\geq 0}$ and $T_C^{(i)} \in \mathbb{R}^{\geq 0}$ are the event and censoring times, respectively. We define $T\text{obs}^{(i)}$ to be the minimum of $T_E^{(i)}$ and $T_C^{(i)}$. This will help us to understand the difference between models trained under the independent censoring assumption and our model. It also gives us an insight into why the model learned under the independence assumption can be biased.

In the remainder of this section, we present a complete derivation of the likelihood of a survival dataset subject to dependent censoring, which is characterized by a copula C.

To begin, we provide a general expression for the survival likelihood. Next, for comparison purposes, we derive the survival likelihood assuming conditional independence, which will illustrate how the likelihood factorizes in a clear manner under this assumption.

Next, we will provide Lemma 1, which enables us to determine the survival likelihood in cases where dependence is defined by a specified copula. Then, we will utilize Lemma 1 to the general form of the survival likelihood, which will produce the learning objective utilized in our dissertation.

3.2.1 The General Likelihood of Survival Data under Right-Censorship

To set the foundation for the following derivations, we explain the reasoning behind the overall probability for survival data that has been right-censored and introduce its expression in Equation 3.6.

Let us consider a survival dataset that comprises N independent and identically distributed samples of the form $\left\{\left(X^{(i)}, T^{(i)}_{\text{obs}}, \delta^{(i)}\right)\right\}_{i=1}^{N} \subset \left(\mathcal{X} \times \mathbb{R}^{\geq 0} \times \{0,1\}\right)^N$. The likelihood, as shown in Equation 3.6, incorporates the $\delta^{(i)}$ terms in the exponent as a binary filter that acts conditionally: raising a term to the power of $\delta^{(i)}$ ensures that it is only non-degenerate when the patient experiences an event while raising a term to the power of $1 - \delta^{(i)}$ ensures that it is only non-degenerate when the patient is censored.

The joint density function of the event and censoring times, conditional on the patients' covariates, can be denoted by $f_{T_E, T_C \mid X}$. We aim to optimize the likelihood for a given patient i in two possible scenarios:

1. **Case 1 (Uncensored):** If $\delta^{(i)} = 1$, maximize the likelihood that $T_E = T^{(i)}_{\text{obs}}$, and $T^{(i)}_C > T^{(i)}_{\text{obs}}$. This corresponds to the observation that the patient experienced the event at time $T^{(i)}_{\text{obs}}$, and was not censored prior to experiencing the event. The probability mass of this likelihood under our density function is $\int_{T^{(i)}_{\text{obs}}}^{\infty} f_{T_E, T_C \mid X}(T^{(i)}_{\text{obs}}, t_c \mid X^{(i)})\, dt_c$.

2. **Case 2 (Censored):** If $\delta^{(i)} = 0$, maximize the likelihood that $T^{(i)}_C = T^{(i)}_{\text{obs}}$, and $T_E > T^{(i)}_{\text{obs}}$. This corresponds to the observation that the patient is censored at time $T^{(i)}_{\text{obs}}$, and did not experience an event prior to being censored. The probability mass of this likelihood under our density function is $\int_{T^{(i)}_{\text{obs}}}^{\infty} f_{T_E, T_C \mid X}(t_e, T^{(i)}_{\text{obs}} \mid X^{(i)})\, dt_e$.

Combining these two cases, and applying the assumption that our data is independent and identically distributed (i.i.d.), yields the following form of the likelihood:

$$\mathcal{L}(\mathcal{D}) = \prod_{i=1}^{N} \underbrace{\left[\int_{T_{\text{obs}}^{(i)}}^{\infty} f_{T_E, T_C \mid X}(T_{\text{obs}}^{(i)}, t_c \mid X^{(i)})\, dt_c \right]^{\delta^{(i)}}}_{\Pr\left(T_E = T_{\text{obs}}^{(i)}, T_C > T_{\text{obs}}^{(i)} \mid X^{(i)}\right)} \cdot \underbrace{\left[\int_{T_{\text{obs}}^{(i)}}^{\infty} f_{T_E, T_C \mid X}(t_e, T_{\text{obs}}^{(i)} \mid X^{(i)})\, dt_e \right]^{1-\delta^{(i)}}}_{\Pr\left(T_C = T_{\text{obs}}^{(i)}, T_E > T_{\text{obs}}^{(i)} \mid X^{(i)}\right)}. \tag{3.6}$$

3.2.2 Derivation of the Likelihood Under Conditional Independence

To derive the likelihood under conditional independence, we start from the general form of the likelihood, as expressed in Equation 3.6. Under the assumption that $T_E \perp T_C \mid X$, the distribution $f_{T_E, T_C \mid X}$ factorizes into $f_{T_E \mid X} f_{T_C \mid X}$, yielding:

$$\begin{aligned}
\mathcal{L}(\mathcal{D}) &= \prod_{i=1}^{N} \left[f_{T_E \mid X}\left(T_{\text{obs}}^{(i)} \mid X^{(i)}\right) \int_{T_{\text{obs}}^{(i)}}^{\infty} f_{T_C \mid X}\left(t_c \mid X^{(i)}\right)\, dt_c \right]^{\delta^{(i)}} \\
&\qquad \left[f_{T_C \mid X}\left(T_{\text{obs}}^{(i)} \mid X^{(i)}\right) \int_{T_{\text{obs}}^{(i)}}^{\infty} f_{T_E \mid X}(t_e \mid X^{(i)})\, dt_e \right]^{1-\delta^{(i)}} \\
&= \prod_{i=1}^{N} \left[f_{T_E \mid X}\left(T_{\text{obs}}^{(i)} \mid X^{(i)}\right) \left(1 - F_{T_C \mid X}\left(T_{\text{obs}}^{(i)} \mid X^{(i)}\right)\right) \right]^{\delta^{(i)}} \\
&\qquad \left[f_{T_C \mid X}\left(T_{\text{obs}}^{(i)} \mid X^{(i)}\right) \left(1 - F_{T_E \mid X}(T_{\text{obs}}^{(i)} \mid X^{(i)})\right) \right]^{1-\delta^{(i)}} \\
&= \prod_{i=1}^{N} \left[f_{T_E \mid X}\left(T_{\text{obs}}^{(i)} \mid X^{(i)}\right) S_{T_C \mid X}\left(T_{\text{obs}}^{(i)} \mid X^{(i)}\right) \right]^{\delta^{(i)}} \\
&\qquad \left[f_{T_C \mid X}\left(T_{\text{obs}}^{(i)} \mid X^{(i)}\right) S_{T_E \mid X}\left(T_{\text{obs}}^{(i)} \mid X^{(i)}\right) \right]^{1-\delta^{(i)}}
\end{aligned} \tag{3.7}$$

3.2.3 Derivation of the Likelihood Under Dependence

First, we start with Sklar's Theorem [58]. Then we will use this theorem to estimate the conditional survival function.

Theorem 1. *Sklar's Theorem Let H be a joint distribution function with margins F and G. Then there exists a copula C such that $\forall x, y \in R$:*

$$H(x, y) = C(F(x), G(y)) \ . \tag{3.8}$$

Lemma 1 (Conditional Survival Function using Sklar's Theorem)**.** *If* $S_{T_E,T_C \mid X}(t_e, t_c \mid x) = C(u_1, u_2)\Big|_{\substack{u_1=S_{T_E \mid X}(t_e \mid x) \\ u_2=S_{T_C \mid X}(t_c \mid x)}}$, *then,*

$$\int_{t_c}^{\infty} f_{T_C \mid T_E,X}(t_c \mid t_e, x) = \frac{\partial}{\partial u_1} C(u_1, u_2)\Big|_{\substack{u_1=S_{T_E \mid X}(t_e \mid x) \\ u_2=S_{T_C \mid X}(t_c \mid x)}} \tag{3.9}$$

Proof.

$$\begin{aligned}
\int_{t_c}^{\infty} f_{T_C \mid T_E,X}(t_c \mid t_e, x) &= \frac{\int_{t_c}^{\infty} f_{T_C,T_E \mid X}(t_c, t_e \mid x) dt_c}{f_{T_E \mid X}(t_e \mid x)} \\
&= \frac{\frac{-\partial}{\partial T_E} \int_{t_e}^{\infty} \int_{t_c}^{\infty} f_{T_C,T_E \mid X}(t_c, t_e \mid x) dt_c dt_e}{f_{T_E \mid X}(t_e \mid x)} \\
&= \frac{\frac{-\partial}{\partial T_E} S_{T_C,T_E \mid X}(t_c, t_e \mid x)}{f_{T_E \mid X}(t_e \mid x)}
\end{aligned} \tag{3.10}$$

Applying Sklar's Theorem [58] to the numerator:

$$= \frac{\frac{-\partial}{\partial T_E} \left(C(u_1, u_2)\Big|_{\substack{u_1=S_{T_E \mid X}(t_e \mid x) \\ u_2=S_{T_C \mid X}(t_c \mid x)}} \right)}{f_{T_E \mid X}(t_e \mid x)} \ , \tag{3.11}$$

Chain rule of differentiation:

$$\begin{aligned}
&= \frac{\frac{-\partial}{\partial u_1} \left(C(u_1, u_2)\Big|_{\substack{u_1=S_{T_E \mid X}(t_e \mid x) \\ u_2=S_{T_C \mid X}(t_c \mid x)}} \right) \frac{\partial}{\partial T_E} S_{T_E \mid X}(t_e \mid x)}{f_{T_E \mid X}(t_e \mid x)} \\
&= \frac{-\partial}{\partial u_1} \left(C(u_1, u_2)\Big|_{\substack{u_1=S_{T_E \mid X}(t_e \mid x) \\ u_2=S_{T_C \mid X}(t_c \mid x)}} \right) \frac{-f_{T_E \mid X}(t_e \mid x)}{f_{T_E \mid X}(t_e \mid x)} \ . \\
&= \frac{\partial}{\partial u_1} \left(C(u_1, u_2)\Big|_{\substack{u_1=S_{T_E \mid X}(t_e \mid x) \\ u_2=S_{T_C \mid X}(t_c \mid x)}} \right)
\end{aligned} \tag{3.12}$$

□

The application of this lemma to the censored case – modeling $f_{T_E \mid T_C, X}$ – follows immediately from a symmetric argument. That argument yields:

$$\int_{t_c}^{\infty} f_{T_E \mid T_C, X}(t_e \mid t_c, x) = \frac{\partial}{\partial u_2} C(u_1, u_2)\Big|_{\substack{u_1 = S_{T_E \mid X}(t_e \mid x) \\ u_2 = S_{T_C \mid X}(t_c \mid x)}} . \tag{3.13}$$

We now have the tools required to derive the likelihood under dependence, as follows. As before, we begin with the general likelihood function from Equation 3.6. Without the assumption $T_E \perp T_C \mid X$, we apply the chain rule instead of factorizing:

$$= \prod_{i=1}^{N} \left(\left[f_{T_E \mid X}\left(T_{\text{obs}}^{(i)} \mid X^{(i)}\right) \int_{T_{\text{obs}}^{(i)}}^{\infty} f_{T_C \mid T_E, X}(t_c \mid T_{\text{obs}}^{(i)}, X^{(i)})\, dt_c \right]^{\delta_i} \right.$$
$$\left. \left[f_{T_C \mid X}\left(T_{\text{obs}}^{(i)} \mid X^{(i)}\right) \int_{T_{\text{obs}}^{(i)}}^{\infty} f_{T_E \mid T_C, X}(t_e \mid T_{\text{obs}}^{(i)}, X^{(i)})\, dt_e \right]^{1-\delta_i} \right) .$$

Under Sklar's Theorem (Survival), we can apply Lemma 1, yielding:

$$= \prod_{i=1}^{N} \left(\left[f_{T_E \mid X}\left(T_{\text{obs}}^{(i)} \mid X^{(i)}\right) \frac{\partial}{\partial u_1} \left(C(u_1, u_2)\Big|_{\substack{u_1 = S_{T_E \mid X}(T_{\text{obs}}^{(i)} \mid X^{(i)}) \\ u_2 = S_{T_C \mid X}(T_{\text{obs}}^{(i)}) \mid X^{(i)}}} \right) \right]^{\delta_i} \right.$$
$$\left. \left[f_{T_C \mid X}\left(T_{\text{obs}}^{(i)} \mid X^{(i)}\right) \frac{\partial}{\partial u_2} \left(C(u_1, u_2)\Big|_{\substack{u_1 = S_{T_E \mid X}(T_{\text{obs}}^{(i)} \mid X^{(i)}) \\ u_2 = S_{T_C \mid X}(T_{\text{obs}}^{(i)}) \mid X^{(i)}}} \right) \right]^{1-\delta_i} \right) . \tag{3.14}$$

For the sake of numerical stability, we instead optimize the *log-likelihood*:

$$\ell(\mathcal{D}) = \sum_{i=1}^{N} \left(\delta_i \left[\log f_{T_E \mid X}\left(T_{\text{obs}}^{(i)} \mid X^{(i)}\right) + \log \frac{\partial}{\partial u_1} \left(C(u_1, u_2)\Big|_{\substack{u_1 = S_{T_E \mid X}(T_{\text{obs}}^{(i)} \mid X^{(i)}) \\ u_2 = S_{T_C \mid X}(T_{\text{obs}}^{(i)}) \mid X^{(i)}}} \right) \right] + \right.$$
$$\left. (1 - \delta_i) \left[\log f_{T_C \mid X}\left(T_{\text{obs}}^{(i)} \mid X^{(i)}\right) + \log \frac{\partial}{\partial u_2} \left(C(u_1, u_2)\Big|_{\substack{u_1 = S_{T_E \mid X}(T_{\text{obs}}^{(i)} \mid X^{(i)}) \\ u_2 = S_{T_C \mid X}(T_{\text{obs}}^{(i)}) \mid X^{(i)}}} \right) \right] \right)$$

In this expression, the first term corresponds to the log-likelihood of observing the event at time $T_{\text{obs}}^{(i)}$. $C(u_1, u_2)$ is the joint survival function estimating the probability that both event and censoring occur after $T_{\text{obs}}^{(i)}$. By applying a partial derivative with respect to u_1 it becomes the conditional probability

that censoring happens after $T_{\text{obs}}^{(i)}$ given that event occurred at $T_{\text{obs}}^{(i)}$. Under independent copula, this term reduces to $S_{T_C|X}(T_{\text{obs}}^{(i)}|X^i)$. The third and fourth terms, by symmetry, represent the same quantities for the censorship time. Despite the visual complexity of Equation 3.15, the partial derivatives of the Clayton and Frank copulas admit closed form solutions, so the log-likelihood function has a closed form and can be maximized via gradient-based methods using software packages like PyTorch.

3.3 Optimization

We assume that we possess know the copula's form. However, optimizing with an assumed copula can prove challenging since the copula parameter may fail to converge to the optimal value that minimizes the objective function. To overcome this difficulty, we experimented with several optimization methods but ultimately decided to use the same learning rate for all model parameters, including the copula parameter. Furthermore, we discovered that multiplying the loss gradient with respect to the copula parameter and constraining it within an acceptable range can enhance the model's performance.

Algorithm 1: Learning Under Dependent Censorship

Input: $\mathcal{D}$: survival dataset of the form $\left\{\left(X^{(i)}, T^{(i)}, \delta^{(i)}\right)\right\}_{i=1}^{N}$; C_θ: a bivariate copula, parameterized by θ; $\mathcal{M}$, a class of survival model parameterized by ϕ that can produce $\hat{S}_{T|X}^{(\mathcal{M})}(t|X)$, $\hat{f}_{T|X}^{(\mathcal{M})}(t|X)$, for each $X^{(i)} \in \mathcal{D}$; α: learning rate, `NUM_EPOCHS`: number of iterations for optimization.

Result: $\hat{\theta}, \hat{\phi}_E, \hat{\phi}_C$: learned parameters of the copula and each marginal survival model.

$\mathcal{M}_E \leftarrow$ `Instantiate`$(\mathcal{M}; \hat{\psi}_E^{(0)})$;
$\mathcal{M}_C \leftarrow$ `Instantiate`$(\mathcal{M}; \hat{\psi}_C^{(0)})$;
$C_\theta \leftarrow$ `Instantiate`$(C; \hat{\theta}^{(0)})$;
for $i = 1, \ldots,$ *NUM_EPOCHS* **do**
  $\mathcal{L}_i \leftarrow \ell\left[\mathcal{D}; \hat{f}_{T|X}^{(\mathcal{M}_E)}, \hat{f}_{T|X}^{(\mathcal{M}_C)}, \hat{S}_{T|X}^{(\mathcal{M}_E)}, \hat{S}_{T|X}^{(\mathcal{M}_C)}, C_{\hat{\theta}^{(i)}}\right]$;
  $\hat{\psi}_C^{(i)} \leftarrow$`Adam`$(\mathcal{L}_{\texttt{EPOCH}}, \hat{\psi}_C, \alpha)$;
  $\hat{\psi}_E^{(i)} \leftarrow$`Adam`$(\mathcal{L}_{\texttt{EPOCH}}, \hat{\psi}_E, \alpha)$;
  $\hat{\theta}^{(i)}.gradient \leftarrow \hat{\theta}^{(i)}.gradient * 1000$;
  $\hat{\theta}^{(i)}.gradient \leftarrow clip(\hat{\theta}^{(i)}.gradient, -0.1, 0.1)$; `# clip limits the input into the range of` $[-0.1, 0.1]$
  $\hat{\theta}^{(i)} \leftarrow$`Adam`$(\mathcal{L}_{\texttt{EPOCH}}, \hat{\theta}, \alpha)$;
end
return $\hat{\theta}^{(i)}$, $\hat{\psi}_E^{(i)}$, $\hat{\psi}_C^{(i)}$

3.4 Evaluation

It is well-known in the literature that informative censoring has the potential to *bias* survival models [53]. However, such bias is often defined informally. In this section, we formalize the notion of bias in an estimated survival function and prove that the standard metrics of evaluation in survival analysis - concordance index [23], [25], [63] and Integrated Brier Score (IBS) [5] - may not always adequately reflect this bias. The significance of bias arises in scenarios that involve the computation of the difference between survival curves. For instance, consider the situation where we aim to determine the extent to which a surgical procedure or therapy can increase a patient's lifespan.

This discussion motivates our introduction and use of the *Survival-*ℓ_1 measure that compares the ground truth curve with the estimation of the curve; we subsequently demonstrate that the *Survival-*ℓ_1 method provides an accurate characterization of bias in estimating a survival function. We also introduce an algorithm to compute the Survival-Δ metric between a pair of survival curves (e.g. one estimated survival curve and one ground-truth survival curve). Finally, we introduce Log-likelihood as another metric to evaluate the performance of our method.

Before introducing evaluation metrics and discussing their shortcomings in identifying the distributional bias, we first provide an example to illustrate what we mean by distributional bias. Let's consider a survival scenario where all men are censored just before the event occurs. As a result, a model will learn a survival curve with a constant value of 1 for all time points. However, this curve does not represent the actual survival function that generated the data for males. The discrepancy between the true survival function and our estimate at each time point is what we refer to as distributional bias.

3.4.1 Concordance Index

Theorem 2. *The Concordance Index is not sensitive to the distributional bias in the survival function, meaning that there exist other survival functions different from the ground truth survival function that can result in the same*

Concordance Index.

We prove the theorem using an example in the rest of this section, but first, we need to have a few definitions. Earlier, we claimed that measuring the bias in survival function estimation provides a general means of measuring the bias in a survival model. To formalize this notion, we introduce the concept of *d-Survival-Δ bias*.

Definition 1 (d-Survival-Δ Bias)**.** *The d-Survival-Δ Bias is the distance between the ground truth and estimated survival curves at each time point T under some distance metric d (in our case, this is the ℓ_1 metric). Given some true survival curve S_T, and some estimated survival curve, $\hat{S}_T$, the d-Survival-Δ bias is:*

$$\Xi_{d\text{-}Survival\text{-}\Delta}(S, \hat{S}) \quad = \quad \int_0^\infty d(S(t), \hat{S}(t))\, dt \ . \tag{3.15}$$

Next, we will introduce Harrell's concordance index, and the concept of the *risk score*, on which the definition of the concordance index relies:

Definition 2 (Risk Score)**.** *A risk score, $\mathcal{R}_i \in \mathbb{R}$ is a real-valued number defined for each instance i. $\mathcal{R}_i$ has the property that a large value of $\mathcal{R}_i$ corresponds to a prediction of a small time-to-event, while a small value of $\mathcal{R}_i$ corresponds to a prediction of a large time-to-event. For the purposes of this proof, we define $\mathcal{R}_i$ as the cumulative hazard function.*

$$\mathcal{R}_i(t) \triangleq \Lambda(t|x_i) \tag{3.16}$$

Definition 3 (Concordance Index)**.** *The concordance index evaluates the relative rank-ordering of each instance's risk score with their observed ordering by time-to-event in the dataset. Given a dataset $\mathcal{D}$, and a corresponding risk score $\mathcal{R}_i$ for each instance, the concordance index is calculated as follows [25]:*

$$\Xi_{Conc.}(\mathcal{D}) = \frac{\sum_{i \neq j \in [1,N]\times[1,N]} \mathbf{1}\left[\mathcal{R}_i(t_i) > \mathcal{R}_j(t_i)\right] \mathbf{1}\left[T_{obs}^{(i)} < T_{obs}^{(j)}\right] \delta^{(i)}}{\sum_{i \neq j \in [1,N]\times[1,N]} \mathbf{1}\left[T_{obs}^{(i)} < T_{obs}^{(j)}\right]} \ . \tag{3.17}$$

Having introduced these terms, below we prove that biased estimates of the survival function may yield the same concordance index as the true survival

curves themselves, under the assumption of proportional hazards.

Consider the following toy example with two patients, 1 and 2. In this example, $\mathcal{D} = \left\{ \left(X^{(1)}, T^{(1)}_{\text{obs}}, \delta^{(1)} \right), \left(X^{(2)}, T^{(2)}_{\text{obs}}, \delta^{(2)} \right) \right\}$. Additionally, assume that 1 and 2 have the following ground-truth survival curves:

$$S^{(1)}_{T_E \,|\, X} = \begin{cases} -t + 1 & \text{if } t < 1 \\ 0 & \text{otherwise} \end{cases} \qquad S^{(1)}_{T_C \,|\, X} = 1 \ , \tag{3.18}$$

$$S^{(2)}_{T_E \,|\, X} = \begin{cases} -2t + 1 & \text{if } t < 0.5 \\ 0 & \text{otherwise} \end{cases} \qquad S^{(2)}_{T_C \,|\, X} = 1 \ . \tag{3.19}$$

Figure 3.1 plots these survival curves, along with their biased counterparts, which will be introduced later.

Moreover, assume that each patient experiences the event at the time corresponding to the survival rate of 0.5 under their respective survival curves. Since in this example, no patient is censored, we can therefore evaluate the inverse of $S^{(1)}_{T_E \,|\, X}, S^{(2)}_{T_E \,|\, X}$ at $T^{(1),\text{Quantile}} = T^{(2),\text{Quantile}} = 0.5$ to obtain:

$$\begin{cases} T^{(1)}_{\text{obs}} & \triangleq \mathbb{E}\left[S^{(1)}_{T_E \,|\, X} \right] = 0.5 \\ \\ T^{(2)}_{\text{obs}} & \triangleq \mathbb{E}\left[S^{(2)}_{T_E \,|\, X} \right] = 0.25 \end{cases} . \tag{3.20}$$

Concordance Under Unbiased Survival Curves. Here, we calculate the concordance index under the true survival curves specified in Equations 3.18 and 3.19. To do so, we compute $\mathcal{R}_1, \mathcal{R}_2$ at $t = 0.25$, to yield:

$$\begin{cases} \mathcal{R}_1 & = & \Lambda(0.25|x_1) & = & 0.125 \\ \mathcal{R}_2 & = & \Lambda(0.25|x_2) & = & 0.3 \end{cases} . \tag{3.21}$$

Thus, we can compute the concordance index as follows:

$$\begin{aligned} \Xi_{\text{Conc.}}(\mathcal{D}) =& \frac{\sum_{i \neq j \in [1,N] \times [1,N]} \mathbf{1}\left[\mathcal{R}_i(t_i) > \mathcal{R}_j(t_i) \right] \mathbf{1}\left[T^{(i)}_{\text{obs}} < T^{(j)}_{\text{obs}} \right] \delta^{(i)}}{\sum_{i \neq j \in [1,N] \times [1,N]} \mathbf{1}\left[T^{(i)}_{\text{obs}} < T^{(j)}_{\text{obs}} \right]} \\ =& \frac{1}{1} = 1 \end{aligned} . \tag{3.22}$$

We, therefore, see that a risk score computed from the ground-truth survival curves yields perfect concordance.

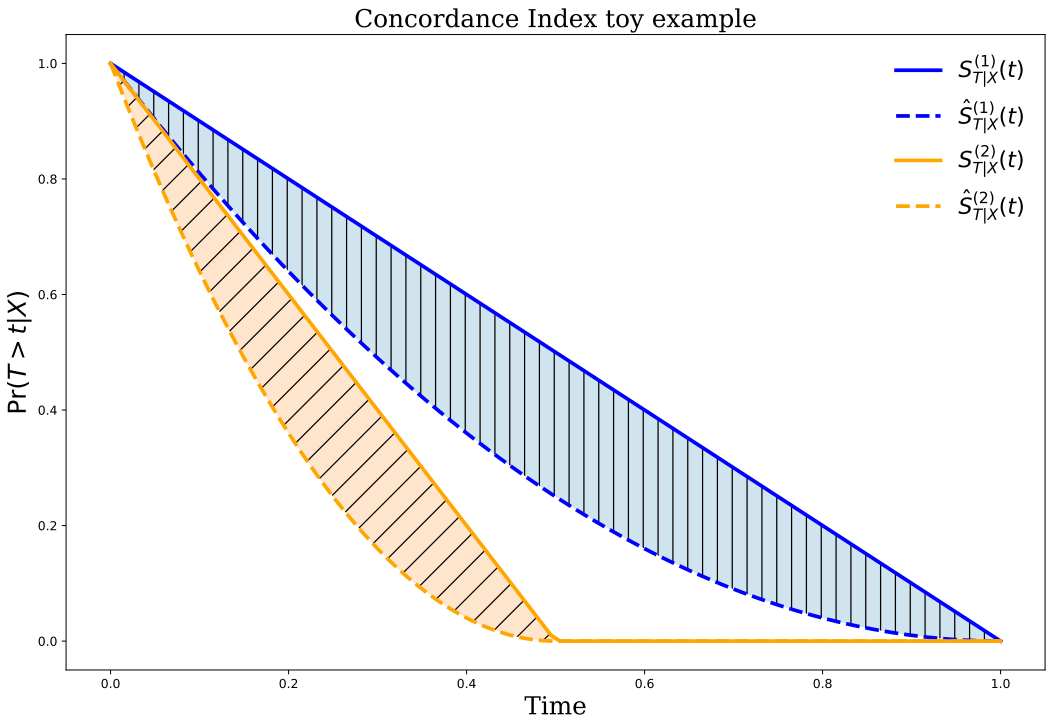

Figure 3.1: A figure representing the ground-truth survival functions for patients 1 and 2 in our toy example. The horizontal axis shows the progression of time, while the vertical axis shows the likelihood that the patient has not yet experienced the event of interest at the current time. The solid lines show $S^{(1)}_{T_E\,|\,X}(t)$, $S^{(2)}_{T_E\,|\,X}(t)$, while the dashed lines show $\hat{S}^{(1)}_{T_E\,|\,X}(t)$, $\hat{S}^{(2)}_{T_E\,|\,X}(t)$. The hatched area highlights the difference between each curve and its biased counterpart.

Concordance Under Biased Survival Curves. Next, we show that a biased set of survival curves can also yield perfect concordance. To do so, we will define the concept of a *biasing function*, apply it to the ground-truth survival curves in Equations 3.18 and 3.19, and then compute the corresponding concordance index.

Definition 4 (Biasing function)**.** *A biasing function, $B : \mathcal{S} \to \mathcal{S}$, is any function mapping the space of survival curves to itself.*

For the second part of this proof, define the biasing function, B, as follows:

$$S'_T \quad \triangleq B(S_T) = \quad S_T^2 \ . \tag{3.23}$$

Our definition of B yields the following biased survival curves. It is trivial to see from these definitions, and from Figure 3.1, for each $k \in \{1, 2\}$, that $\Xi_{\ell_1\text{-Survival-}\Delta}(S^{(k)}_{T_E}, S'^{(k)}_{T_E}) > 0$.

$$S'^{(1)}_{T_E \,|\, X} = \begin{cases} (-t+1)^2 & \text{if } t < 1 \\ 0 & \text{otherwise} \end{cases} \qquad S'^{(1)}_{T_C \,|\, X} = 1 \tag{3.24}$$

$$S'^{(2)}_{T_E \,|\, X} = \begin{cases} (-2t+1)^2 & \text{if } t < 0.5 \\ 0 & \text{otherwise} \end{cases} \qquad S'^{(2)}_{T_C \,|\, X} = 1 \tag{3.25}$$

Next, we calculate the concordance index under the biased survival functions above. As before, we use the function $\Lambda(t|x_i)$ to obtain risk scores from the biased survival functions:

$$\begin{cases} \mathcal{R}'_1 & = \Lambda(0.25|x_1) \approx 0.25 \\ \mathcal{R}'_2 & = \Lambda(0.25|x_2) \approx 0.6 \end{cases} \tag{3.26}$$

As before, we can compute the concordance index:

$$\begin{aligned} \Xi_{\text{Conc.}}(\mathcal{D}) =& \frac{\sum_{i \neq j \in [1,N]\times[1,N]} \mathbf{1}\left[\mathcal{R}'_i(t_i) > \mathcal{R}'_j(t_i)\right] \mathbf{1}\left[T^{(i)}_{\text{obs}} < T^{(j)}_{\text{obs}}\right] \delta^{(i)}}{\sum_{i \neq j \in [1,N]\times[1,N]} \mathbf{1}\left[T^{(i)}_{\text{obs}} < T^{(j)}_{\text{obs}}\right]} \\ =& \frac{1}{1} = 1 \end{aligned} \ . \tag{3.27}$$

Since the concordance index is the same under both the biased and unbiased survival functions, this proves that bias in the underlying estimation of a survival function may not manifest itself in the concordance index.

3.4.2 Brier Score and Integrated Brier Score

Now we will explain how the bias in the survival curve can not be identified using IBS.

Theorem 3. *The Integrated Brier Score (IBS) is not sensitive to the distributional bias in the survival function, meaning that there exist other survival functions different from the ground-truth survival function that can result in the same IBS.*

We need to first define Brier Score(BS) [5], based on which we can define Integrated Brier Score(IBS).

Definition 5 (Brier Score)**.** *The Brier Score evaluates the accuracy of the survival curve at each time t, representing the average squared distance between the predicted survival probability and the survival status at time t. Given a dataset $\mathcal{D} = \left\{(X^{(i)}, T^{(i)}_{obs}, \delta^{(i)})\right\}_{i=1}^{N}$, and the predicted survival curve $\hat{S}(t, x)$, the Brier score is:*

$$BS(t, D, \hat{S}) = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{1}\left(t_i > t\right) - \hat{S}\left(t, x_i\right)\right)^2 . \tag{3.28}$$

BS is always a number between 0 and 1, where 0 means the best possible performance and 1 means the worst. BS defined in Equation 3.28 is valid for a scenario without censoring. In the presence of right censoring, the score must be adjusted using inverse probability of censoring weights. The adjusted score is defined as below [24]:

$$BS(t, D, \hat{S}, \hat{G}) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\left(0 - \hat{S}\left(t, x_i\right)\right)^2 \cdot \mathbf{1}\left(t_i \le t\right) \cdot \delta_i}{\hat{G}\left(t_i^-\right)} + \frac{\left(1 - \hat{S}\left(t, x_i\right)\right)^2 \cdot \mathbf{1}\left(t_i > t\right)}{\hat{G}\left(t\right)}\right), \tag{3.29}$$

where t_i is the time of the event or censoring for sample i and $\hat{G}(t) = P(C > t)$ is the probability of censoring after time t calculated using a Kaplan-Meier algorithm with censoring bit reversed(consider the censored patient as patients who experienced an event and patients who experienced an event as censored).

Definition 6 (Integrated Brier Score (IBS))**.** *The Integrated Brier Score (IBS) [24] provides a comparison of a model to the perfect model, which is a step function that drops from* 1 *to* 0 *at the time of the event.*

$$IBS(t_{max}, D, \hat{S}, \hat{G}) = \int_0^{t_{max}} BS(t, D, \hat{S}, \hat{G}) dt \tag{3.30}$$

Consider the following toy example with one patient $\mathcal{D} = \left\{ \left(X^{(1)}, T^{(1)}_{\text{obs}}, \delta^{(1)} \right) \right\}$ and assume that the patient has the following ground-truth survival curve:

$$S^{(1)}_{T_E \,|\, X} = \begin{cases} 1 - 0.5t^2 & t \leq 1 \\ 1 - 0.5t & 1 < t \leq 2 \\ 0 & \text{otherwise} \end{cases} \qquad S^{(1)}_{T_C \,|\, X} = 1 \ . \tag{3.31}$$

Assume that the event happened at $t = 1$, therefore, $T^{(1)}_{obs} = 1$. Then consider the following survival curve as our estimation:

$$\hat{S}^{(1)}_{T_E \,|\, X} = \begin{cases} 1 - 0.5t & t \leq 1 \\ 0.5(t-2)^2 & 1 < t \leq 2 \\ 0 & \text{otherwise} \end{cases} \qquad \hat{S}^{(1)}_{T_C \,|\, X} = 1 \ . \tag{3.32}$$

Figure 3.2 plots biased and unbiased survival curves for the example alongside the step Function we use to calculate BS and IBS. Both S and $\hat{S}$ have the same BS (at event time which is $t = 1$) and IBS. Since the BS and IBS are the same under both the biased and unbiased survival functions, this proves that bias in the underlying estimation of a survival function may not manifest itself in BS and IBS.

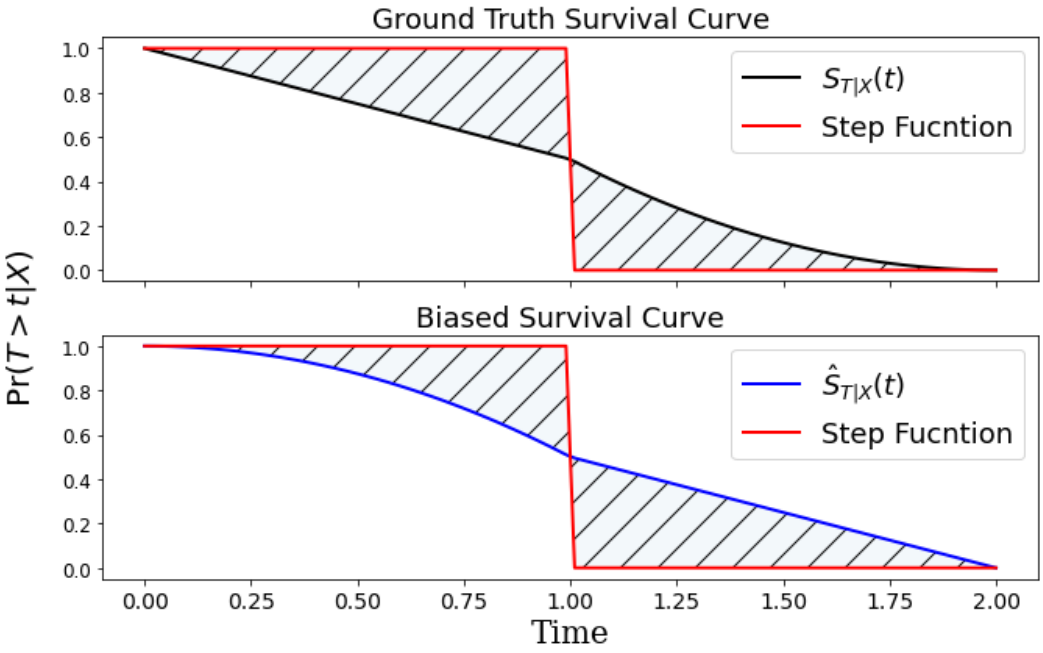

Figure 3.2: A figure representing the ground-truth survival function, $S_{T|X}(t)$, for a patient. The horizontal axis shows the progression of time, while the vertical axis shows the likelihood that the patient has not yet experienced the event of interest at the current time. $\hat{S}_{T|X}(t)$ shows a biased survival curve and Step Function shows the perfect survival curve in terms of the IBS metric. The hatched area in each subplot shows the region used for calculating IBS for $S_{T|X}(t)$, $\hat{S}_{T|X}(t)$. This visualization showcases how the BS and IBS are the same under both $S_{T|X}(t)$, $\hat{S}_{T|X}(t)$.

3.4.3 The Survival-ℓ_1 Metric

Due to the drawbacks of conventional scoring rules, we introduce the *Survival-ℓ_1* measure as a means of quantifying bias in survival analysis due to dependent censoring. The *Survival-ℓ_1* metric $\mathcal{C}_{\text{Survival-}\ell_1} : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$, is the ℓ_1 distance between the ground-truth survival curve, $S_{T|X}$, and the estimate achieved by a survival model, $\hat{S}_{T|X}$ (Figure 3.3), over the entire range of the curves. Below we prove that the *Survival-ℓ_1* score is proper under dependent censoring.

However, the scale of the naive ℓ_1 measure between survival curves is proportional to the total amount of time taken by each survival curve. To ensure that survival curves over a longer range do not contribute proportionally more to the evaluation metric than those over a shorter range, we define the small constant *normalizing quantile*, $Q_{\|\cdot\|}$ (in our experiments, $Q_{\|\cdot\|} = 0.01$). We can loosely think of the time when each survival curve reaches the normalizing quantile as the "end time" of that survival curve. By normalizing the area between the survival curves by the *temporal normalization* value $T^{(i)}_{\max} = S^{-1}_{T|X^{(i)}}\left(Q_{\|\cdot\|}\right)$, we ensure that the duration spanned by a patient's survival curve does not influence that patient's contribution to $\mathcal{C}_{\text{Survival-}\ell_1}$ relative to other patients.

Then, our *Survival-ℓ_1* metric takes the following form:

$$\mathcal{C}_{\text{Survival-}\ell_1}(S_{T|X}, \hat{S}_{T|X}) = \sum_{i=1}^{N} \frac{1}{N \times T^{(i)}_{\max}} \int_0^{\infty} \left| S_{T\,|\,X}(t \,|\, X^{(i)}) - \hat{S}_{T\,|\,X}(t \,|\, X^{(i)}) \right| dt \ . \tag{3.33}$$

Theorem 4. *The Survival-ℓ_1 metric is sensitive to the distributional bias in the survival function, meaning that there is no function different from the ground truth survival function that can result in the same score under this metric.*

First, we prove the following theorem.

Theorem 5. *For any $S_i, S_j \in \mathcal{S}$, if $S_i \neq S_j$, then $\mathcal{C}_{Survival\text{-}\ell_1}(S_i, S_j) > 0$.*

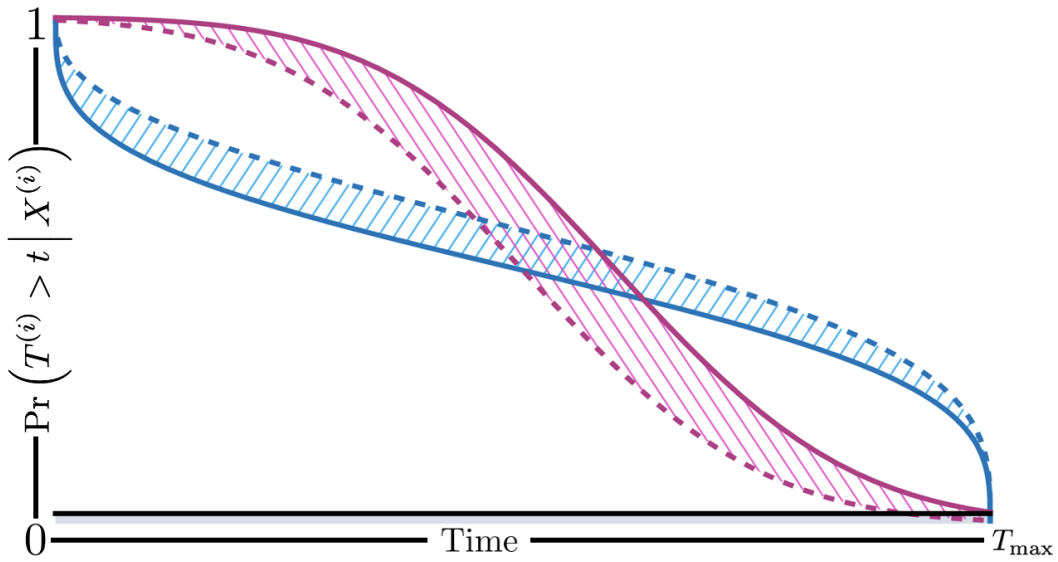

Figure 3.3: The *Survival-ℓ_1* metric, $\mathcal{C}_{\text{Survival-}\ell_1}(S, \hat{S})$, for event and censoring distributions. Dashed lines represent the predicted survival curves, $\hat{S}_{T_E \mid X}$, and $\hat{S}_{T_C \mid X}$, while solid lines represent the corresponding ground-truth survival curves, $S_{T_E \mid X}$, and $S_{T_C \mid X}$. The black horizontal line represents the normalizing quantile, $Q_{\|\cdot\|}$. The area of the hatched blue region above $Q_{\|\cdot\|}$ is the value of $\mathcal{C}_{\text{Survival-}\ell_1}(S_{T_E \mid X}, \hat{S}_{T_E \mid X})$, while that of the hatched pink region is the value of $\mathcal{C}_{\text{Survival-}\ell_1}(S_{T_C \mid X}, \hat{S}_{T_C \mid X})$.

Proof. If $S_i \neq S_j$, then this means that there is some time $t \in [0, \infty)$, and some $\epsilon \in \mathbb{R} \backslash \{0\}$ for which:

$$S_i(t) = S_j(t) + \epsilon \ . \tag{3.34}$$

Therefore, at this time, the inner term of the integral in $\mathcal{C}_{\text{Survival-}\ell_1}$ will be equal to $|\epsilon|$. Then, because N and $T^{(i)}_{\max}$ are non-negative, and because $\left|S^{(i)}_T(t) - \hat{S}^{(i)}_T(t)\right|$ is non-negative, we know that all other terms summed into the computation of $\mathcal{C}_{\text{Survival-}\ell_1}(S_i, S_j)$ must be at least zero. Therefore, $\mathcal{C}_{\text{Survival-}\ell_1}(S_i, S_j) \geq |\epsilon|$, which, since $\epsilon \neq 0$, implies that $\mathcal{C}_{\text{Survival-}\ell_1}(S_i, S_j) > 0$. □

Now Considering $\mathcal{C}_{\text{Survival-}\ell_1}(S_i, S_j) = 0$ if $S_i = S_j$, it is obvious that the Survival-ℓ_1 metric is sensitive to the distributional bias in the survival function. This means that for any survival curve other than the ground truth curve the value of the metric is greater than zero, which means Survival-ℓ_1 is a proper score.

Computing the Survival-ℓ_1 Metric Algorithm

Here, we expand on the computation of the Survival-ℓ_1 metric by providing an algorithm for the explicit computation of the inner term of the Survival-ℓ_1 metric, stated in Equation 3.33, as well as the value $T_{\max}$ for the given pair of survival curves, $S, \hat{S}$:

Algorithm 2: Discrete Approximation of the Inner Term of the Survival-ℓ_1

Input:

1. S_1, S_2: Survival curves to compare under the Survival-ℓ_1 metric. Here, we assume S_1 is the ground-truth survival curve, and S_2 is the estimated curve.
2. $Q_{\|\cdot\|}$: Normalizing quantile.
3. N_{steps}: Number of discretization steps.

Result:

1. Δ_{total}: a discretized approximation of the integral $\int_0^\infty \left| S_1(t \mid X^{(i)}) - \hat{S}_2(t \mid X^{(i)}) \right| dt$
2. $T_{\max}$: This is used as a normalization weight when computing the full expression for the Survival-ℓ_1 metric.

$T_{\max} \leftarrow S_1^{-1}\left(Q_{\|\cdot\|}\right)$;
$\Delta_{\text{total}} \leftarrow 0$
for $i = 1, \ldots, N_{steps}$ **do**
  $\Delta_{i;S_1,S_2} \leftarrow \frac{T_{\max}}{N_{\text{steps}}} \times \ell_1 \left[S_1\left(\frac{i \times T_{\max}}{N_{\text{steps}}}\right), S_2\left(\frac{i \times T_{\max}}{N_{\text{steps}}}\right) \right]$;
  $\Delta_{\text{total}} \leftarrow \Delta_{\text{total}} + \Delta_{i;S_1,S_2}$;
end
return Δ_{total}, T_{max}

3.4.4 Log-likelihood

The properness of log-likelihood as a scoring rule has been demonstrated for independent censoring [53]. However, its suitability for dependent censoring is contingent upon the copula's identifiability, whereby identifiability means there exists a unique copula that can fit the data. We have not been able to

prove that log-likelihood is a strictly proper score for dependent censoring, but our experiments have shown that it can be a trustable metric. In this case, we presume the copula is identifiable and present our approach's log-likelihood-based performance evaluation.

Chapter 4

Experiments and Results

In this chapter, we show the results of our experiments, highlighting how our approach improves the learning of patient-specific survival curves. Furthermore, we demonstrate the proficiency of our method in estimating Kendall's tau (τ), which is a critical metric used for evaluating the dependency intensity within the dataset.

The *Survival*-ℓ_1 metric places strong assumptions on our knowledge of the data-generating process by assuming access to the ground-truth survival functions for each outcome. For this reason, we predominantly make use of synthetic data to evaluate the merits of our approach. We also compare the performance of our methods based on log-likelihood.

Furthermore, we demonstrated that our proposed method is not restricted to a singular type of copula, and instead can proficiently handle a convex combination of Archimedean copula members. This ability to relax the assumption of prior knowledge regarding the copula type highlights the versatility and adaptability of our methodology.

4.1 Synthetic Data Experiments

Initially, we present an algorithm to generate synthetic data while adhering to a prescribed copula C with Weibull CoxPH margins. Then we will explain the results that we present for each set of experiments.

4.1.1 Synthetic Data Generating Algorithm

Algorithm 3 has the capability to generate data under risk functions that can be either linear or non-linear, depending on what function is provided in the input.

Algorithm 3: Generating Synthetic Dependent Survival Data

Input: $X \in \mathbb{R}^{N \times d}$: a set of covariates, $g_\psi : \mathcal{X} \to \mathbb{R}$: a class of risk function parameterized by ψ, C_θ: a class of copula parameterized by θ to impose upon the data, $(\nu_E^*, \rho_E^*, \psi_E^*), (\nu_C^*, \rho_C^*, \psi_C^*), \theta^*$: data-generating parameters associated with each outcome model and the copula, respectively.

Result: $\mathcal{D}$, a survival dataset with the desired dependence.

$\mathcal{D} = \emptyset$;
for $i = 1, \ldots, N$ **do**
  $u_1^{(i)}, u_2^{(i)} \sim C_{\theta^*}$;
  $T_E^{(i)} \leftarrow \left(\frac{-\log(u_1)}{g_{\psi_E^*}(X^{(i)})} \right)^{\frac{1}{\nu_E^*}} \rho_E^*$;
  $T_C^{(i)} \leftarrow \left(\frac{-\log(u_2)}{g_{\psi_C^*}(X^{(i)})} \right)^{\frac{1}{\nu_C^*}} \rho_C^*$;
  $\mathcal{D} \leftarrow \mathcal{D} \cup \left\{ \left(X^{(i)}, \min\left(T_E^{(i)}, T_C^{(i)}\right), \mathbb{1}\left[T_E^{(i)} < T_C^{(i)}\right] \right) \right\}$;
end
return $\mathcal{D}$

4.1.2 Results Explanation

We report the performance of methods for each experimental set using the *Survival-*ℓ_1 metric and Negative Log-likelihood, across various values of Kendall's τ. Additionally, we evaluate Kendall's τ estimated by our method against the true Kendall's τ employed during data generation. Lastly, we include a graph displaying the proportion of samples that have encountered the event (uncensored). We used Kendall's τ as a measure of dependence as it allows for the comparison of copulas with the same level of dependency, whereas the same Copula's parameter does not necessarily indicate the same level of dependency for different Copulas.

- *Survival-*ℓ_1 plots: These plots show the $\mathcal{C}_{\text{Survival-}\ell_1}$ bias of the model, as a function of the dependence (true Kendall's τ), for both independence-assuming and copula-based models on synthetic data. Going from left to right on the x-axis denotes stronger dependence between the survival and event time in the data-generating process. The y-axis is overloaded; the

scales on the left-hand side of each y-axis correspond to bias incurred in the prediction of the event times and the scales on the right-hand side correspond to bias incurred in the prediction of the censoring times. Dotted lines represent the bias in the event and censoring survival curves incurred by independence-assuming models, while solid lines represent the bias incurred by our copula-based approach.

- Estimated vs True τ plots: For each value of τ, we plot the estimated copula value $\hat{\tau}$ as a function of the dependence, τ^*. The dotted line, representing $\hat{\tau} = \tau^*$, is plotted for reference. Points close to the line indicate that the learned dependence parameter was close to that of the data-generating process.

- Negative Log-likelihood plots: These plots compare the negative log-likelihood of our method, referred to as the Dependent Model, the baseline method, labeled as the Independent Model, and the data generating model, DGP[1]. The Negative log-likelihood values obtained from each model are plotted as a function of τ, where a lower metric indicates superior performance. We believe that the DGP has the lowest possible negative log-likelihood and can be considered as a lower bound for other methods.

- Event Percentage: Lastly, we display a plot presenting the mean of the indicator beat, δ, across the dataset, demonstrating the percentage of samples that encounter the event, as a function of Kendall's τ. This plot illustrates the impact of dependence strength on the distribution of samples that experience event or censorship.

4.2 Linear Risk Experiments

For the `Linear-Risk` experiments, we generate data according to Algorithm 3 with $X \in \mathbb{R}^{N\times 10} \sim \mathcal{U}_{[0,1]}$, $\nu_E^* = 4, \rho_E^* = 14, \psi_E^*(X) = \beta_E^T(X)$, $\nu_C^* = 3, \rho_C^* = 16, \psi_C^*(X) = \beta_C^T(X)$, where $\beta_E, \beta_C \in [0,1]^{10} \sim \mathcal{U}_{[0,1]}$. Experiments

[1]DGP: Data generating process

were performed on $20,000$ train, $10,000$ validation, and $10,000$ test samples. Our method was subjected to testing with five varying levels of dependency ($\tau = [0.01, 0.2, 0.4, 0.6, 0.8]$) for both Clayton and Frank copula.

Results for this set of experiments are presented in Figure 4.1. As shown, our approach effectively reduces the Survival-ℓ_1 bias, while the Survival-ℓ_1 bias under the assumption of independence increases as the strength of dependency (Kendall's τ) grows. In comparison with the model trained under the independence assumption, our method demonstrates superior performance in terms of negative log-likelihood. Furthermore, our approach successfully retrieves Kendall's τ used in data generation.

4.3 Nonlinear Risk Experiments

For the `NonLinear-Risk` experiments, we generate data according Algorithm 3 with $X \in \mathbb{R}^{N\times 10} \sim \mathcal{U}_{[0,1]}$, $\nu_E^* = 4, \rho_E^* = 17, \psi_E^*(X) = \sum_{i=1}^{10} X_i^2$, $\nu_C^* = 3, \rho_C^* = 16, \psi_C^*(X) = \sum_{i=1}^{10} \beta_{Ci} X_i^2$, where $\beta_C \in [0,1]^{10} \sim \mathcal{U}_{[0,1]}^{10}$. Experiments were performed on $20,000$ train, $10,000$ validation, and $10,000$ test samples.
For the `NonLinear-Risk` experiments, we employed a neural network with hidden layers composed of $[10, 4, 4, 4, 2]$ nodes and an ELu activation function. Additionally, we included an $l2$ regularizer with a weight of $1e^{-3}$ during the training process. Similar to `Linear-Risk` experiments, we tested our method on five different levels of dependency for both Clayton and Frank copula.

Figure 4.2 demonstrates that our method can handle arbitrary risk functions, for which a closed-form solution is not available, using neural networks. Our approach outperformed the model trained under the independence assumption in both Survival-ℓ_1 bias and negative log-likelihood metric. Additionally, we accurately estimated Kendall's τ used in data generation.

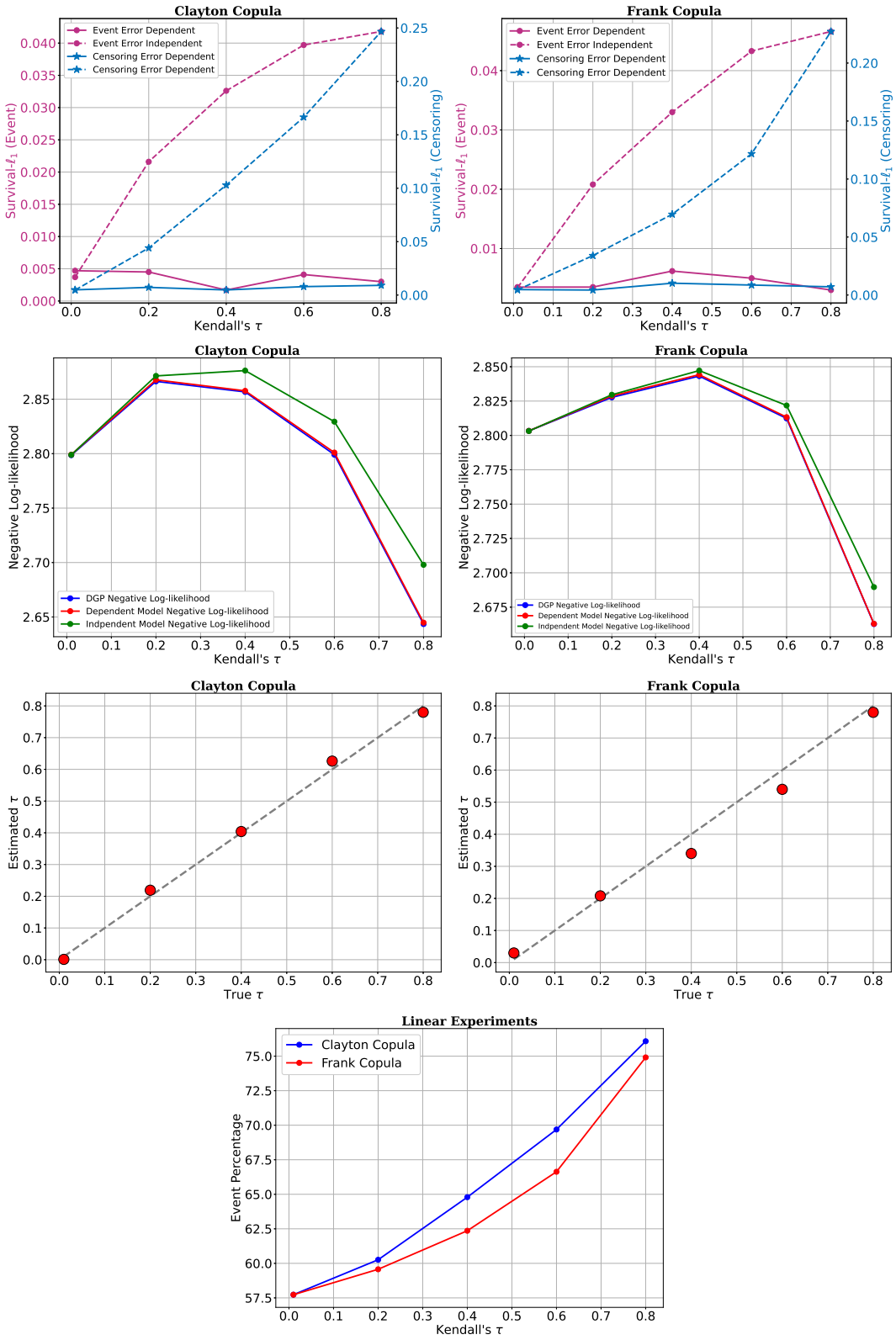

Figure 4.1: Linear experiments plots. The first row of plots exhibits $\mathcal{C}_{\text{Survival-}\ell_1}$ bias. The second and third rows display negative log-likelihood and true versus estimated Kendall's τ. The final row's plot depicts the percentage of events.

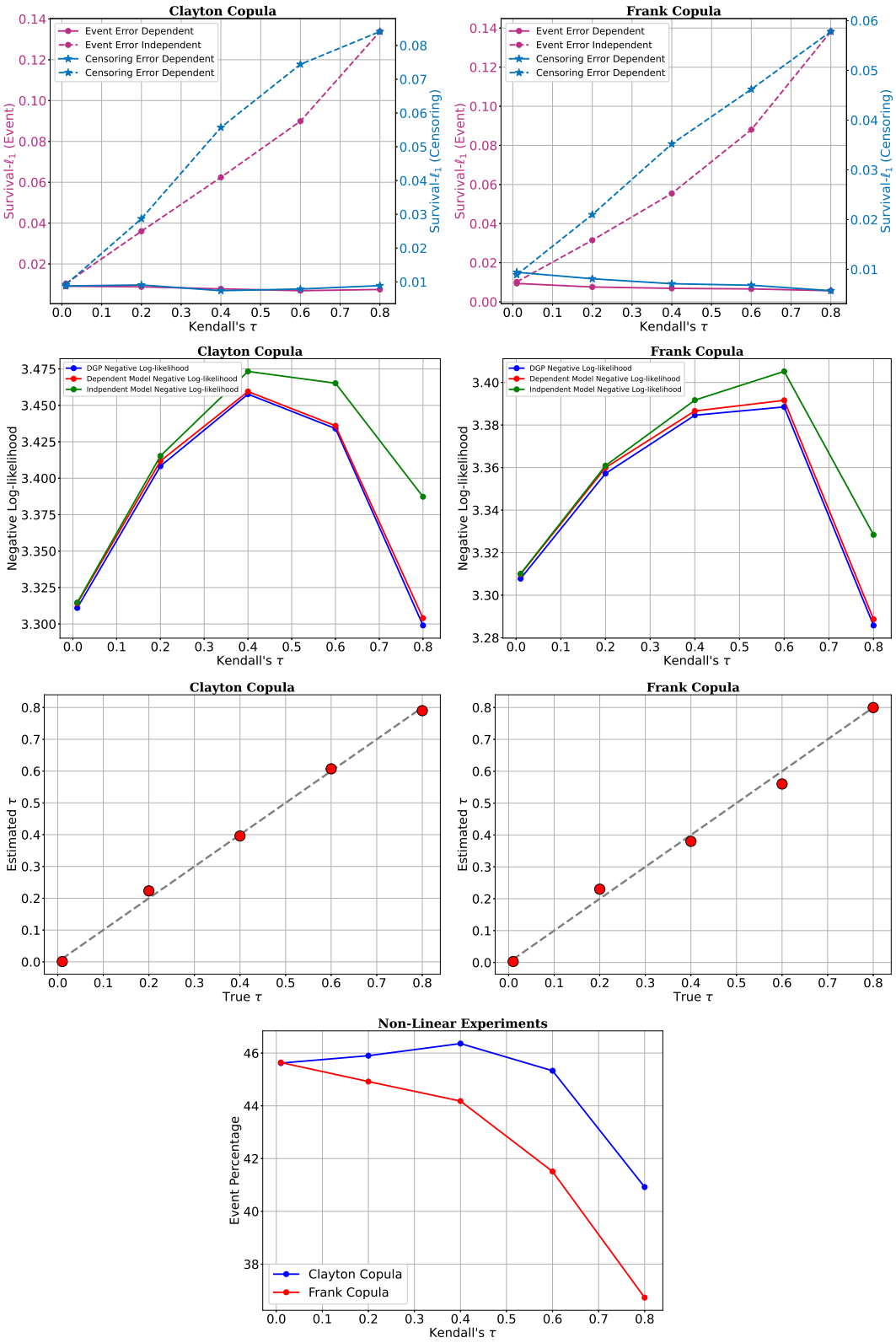

Figure 4.2: Non-Linear experiments plots. The first row of plots exhibits $\mathcal{C}_{\text{Survival-}\ell_1}$ bias. The second and third rows display negative log-likelihood and true versus estimated Kendall's τ. The final row's plot depicts the percentage of events.

4.4 Convex copula with Linear Risk

Requiring prior knowledge of the copula type can be a limiting assumption. To address this, we proposed utilizing a convex combination of the Frank and Clayton copulas, which enables us to handle scenarios where the copula family is unknown. We replicated the experimental setup from the `Linear-Risk` experiments, except for using a combination of a Clayton and a Frank copula, $C_{\theta_1,\theta_2}(u,v) = 0.5 * \mathit{Clayton}_{\theta_1}(u,v) + 0.5 * \mathit{Frank}_{\theta_2}(u,v)$, instead of using a Frank or Clayton copula to generate the data. Here we used a simple combination of Frank and Clayton copula with the same Kendall's τ for simplicity. But our method can handle any number of copulas with different Kendall's τ.

The results of experiments with a convex copula, as shown in Figure 4.3, demonstrate that our method effectively reduces the Survival-ℓ_1 bias in survival curves while accurately retrieving the strength level of the dependency used in data generation. Moreover, our method achieves a negative log-likelihood score that is similar to the ground truth model's score, regardless of the dependence strength, whereas the difference between the model trained under the independence assumption and the ground truth model's score increases in this metric as the dependence becomes stronger.

4.5 Semi-Synthetic Experiments

In addition to conducting fully synthetic experiments, we also attempted a `semi-synthetic` approach where we incorporated covariates from a real-world dataset and substituted the outcomes with a synthetic function. To do so, we utilized the **SUPPORT** dataset [37] renowned in survival analysis, which was curated for creating a predictive model for the survival of hospitalized adults. This dataset comprises 8,873 patients, and we were able to access 11 covariates for each patient.

For this set of experiments we used generated data according to Algorithm 3 with $X \in \mathbb{R}^{N*11}$,where X are from **SUPPORT** dataset, $\nu_E^* = 4, \rho_E^* = 17, \psi_E^*(X) = \log(\sum_{i=1}^{10} \beta_{Ei} X_i)$, $\nu_C^* = 3, \rho_C^* = 16, \psi_C^*(X) = \log(\sum_{i=1}^{10} \beta_{Ci} X_i)$,

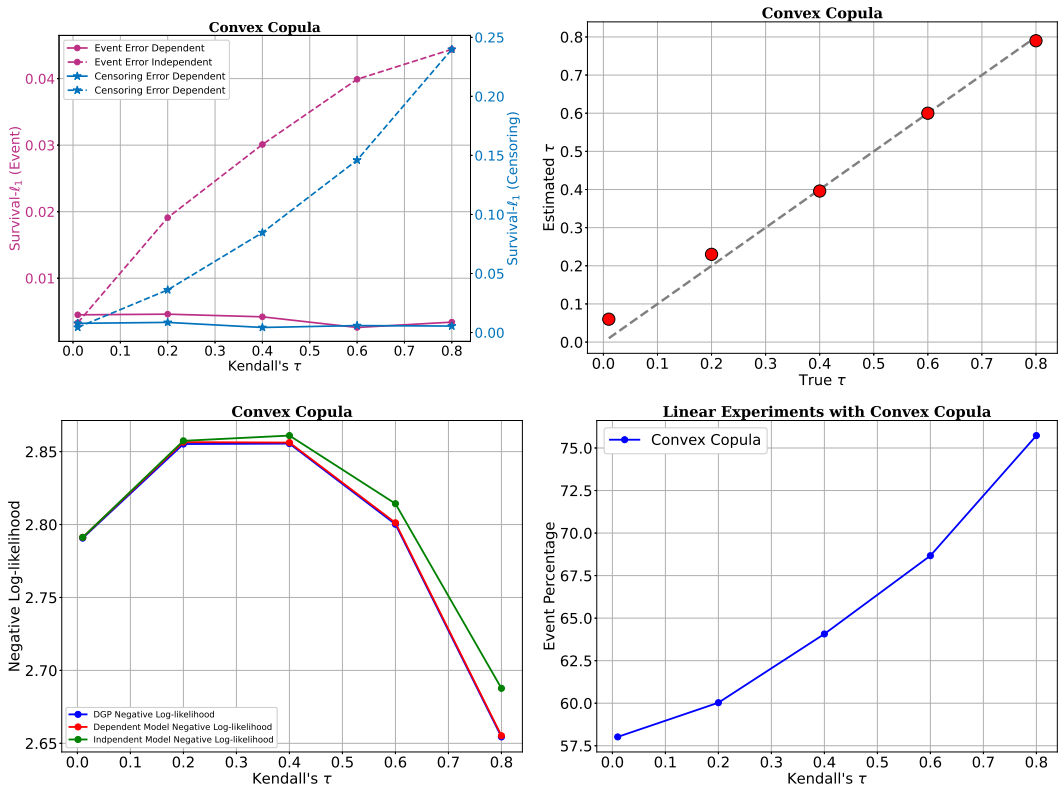

Figure 4.3: Plots for linear experiments with convex copulas. The plots on the first row consist of the left and right plots, displaying $\mathcal{C}_{\text{Survival-}\ell_1}$ bias and true versus estimated Kendall's τ plot, respectively. The left plot on the second row displays the negative log-likelihood plot, and the last plot shows the event percentage for each experiment.

where $\beta_E, \beta_C \in [0,1]^{10} \sim \mathcal{U}^{10}_{[0,1]}$. We split the dataset into 60% for training, 20% for validation, and the last 20% for the test set.

For the `semi-synthetic` experiments, we employed a neural network with hidden layers composed of $[10, 8, 4]$ and an ELu activation function. Additionally, we included an $l2$ regularizer with a weight of $1e^{-3}$ during the training process. Similar to other experiments, we tested our method on five different levels of dependency for both Clayton and Frank copula.

Figure 4.4 shows the outcomes of the experiments, using both Frank and Clayton Copulas, demonstrated identical performance to that of the fully synthetic experiments. Furthermore, our approach has demonstrated the ability to generate satisfactory results even in smaller datasets with a high percentage of censoring.

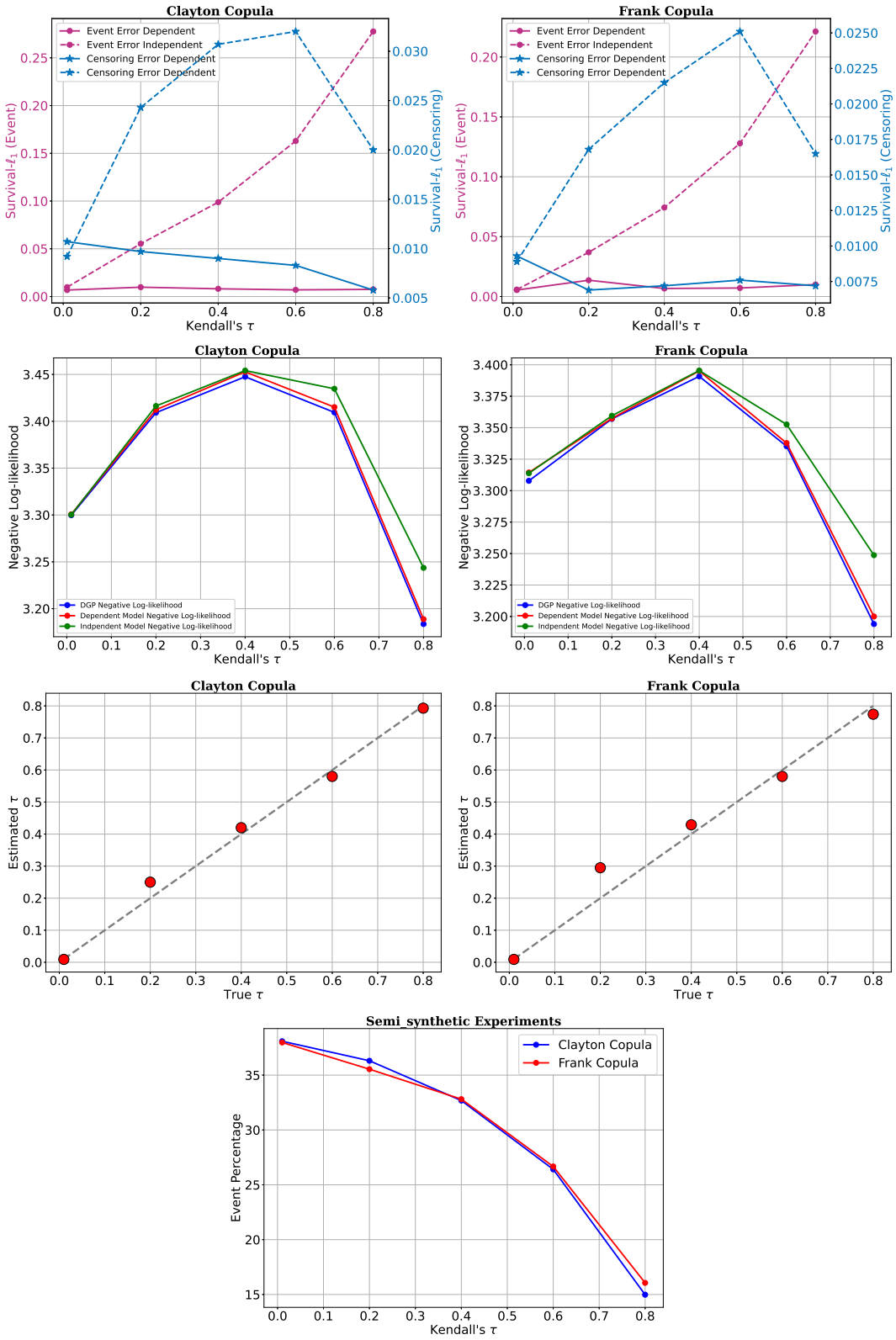

Figure 4.4: Semi-Synthetic experiments plots. The first row of plots exhibits $\mathcal{C}_{\text{Survival-}\ell_1}$ bias. The second and third rows display negative log-likelihood and true versus estimated Kendall's τ. The final row's plot depicts the percentage of events.

4.6 Survival Analysis Metrics

While we demonstrated that the Concordance Index (CI) and Integrated Brier Score (IBS) are insufficient for fully capturing the Survival-ℓ_1 bias, these metrics are widely accepted within the survival community. As such, we include these metrics in our experimental results. Our findings indicate that our approach does not compromise either of these two metrics in order to achieve a lower Survival-ℓ_1 bias.

4.6.1 Concordance Index

We start with the Concordance Index, which is a well-known metric in survival analysis. We present the performance of the ground truth model, labeled as DGP, our model, labeled as Dependent, and the baseline model trained under the independence assumption, labeled as Independent. Since we work with synthetic data, we have access to the ground truth time for both events and censoring. Therefore, we report the C-index for two scenarios: in the first scenario, we do not consider censoring and assume that we have access to both event and censoring time. In the second scenario, we include censoring and only observe the minimum of the event and censoring time. We report the C-index for 5 different levels of dependency (Kendall's $\tau = [0.01, 0.2, 0.4, 0.6, 0.8]$). For each level of dependency, we provide six bars where the three bars on the left side represent the C-index based on the no-censoring scenario, and the three bars on the right represent the C-index based on the observational dataset.

We expect that the DGP and the model trained using a copula exhibit better performance compared to the model trained based on the independence assumption. However, it is noted that the C-index values are quite similar in most cases, indicating that the C-index is not effective in detecting any potential bias in the survival curve.

It can be observed that the C-index derived from the observational dataset differs significantly from the C-index calculated based on the no-censoring dataset. Furthermore, this difference tends to increase with stronger depen-

dence. Additionally, the C-index based on the observational dataset does not align with the ordering of methods provided by the C-index based on the no-censoring dataset.

Linear Risk Experiments

Clayton Copula

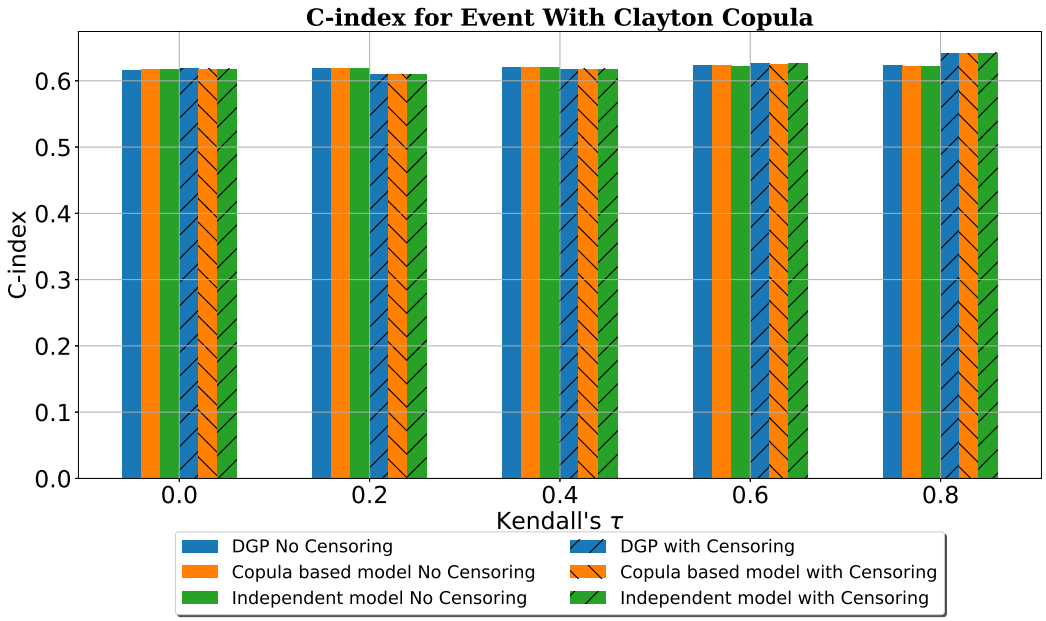

Figure 4.5: C-Index for Event in Linear Risk Experiments with Clayton Copula

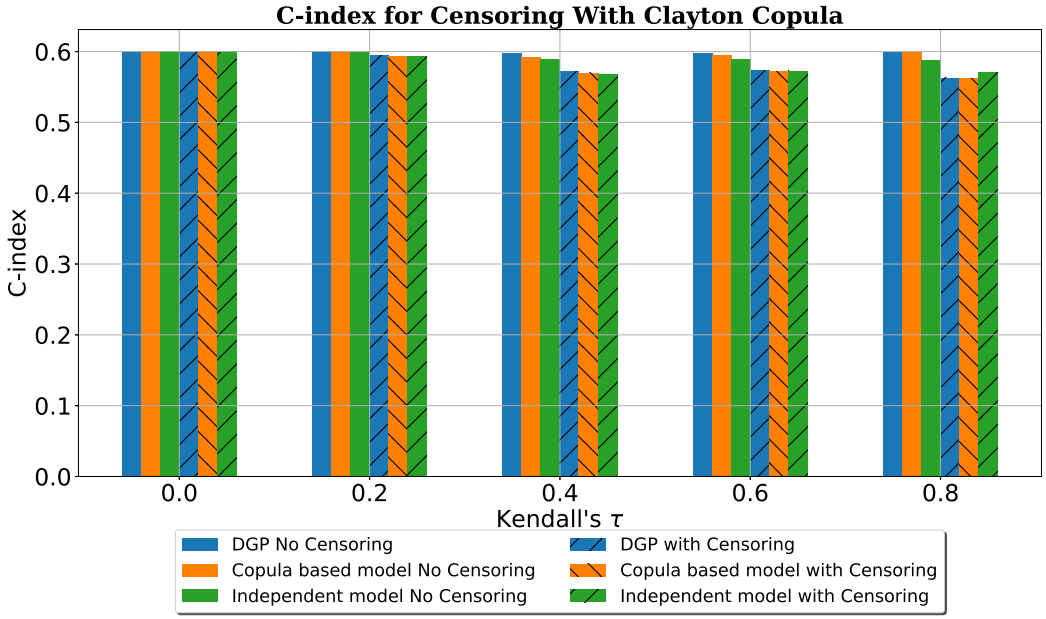

Figure 4.6: C-Index for Censoring in Linear Risk Experiments with Clayton Copula

Frank Copula

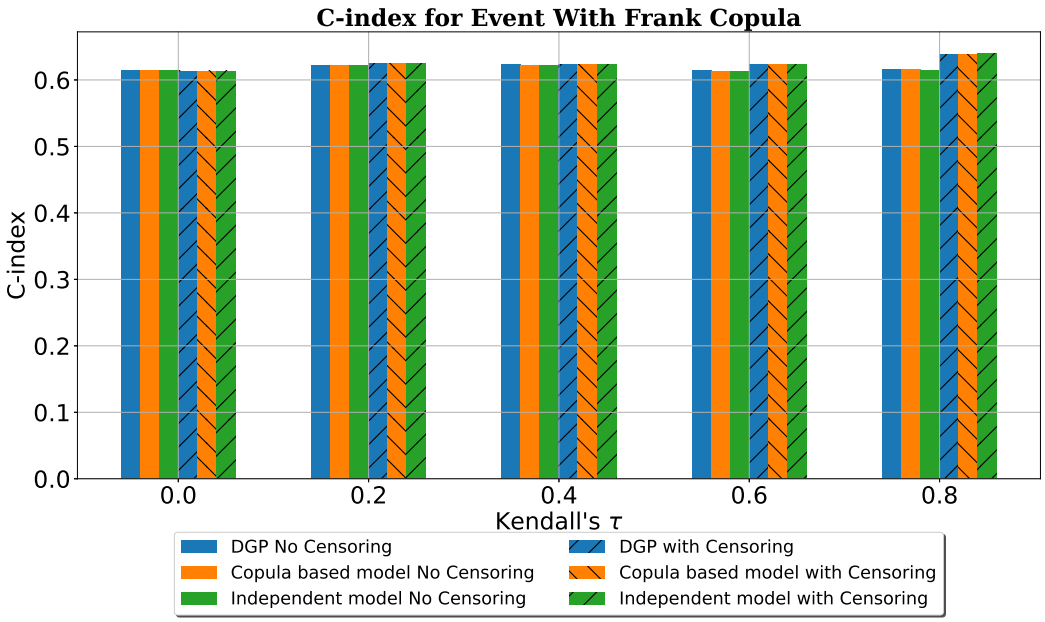

Figure 4.7: C-Index for Event in Linear Risk Experiments with Frank Copula

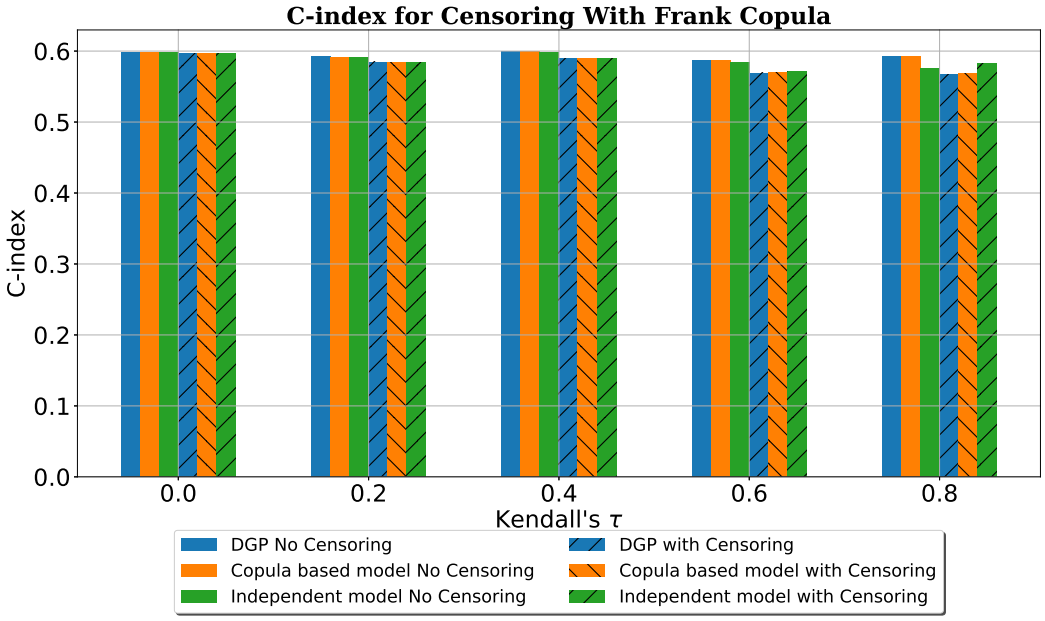

Figure 4.8: C-Index for Censoring in Linear Risk Experiments with Frank Copula

Non-Linear Risk Experiments

Clayton Copula

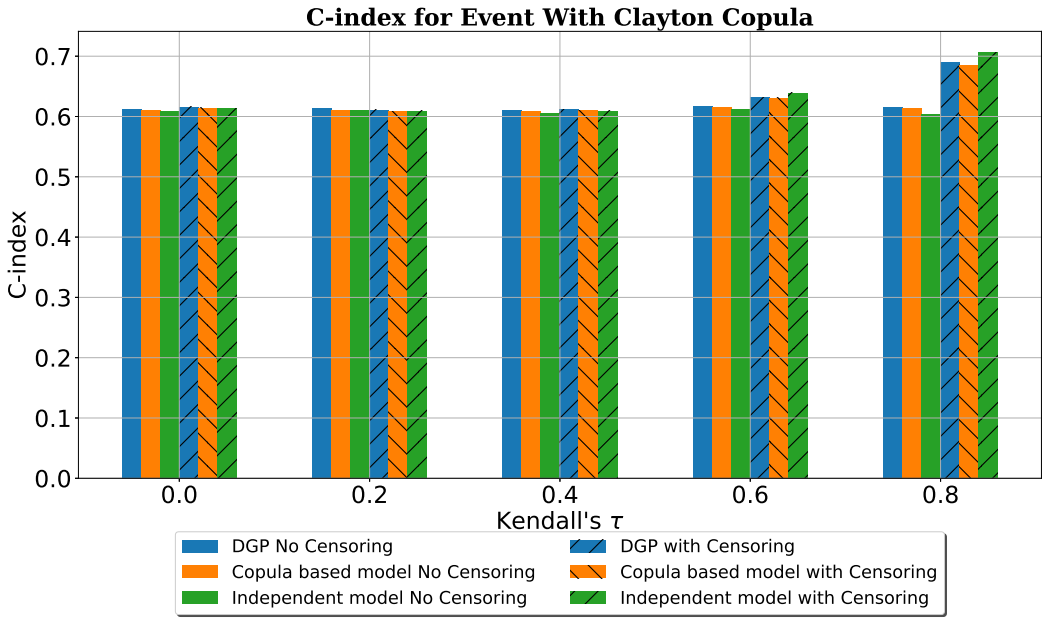

Figure 4.9: C-Index for Event in Non-Linear Risk Experiments with Clayton Copula

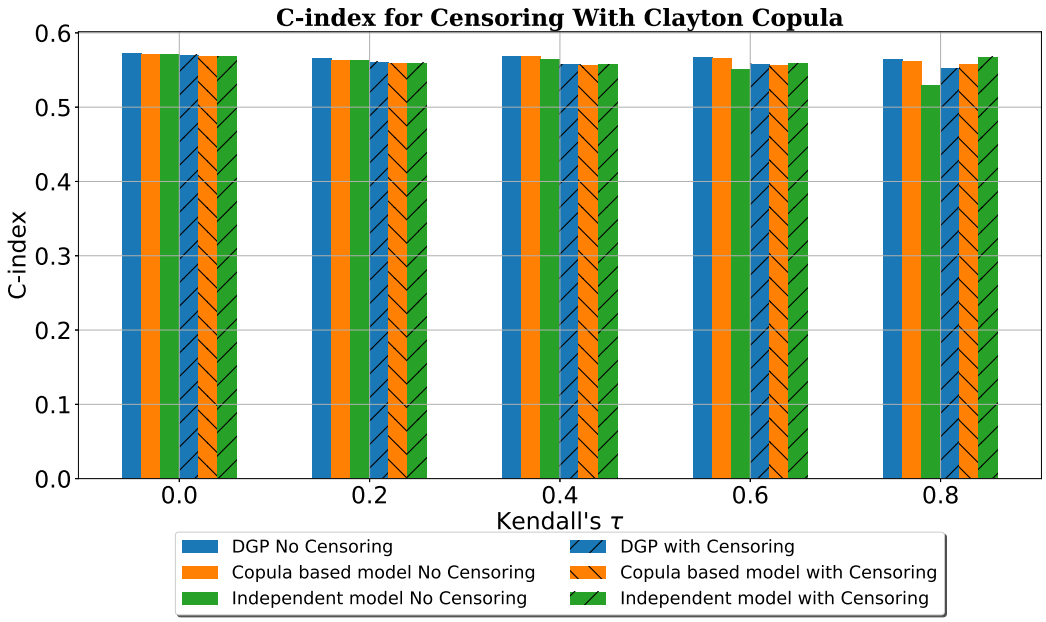

Figure 4.10: C-Index for Censoring in Non-Linear Risk Experiments with Clayton Copula

Frank Copula

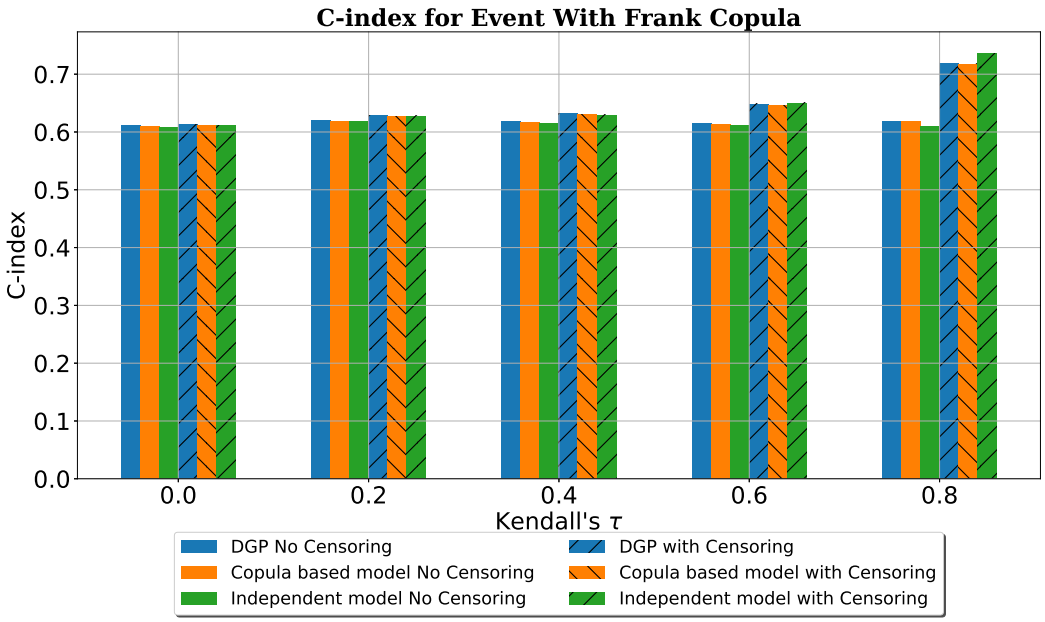

Figure 4.11: C-Index for Event in Non-Linear Risk Experiments with Frank Copula

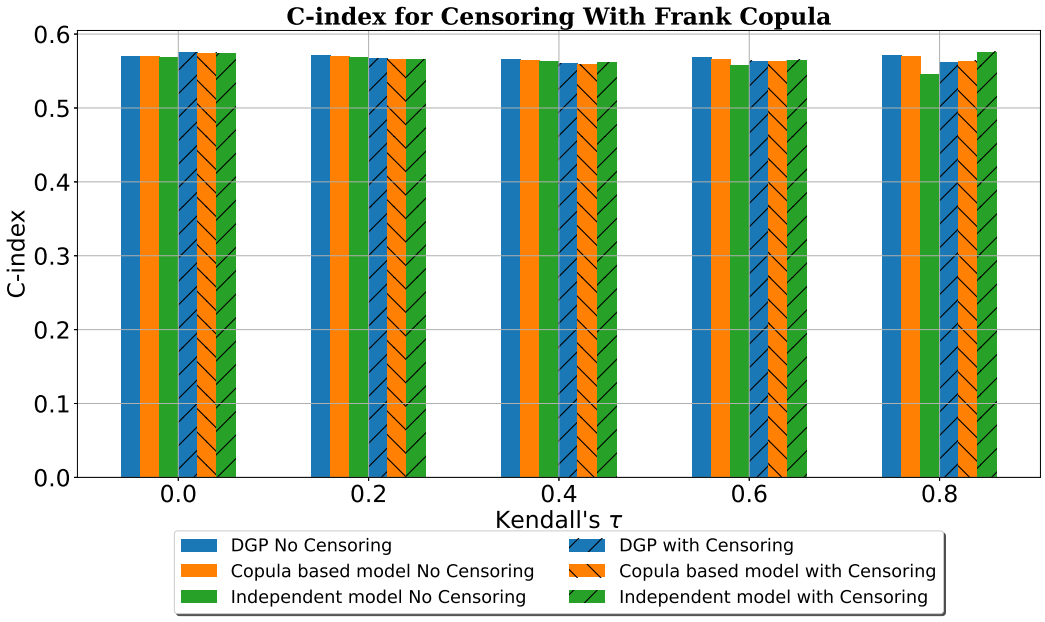

Figure 4.12: C-Index for Censoring in Non-Linear Risk Experiments with Frank Copula

Convex Experiments

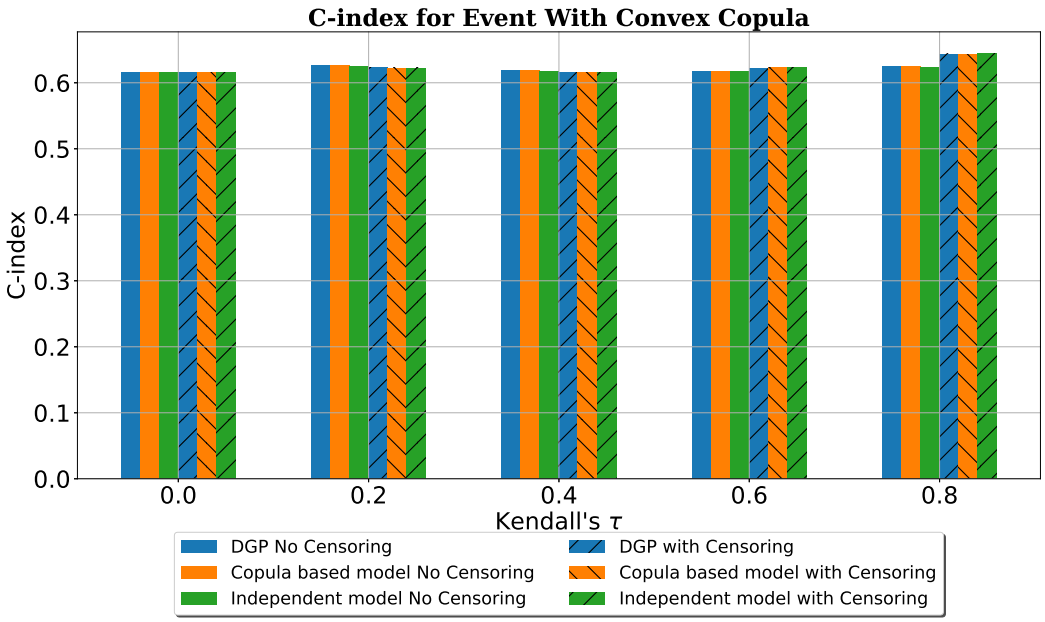

Figure 4.13: C-Index for Event in Linear Risk Experiments with Convex Copula

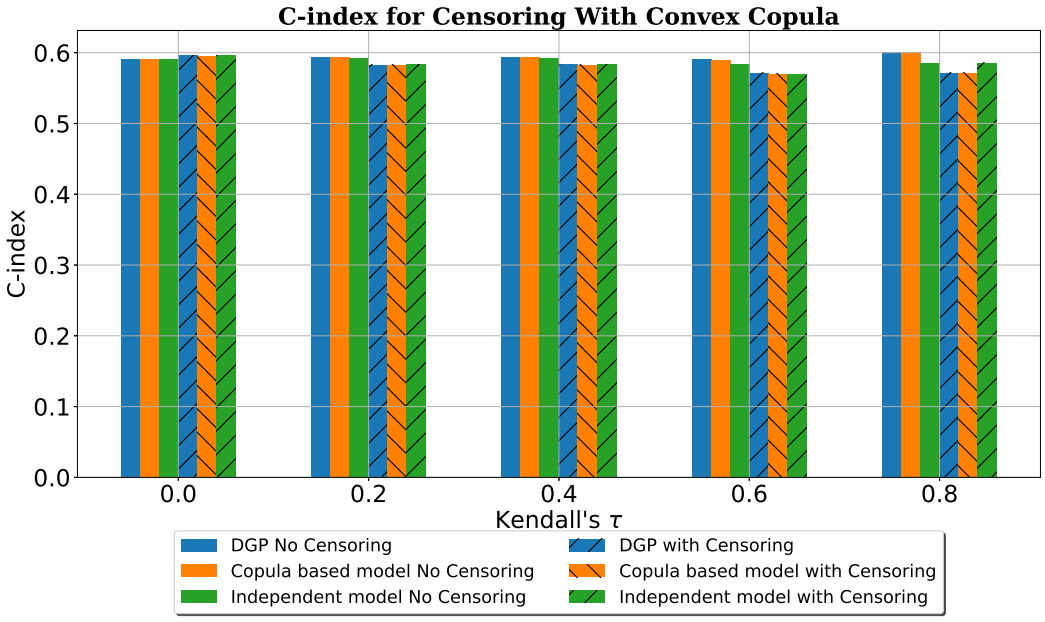

Figure 4.14: C-Index for Censoring in Linear Risk Experiments with Convex Copula

Semi-Synthetic Experiments

Clayton Copula

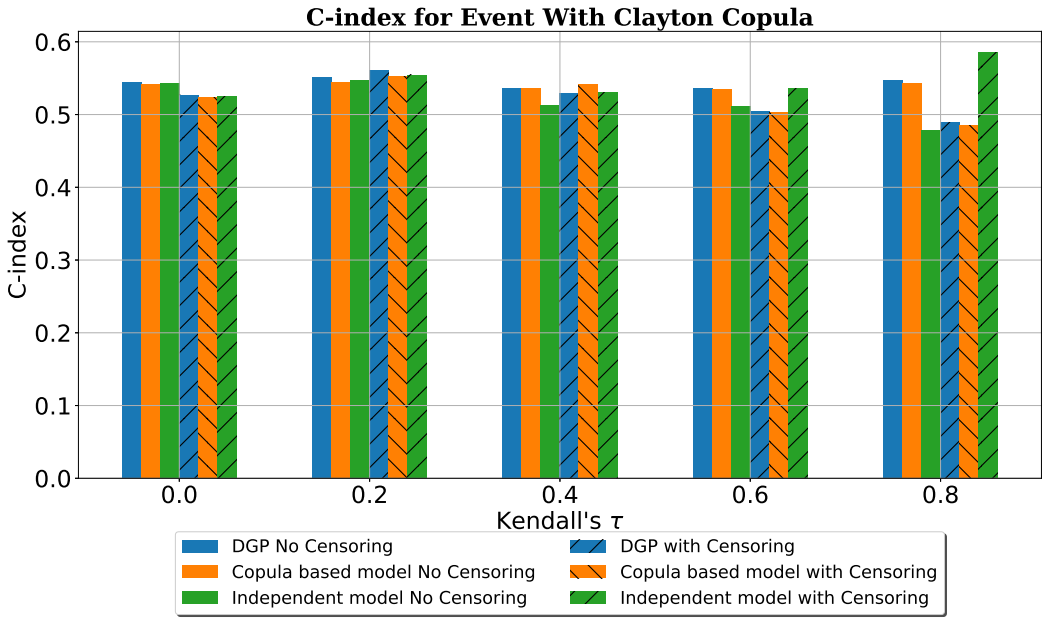

Figure 4.15: C-Index for Event in Semi-Synthetic Experiments with Clayton Copula

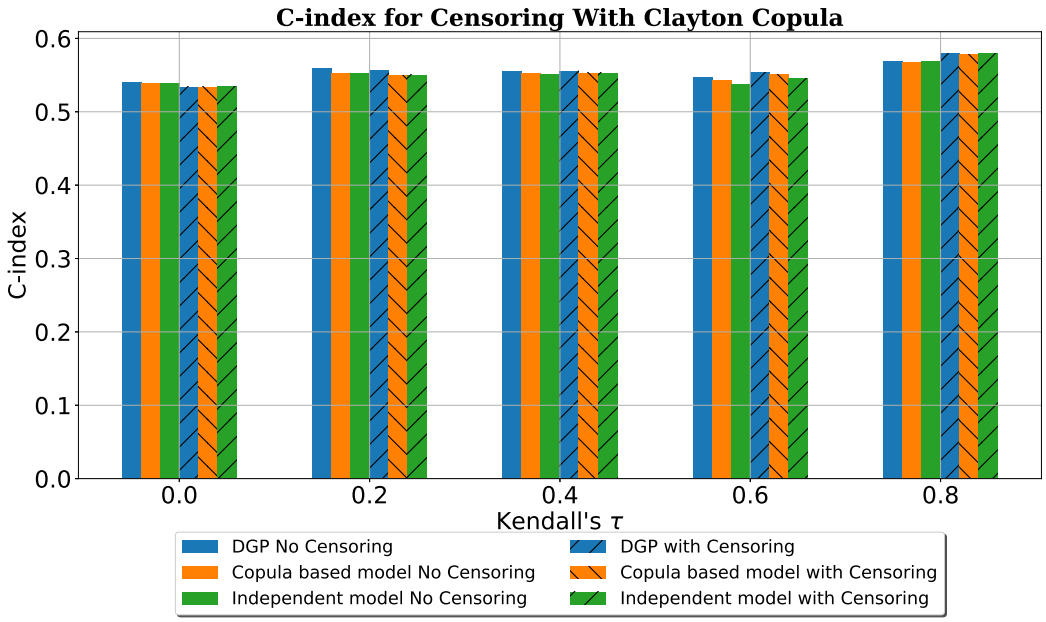

Figure 4.16: C-Index for Censoring in Semi-Synthetic Experiments with Clayton Copula

Frank Copula

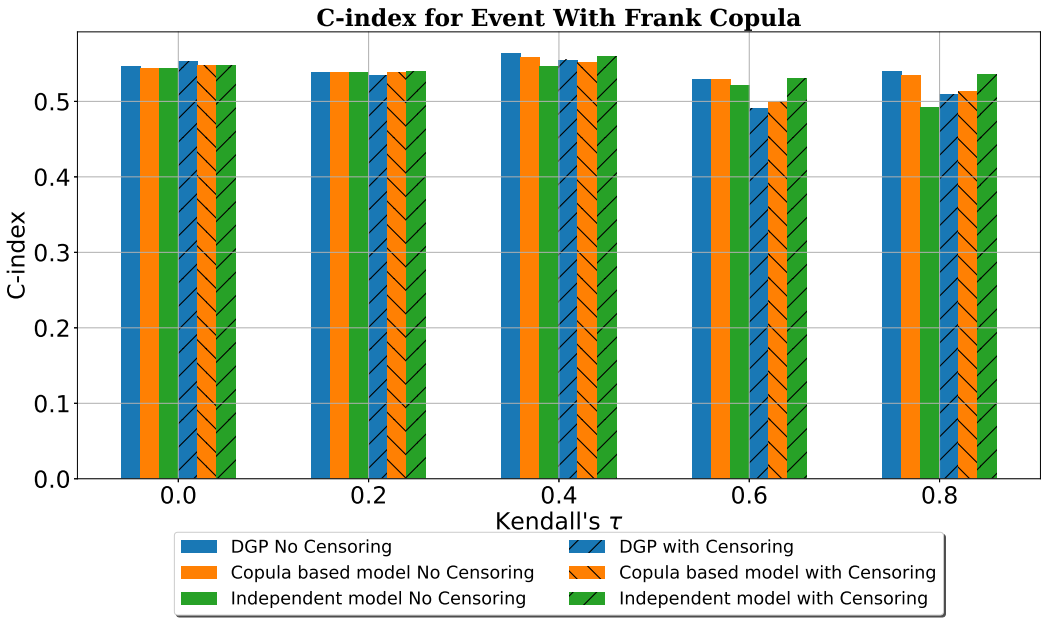

Figure 4.17: C-Index for Event in Semi-Synthetic Experiments with Frank Copula

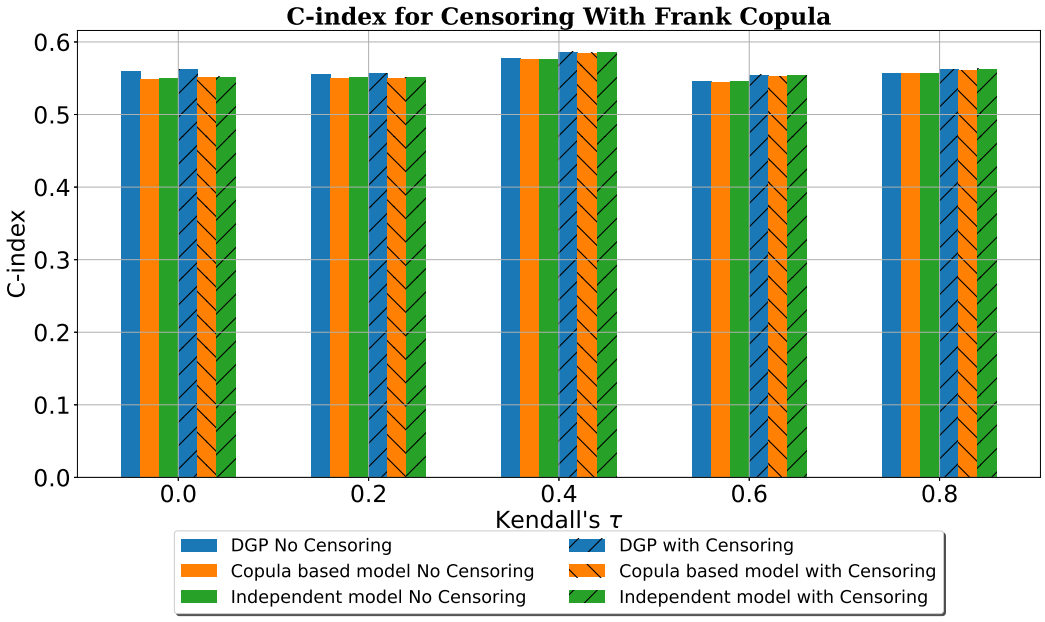

Figure 4.18: C-Index for Censoring in Semi-Synthetic Experiments with Frank Copula

4.6.2 Integrated Brier Score

IBS is a metric frequently used in survival analysis. Similar to our methodology for the C-index, we will evaluate the performance of the DGP, dependent, and independent models for both fully observed scenarios and scenarios where we can only observe the minimum of the event and censoring time.

Our observations suggest that IBS is a reliable metric when there is no censoring, and the model trained with a copula performs comparably well across all levels of dependence. However, we note that the performance gap between the DGP and the independent model widens with increasing dependence.

Based on our observations, we found that IBS is a biased metric under censoring, meaning that the IBS calculated based on the observational dataset differs from the IBS for the dataset without censoring. This bias tends to increase as the level of dependence increases. Additionally, we found that the ranking of methods based on the IBS calculated from the observational dataset does not align with the true IBS calculated based on the no-censoring dataset.

Linear Risk Experiments

Clayton Copula

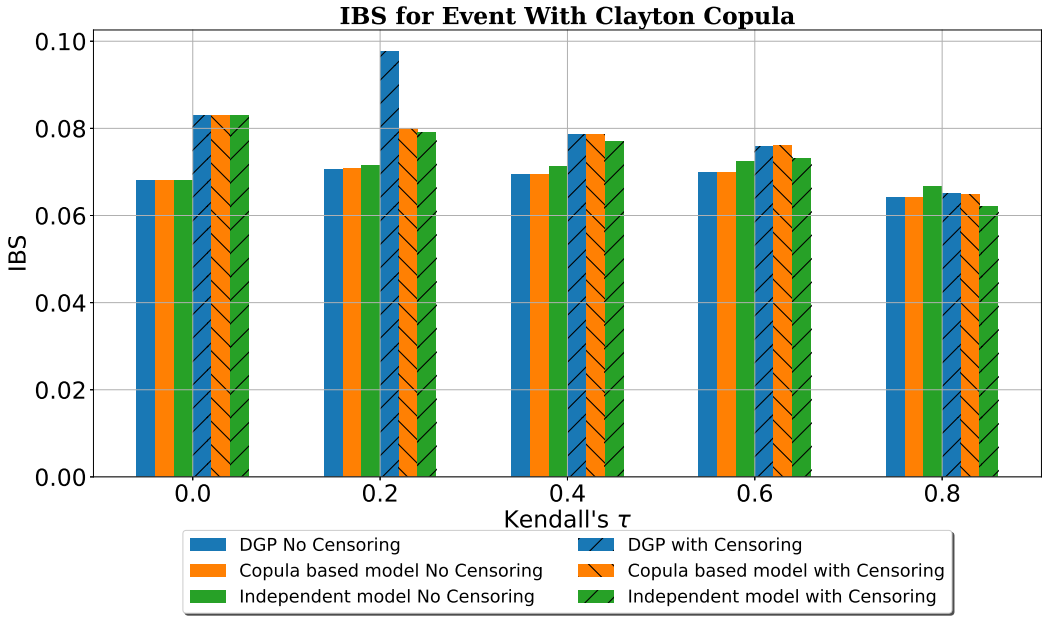

Figure 4.19: IBS for Event in Linear Risk Experiments with Clayton Copula

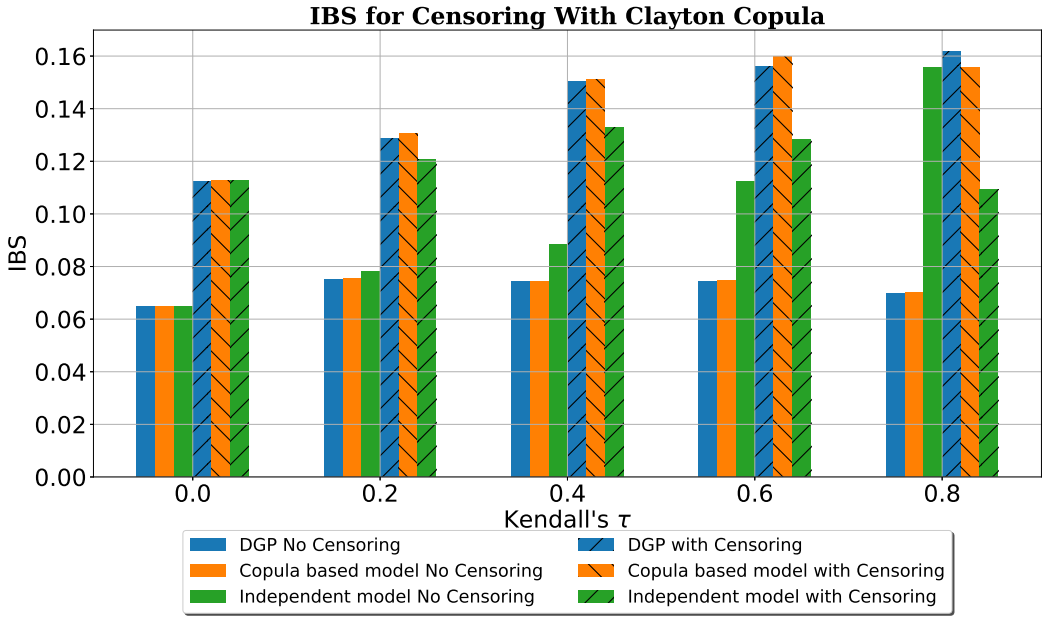

Figure 4.20: IBS for Censoring in Linear Risk Experiments with Clayton Copula

Frank Copula

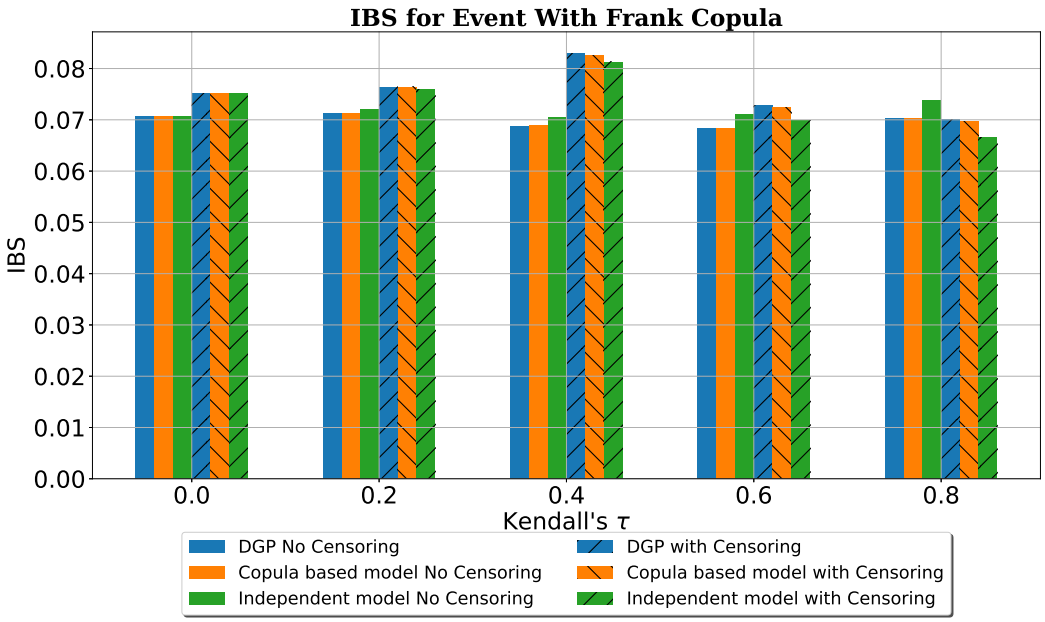

Figure 4.21: IBS for Event in Linear Risk Experiments with Frank Copula

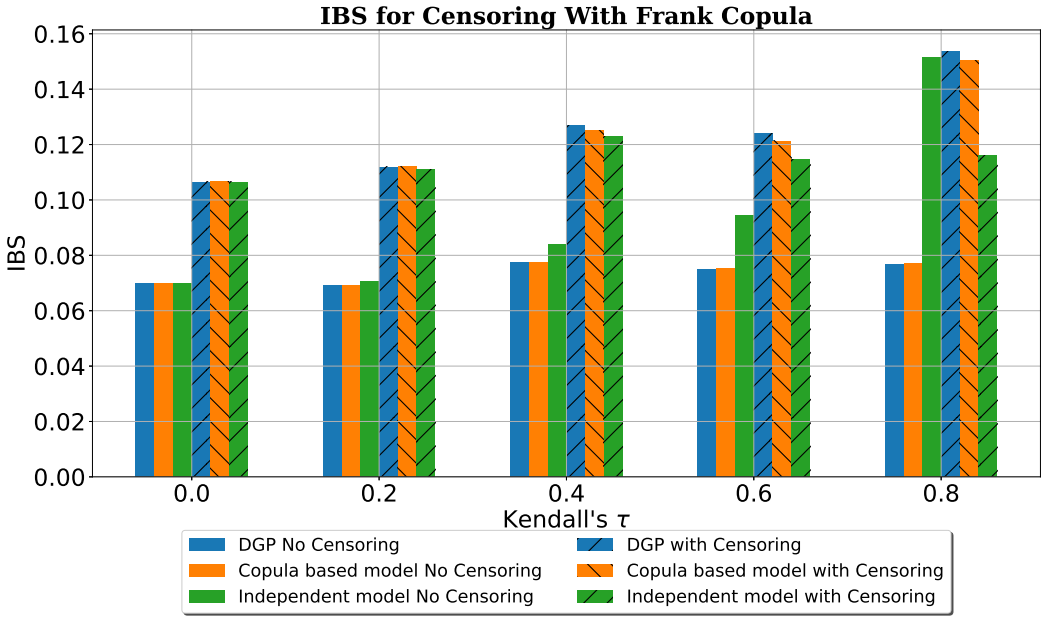

Figure 4.22: IBS for Censoring in Linear Risk Experiments with Frank Copula

Non-Linear Risk Experiments

Clayton Copula

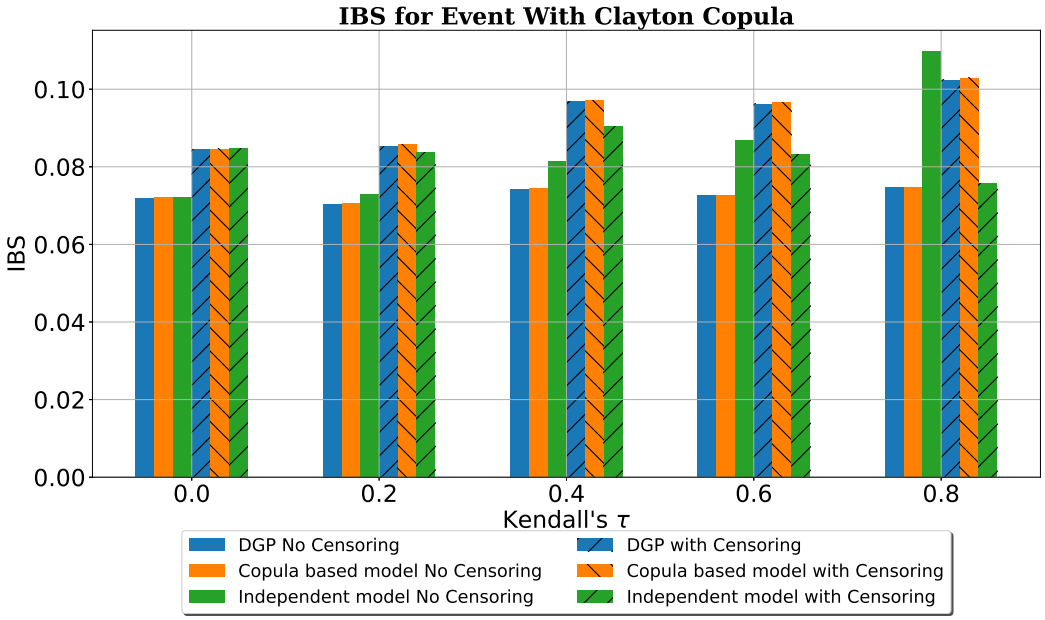

Figure 4.23: IBS for Event in Non-Linear Risk Experiments with Clayton Copula

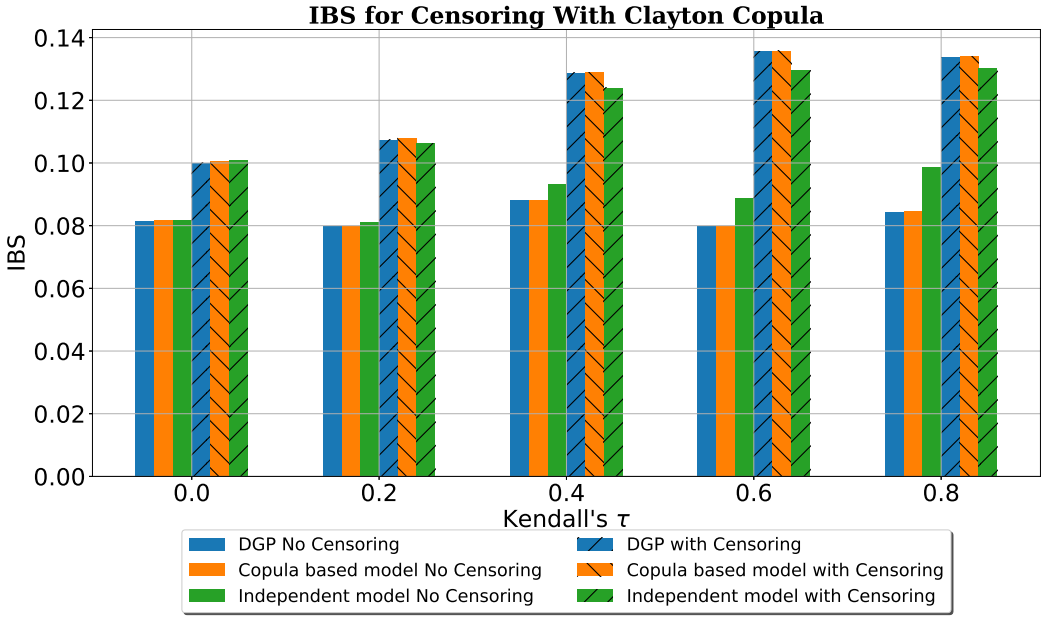

Figure 4.24: IBS for Censoring in Non-Linear Risk Experiments with Clayton Copula

Frank Copula

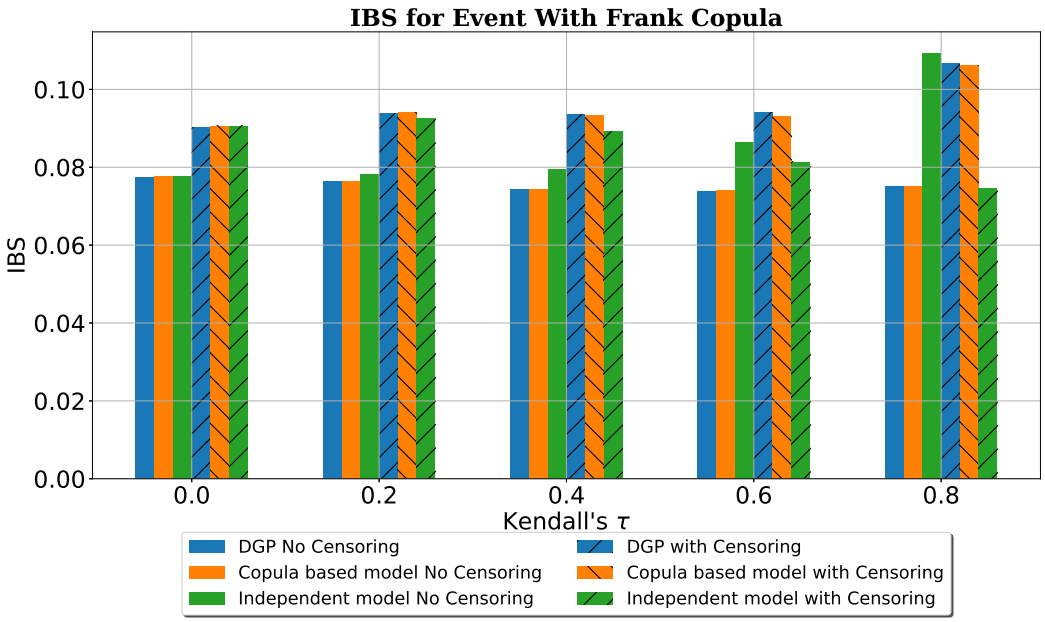

Figure 4.25: IBS for Event in Non-Linear Risk Experiments with Frank Copula

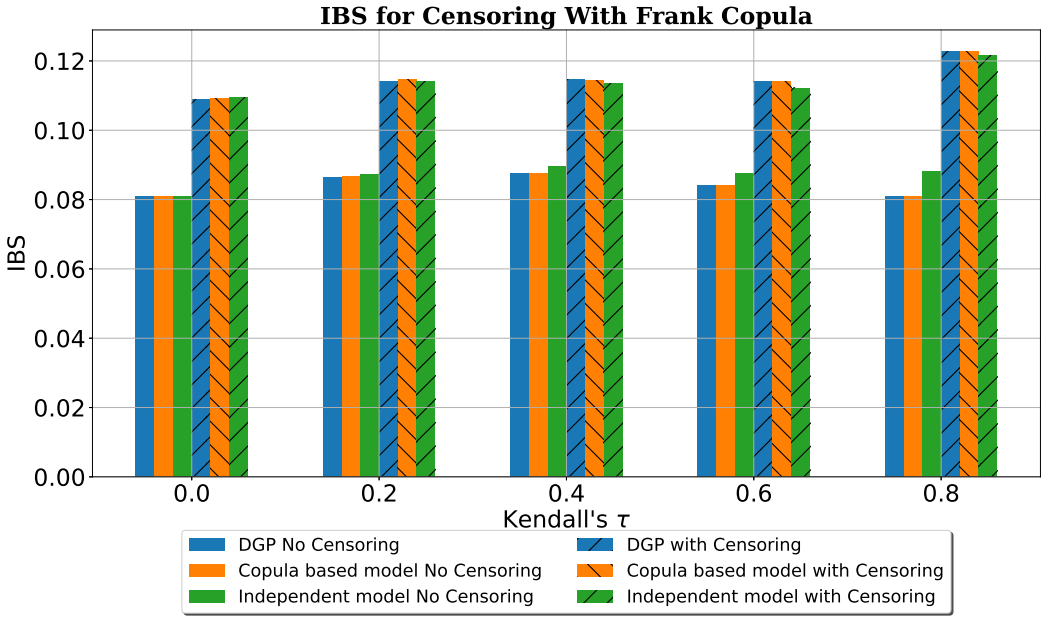

Figure 4.26: IBS for Censoring in Non-Linear Risk Experiments with Frank Copula

Linear Risk Experiments with Convex Copula

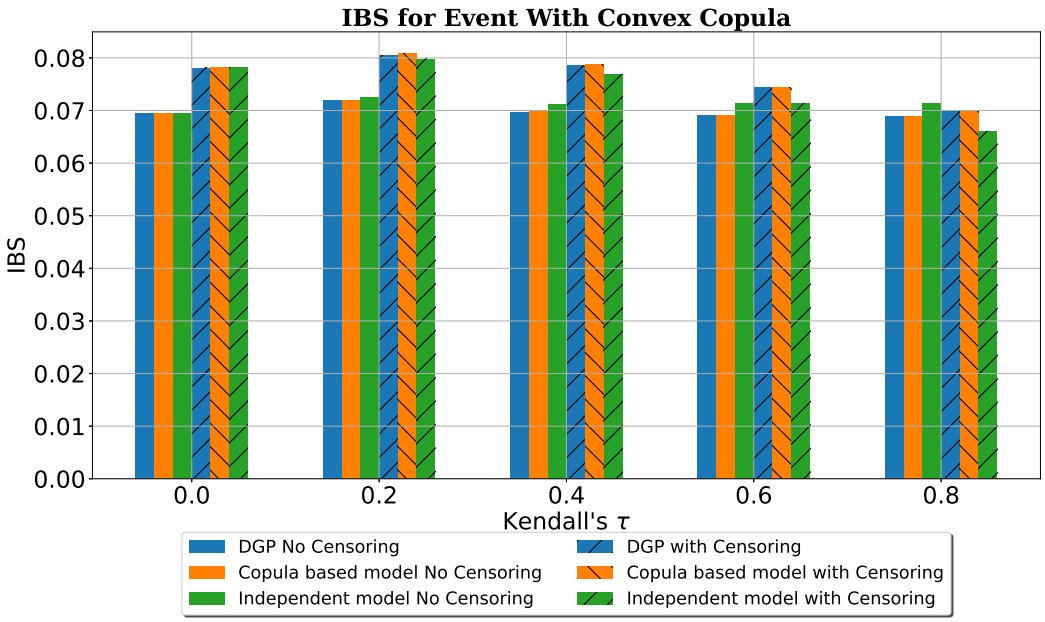

Figure 4.27: IBS for Event in Linear Risk Experiments with Convex Copula

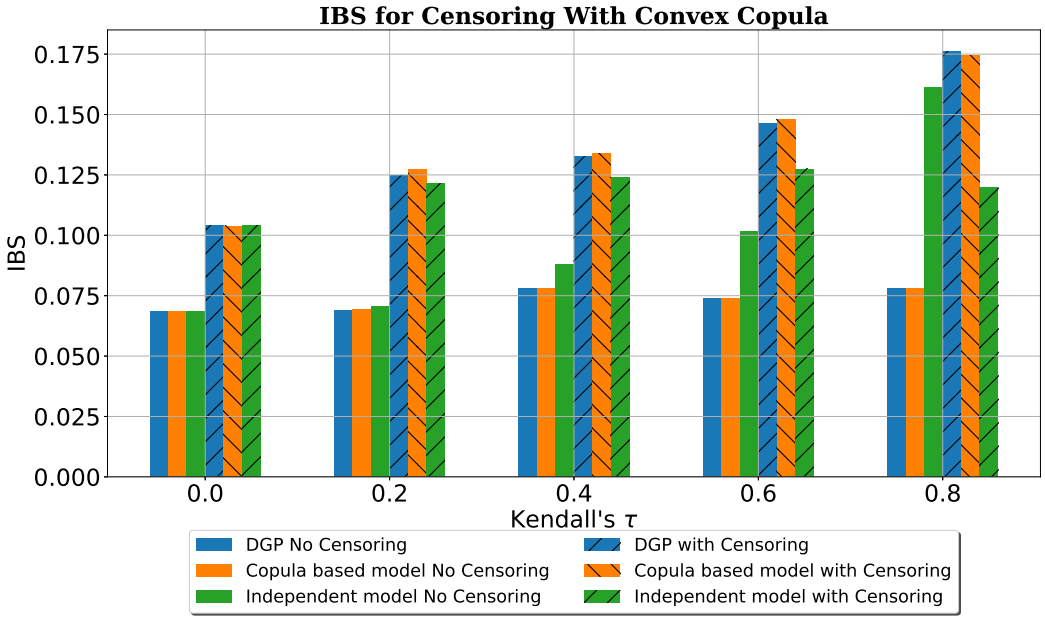

Figure 4.28: IBS for Censoring in Linear Risk Experiments with Convex Copula

Semi-Synthetic Experiments

Clayton Copula

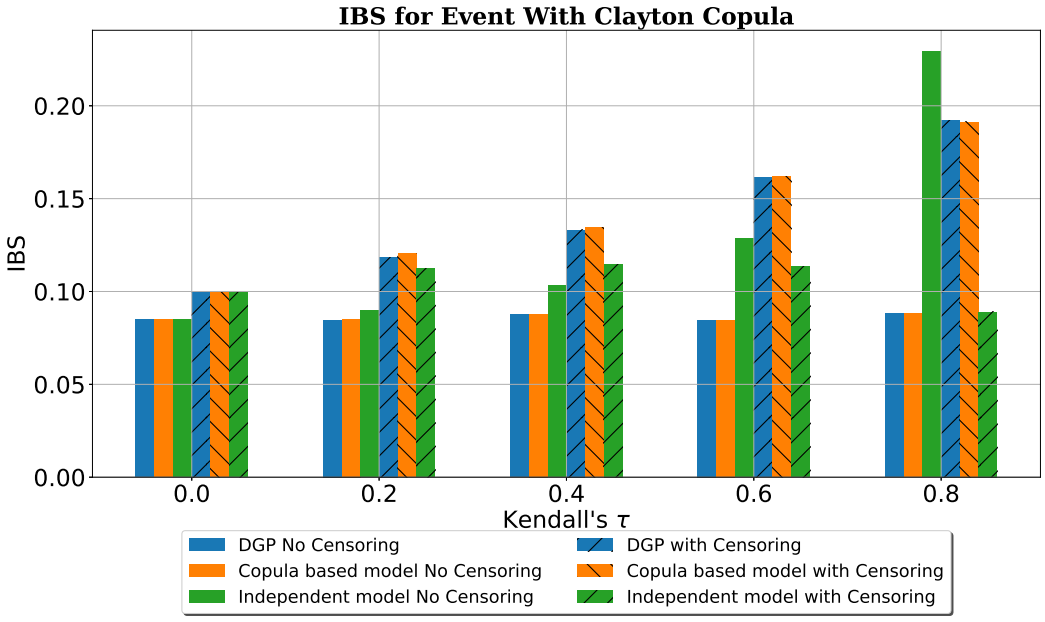

Figure 4.29: IBS for Event in Semi-Synthetic Experiments with Clayton Copula

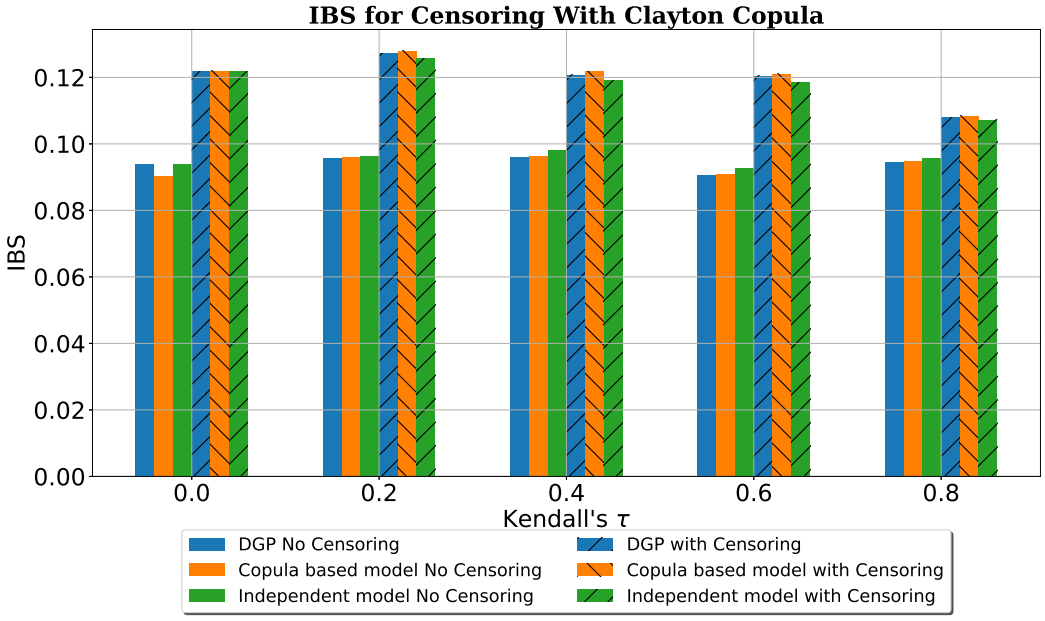

Figure 4.30: IBS for Censoring in Semi-Synthetic Experiments with Clayton Copula

Frank Copula

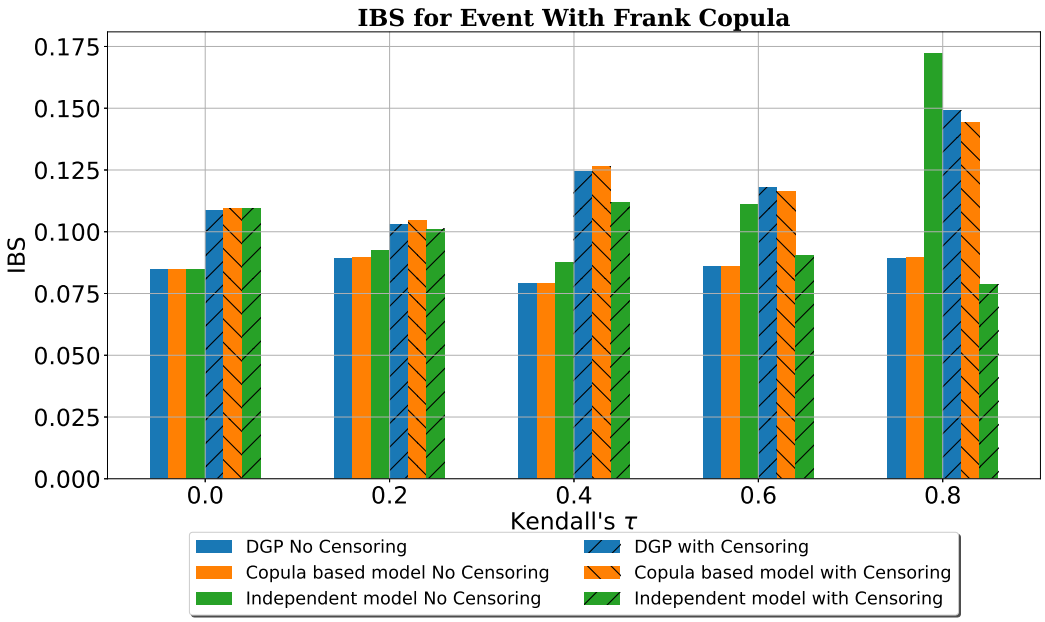

Figure 4.31: IBS for Event in Semi-Synthetic Experiments with Frank Copula

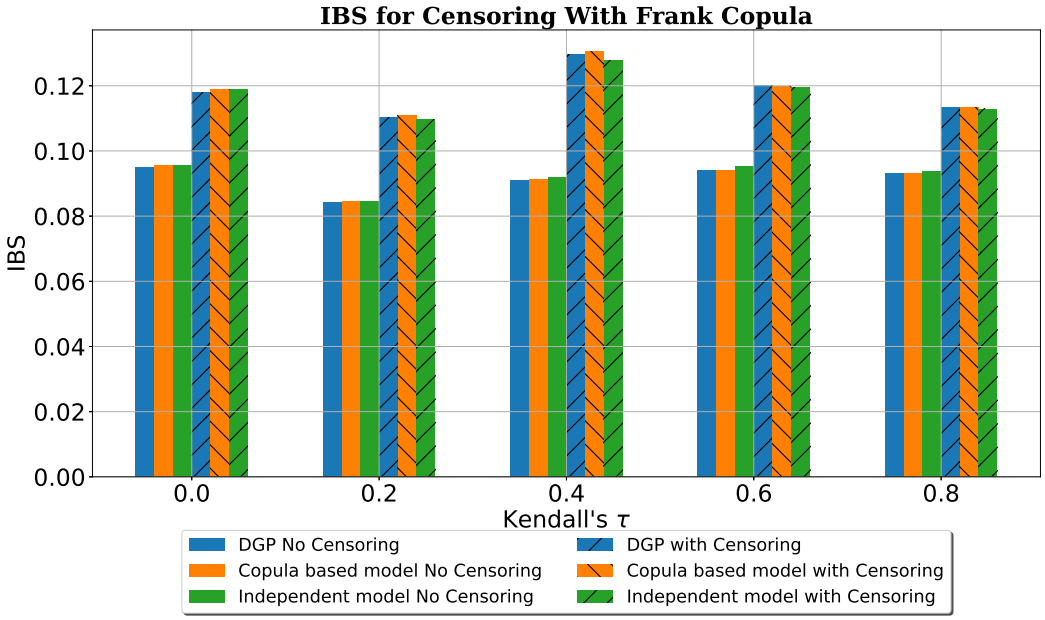

Figure 4.32: IBS for Censoring in Semi-Synthetic Experiments with Frank Copula

Chapter 5

Conclusion

5.1 Contributions

Modern statistical methods in survival analysis increasingly rely on complex, nonlinear functions of risk; however, existing applications of deep learning to survival analysis do not accommodate dependent censoring that may be present in the data. This work relaxes this key assumption, and presents the first neural network-based model of survival to accommodate dependent censoring.

Our experimental results demonstrate the promise of our method: our approach significantly reduces the *Survival*-ℓ_1 (bias) in estimation and shows superior performance in terms of negative log-likelihood. Furthermore, our optimization technique is reliably able to recover the underlying dependence parameter (Kendall's τ) in survival data across datasets of varying feature sizes and different levels and types of dependencies.

We have demonstrated that two commonly used metrics in survival analysis, namely C-index, and IBS, can not identify the bias in survival curves properly. Furthermore, these metrics are biased in the presence of dependent censoring and can be misleading when evaluating the performance of different methods.

5.2 Future Works

The method of using copulas to couple marginal survival distributions is a general one. As future work, we consider extending this approach to other classes of survival models, such as those that do not assume either proportional hazards or a Weibull baseline hazard. We also explore the utilization of neural network-based non-parametric copulas(Archimedean or non-Archimedean) to expand the dependence structure that our method can accommodate.

Though the *Survival*-ℓ_1 metric is a sufficient metric to demonstrate the promise of our approach, it relies on knowledge of the complete survival curve for each instance. In real-world data, we instead typically only have access to point-wise time-of-event or censoring time. The careful study of the behavior of conventional evaluation metrics under dependence, and the design of novel metrics that are faithful reflections of model performance under dependent censoring, remains an open avenue for future work.

Bibliography

[1] O. Aalen, "Nonparametric inference for a family of counting processes," *The Annals of Statistics*, pp. 701–726, 1978.

[2] S. W. Biggins *et al.*, "Serum sodium predicts mortality in patients listed for liver transplantation," *Hepatology*, vol. 41, no. 1, pp. 32–39, 2005.

[3] J. Bona-Pellissier *et al.*, "Parameter identifiability of a deep feedforward relu neural network," *arXiv preprint arXiv:2112.12982*, 2021.

[4] N. E. Breslow, N. E. Day, and E. Heseltine, "Statistical methods in cancer research," 1980.

[5] G. W. Brier *et al.*, "Verification of forecasts expressed in terms of probability," *Monthly weather review*, vol. 78, no. 1, pp. 1–3, 1950.

[6] J. F. Carrière, "Removing cancer when it is correlated with other causes of death," *Biometrical Journal*, vol. 37, no. 3, pp. 339–350, 1995.

[7] Y.-H. Chen, "Semiparametric marginal regression analysis for dependent competing risks under an assumed copula," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 2, pp. 235–251, 2010.

[8] E. Cholongitas *et al.*, "Female liver transplant recipients with the same gfr as male recipients have lower meld scores—a systematic bias," *American journal of transplantation*, vol. 7, no. 3, pp. 685–692, 2007.

[9] D. G. Clayton, "A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence," *Biometrika*, vol. 65, no. 1, pp. 141–151, 1978.

[10] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[11] D. R. Cox, "Partial likelihood," *Biometrika*, vol. 62, no. 2, pp. 269–276, 1975.

[12] M. Crowder, "On the identifiability crisis in competing risks analysis," *Scandinavian Journal of Statistics*, pp. 223–233, 1991.

[13] N. W. Deresa and I. Van Keilegom, "Copula based cox proportional hazards models for dependent censoring," *Journal of the American Statistical Association*, no. just-accepted, pp. 1–23, 2022.

[14] O. J. Dunn and V. A. Clark, *Basic statistics: a primer for the biomedical sciences*. John Wiley & Sons, 2009.

[15] S. Elwir and J. Lake, "Current status of liver allocation in the united states," *Gastroenterology & hepatology*, vol. 12, no. 3, p. 166, 2016.

[16] J. Emmerson *et al.*, "Understanding survival analysis in clinical trials," *Clinical Oncology*, vol. 33, no. 1, pp. 12–14, 2021.

[17] T. Emura *et al.*, "A joint frailty-copula model between tumour progression and death for meta-analysis," *Statistical methods in medical research*, vol. 26, no. 6, pp. 2649–2666, 2017.

[18] T. Emura and Y.-H. Chen, *Analysis of survival data with dependent censoring: Copula-Based Approaches*. Springer, 2018.

[19] G. Escarela and J. F. Carriere, "Fitting competing risks with an assumed copula," *Statistical Methods in Medical Research*, vol. 12, no. 4, pp. 333–349, 2003.

[20] D. Faraggi and R. Simon, "A neural network model for survival data," *Statistics in medicine*, vol. 14, no. 1, pp. 73–82, 1995.

[21] M. J. Frank, "On the simultaneous associativity off (x, y) andx+ y- f (x, y)," *Aequationes mathematicae*, vol. 19, no. 1, pp. 194–226, 1979.

[22] D. Geiger *et al.*, "D-separation: From theorems to algorithms," in *Machine Intelligence and Pattern Recognition*, vol. 10, Elsevier, 1990, pp. 139–148.

[23] T. A. Gerds, M. W. Kattan, M. Schumacher, and C. Yu, "Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring," *Statistics in medicine*, vol. 32, no. 13, pp. 2173–2184, 2013.

[24] E. Graf *et al.*, "Assessment and comparison of prognostic classification schemes for survival data," *Statistics in medicine*, vol. 18, no. 17-18, pp. 2529–2545, 1999.

[25] F. E. Harrell *et al.*, "Evaluating the yield of medical tests," *Jama*, vol. 247, no. 18, pp. 2543–2546, 1982.

[26] S. Hu *et al.*, "Transformer-based deep survival analysis," in *Survival Prediction - Algorithms, Challenges and Applications*, PMLR, 2021, pp. 132–148.

[27] A. J. Hung *et al.*, "A deep-learning model using automated performance metrics and clinical features to predict urinary continence recovery after robot-assisted radical prostatectomy," *BJU international*, vol. 124, no. 3, pp. 487–495, 2019.

[28] X. Jia *et al.*, "A cox-based risk prediction model for early detection of cardiovascular disease: Identification of key risk factors for the development of a 10-year cvd risk prediction," *Advances in preventive medicine*, vol. 2019, 2019.

[29] J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data*. John Wiley & Sons, 2011.

[30] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[31] J. L. Katzman *et al.*, "Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC medical research methodology*, vol. 18, no. 1, pp. 1–12, 2018.

[32] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[33] D. W. Kim *et al.*, "Deep learning-based survival prediction of oral cancer patients," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[35] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," *NeurIPS*, vol. 30, 2017.

[36] J. P. Klein and M. L. Moeschberger, *Survival Analysis Techniques for Censored and Truncated Data*, Second. 2003.

[37] W. A. Knaus, F. E. Harrell, J. Lynn, *et al.*, "The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults," *Annals of internal medicine*, vol. 122, no. 3, pp. 191–203, 1995.

[38] E. Kuntz and H.-D. Kuntz, *Hepatology: Textbook and atlas*. Springer Science & Business Media, 2009.

[39] J. C. Lai *et al.*, "Height contributes to the gender difference in waitlist mortality under the meld-based liver allocation system," *American Journal of Transplantation*, vol. 10, no. 12, pp. 2658–2664, 2010.

[40] C. Lee *et al.*, "Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data," *IEEE TBME*, vol. 67, no. 1, pp. 122–133, 2019.

[41] C. Lee, W. Zame, J. Yoon, and M. Van Der Schaar, "Deephit: A deep learning approach to survival analysis with competing risks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[42] E. T. Lee and J. Wang, *Statistical methods for survival data analysis*. John Wiley & Sons, 2003, vol. 476.

[43] C. R. Lesko *et al.*, "When to censor?" *American journal of epidemiology*, vol. 187, no. 3, pp. 623–632, 2018.

[44] K.-M. Leung *et al.*, "Censoring issues in survival analysis," *Annual review of public health*, vol. 18, no. 1, pp. 83–104, 1997.

[45] I. Lipkovich *et al.*, "Sensitivity to censored-at-random assumption in the analysis of time-to-event endpoints," *Pharmaceutical statistics*, vol. 15, no. 3, pp. 216–229, 2016.

[46] L. Mariani *et al.*, "Prognostic factors for metachronous contralateral breast cancer: A comparison of the linear cox regression model and its artificial neural network extension," *Breast cancer research and treatment*, vol. 44, no. 2, pp. 167–178, 1997.

[47] A. L. Mindikoglu *et al.*, "Impact of estimated liver volume and liver weight on gender disparity in liver transplantation," *Liver Transplantation*, vol. 19, no. 1, pp. 89–95, 2013.

[48] R. Nabi *et al.*, "Full law identification in graphical models of missing data: Completeness results," in *ICML*, PMLR, 2020, pp. 7153–7163.

[49] C. Nagpal *et al.*, "Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks," *IEEE JBHI*, vol. 25, no. 8, pp. 3163–3175, 2021.

[50] R. B. Nelsen, *An introduction to copulas*. Springer Science & Business Media, 2007.

[51] W. Nelson, "Hazard plotting for incomplete failure data," *Journal of Quality Technology*, vol. 1, no. 1, pp. 27–52, 1969.

[52] W. Nelson, "Theory and applications of hazard plotting for censored failure data," *Technometrics*, vol. 14, no. 4, pp. 945–966, 1972.

[53] D. Rindt, R. Hu, D. Steinsaltz, and D. Sejdinovic, "Survival regression with proper scoring rules and monotonic neural networks," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2022, pp. 1190–1205.

[54] L.-P. Rivest and M. T. Wells, "A martingale approach to the copula-graphic estimator for the survival function under dependent censoring," *Journal of Multivariate Analysis*, vol. 79, no. 1, pp. 138–155, 2001.

[55] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[56] M. Schwarz *et al.*, "On the identifiability of copulas in bivariate competing risks models," *Canadian Journal of Statistics*, vol. 41, no. 2, pp. 291–303, 2013.

[57] Y. She *et al.*, "Development and validation of a deep learning model for non–small cell lung cancer survival," *JAMA network open*, vol. 3, no. 6, e205842–e205842, 2020.

[58] M. Sklar, "Fonctions de repartition an dimensions et leurs marges," *Publ. Inst. Statist. Univ. Paris*, vol. 8, pp. 229–231, 1959.

[59] W. Tang *et al.*, "Soden: A scalable continuous-time survival model through ordinary differential equation networks.," *J. Mach. Learn. Res.*, vol. 23, pp. 34–1, 2022.

[60] A. Tsiatis, "A nonidentifiability aspect of the problem of competing risks.," *PNAS*, vol. 72, no. 1, pp. 20–22, 1975.

[61] A. A. Tsiatis, "Semiparametric theory and missing data," 2006.

[62] J. de Uña-Álvarez and N. Veraverbeke, "Generalized copula-graphic estimator," *Test*, vol. 22, no. 2, pp. 343–360, 2013.

[63] H. Uno *et al.*, "On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," *Statistics in medicine*, vol. 30, no. 10, pp. 1105–1117, 2011.

[64] Z. Wang *et al.*, "Survtrace: Transformers for survival analysis with competing events," in *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2022, pp. 1–9.

[65] A. Xiang *et al.*, "Comparison of the performance of neural network methods and cox regression for censored survival data," *Computational statistics & data analysis*, vol. 34, no. 2, pp. 243–257, 2000.

[66] C.-N. Yu, R. Greiner, H.-C. Lin, and V. Baracos, "Learning patient-specific cancer survival distributions as a sequence of dependent regressors," *Advances in neural information processing systems*, vol. 24, 2011.

[67] M. Zheng and J. P. Klein, "Estimates of marginal survival for dependent competing risks based on an assumed copula," *Biometrika*, vol. 82, no. 1, pp. 127–138, 1995.

[68] M. Zheng and J. P. Klein, "Identifiability and estimation of marginal survival functions for dependent competing risks assuming the copula is known," in *Lifetime Data: Models in Reliability and Survival Analysis*, Springer, 1996, pp. 401–408.