Copula-Based Deep Survival Models for Dependent Censoring

Abstract

Survival datasets describe a set of instances (*e.g.*, patients), and provide, for each, either the time until an event (*e.g.*, death), or the censoring time (*e.g.*, when lost to follow-up – which is a lower bound on the time until the event). We consider the challenge of survival prediction: learning, from such data, a predictive model that can produce an individual survival distribution for a novel instance. Many contemporary methods of survival prediction implicitly assume that the event and censoring distributions are independent conditional on the instance's covariates – a strong assumption that is difficult to verify (as we observe only one outcome for each instance) and which can induce significant bias when it does not hold. This paper presents a parametric model of survival that extends modern non-linear survival analysis by relaxing the assumption of conditional independence. Experiments on synthetic and semi-synthetic data demonstrate that our approach significantly improves estimates of survival distributions compared to the standard that assumes conditional independence in the data.[a]

[a]Code available at: [we will link our code repository here at the conclusion of the anonymous review process].

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 an experimental therapy, against that of the patients in the control arm, who did not [Emmerson et al., 2021], 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 [Jia et al., 2019].

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 to patients who did not die prior to the conclusion of the study (administrative censoring) [Leung et al., 1997, Lesko et al., 2018].

Survival prediction refers to the development of statistical models that support time-to-event prediction under right-censorship. Rather than discarding such censored instances, methods in survival analysis methods instead leverage the censoring time as a *lower bound* on that individual's time-of-event [Kalbfleisch and Prentice, 2011]. Let $X^{(i)} \in \mathcal{X}$ refer to a patient's covariates, and let $T_{\text{obs}}^{(i)} \in \mathbb{R}_+$ refer to their time of last observation, taken to be the minimum of the (potentially unobserved) event time $T_E^{(i)} \in \mathbb{R}_+$ and censorship time $T_C^{(i)} \in \mathbb{R}_+$. A common assumption in survival analysis is *conditionally independent censoring* [Kalbfleisch and Prentice, 2011]. That is,

$$T_E \perp T_C \mid X. \tag{1}$$

Equation 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; it is also difficult to verify in practice due to the challenges of high-dimensional conditional independence testing [Shah and Peters, 2020], and because we only observe one outcome (either event or censorship) per instance, but never both. Figure 1 shows 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 does not hold, we say that the data features *dependent censorship*, which is a common feature of survival data that is often unaccounted for in modern methods of survival prediction.

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

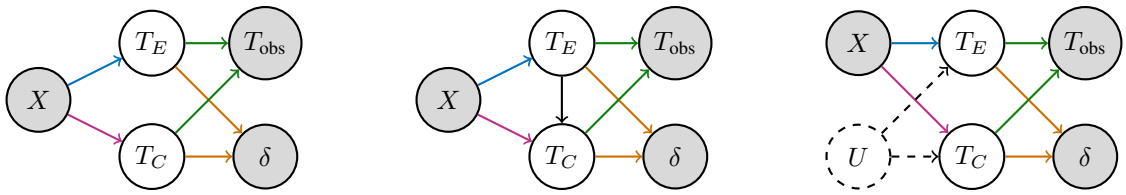

Figure 1: Three graphical models of survival analysis, showcasing the dependencies between covariates X,event/censorship times T_E / T_C, time of last observation T_{obs} and event indicator δ. Shaded nodes represent variables whose values we can observe. Blue and magenta arrows represent the event and censoring functions $f_E : \mathcal{X} \to \mathbb{R}_+$, $f_C : \mathcal{X} \to \mathbb{R}_+$, respectively, of arbitrary functional form. Green arrows into a node T represent the function $\mathbb{R}_+^2 \to \mathbb{R}_+$ defined by $\min(t_e, t_c)$. Orange arrows into a node δ represent the indicator function $\mathbb{R}_+^2 \to \{0, 1\}$ defined by $\mathbb{1}[t_e < t_c]$. The leftmost graph demonstrates the case of conditionally independent censoring (CAR), because conditioning on X d-separates [Geiger et al., 1990] T_E and T_C. The center- and right-most graphs represent cases in which the censoring and event times may be conditionally dependent (CNAR): in the center graph, this is through a direct dependency between T_E and T_C, while in the rightmost graph, this is via the unobserved confounding node, U, that affects both T_E and T_C.

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 represent a life saving intervention for patients suffering from end-stage liver disease [Kuntz and Kuntz, 2009]. Under the Final Rule, the American federal policy of organ allocation, patients are prioritized according to their degree of hepatic dysfunction [Elwir and Lake, 2016]. 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 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 [Biggins et al., 2005], 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 does not hold – *e.g.*, patient sex. Specifically, serum creatinine – a biomarker that tends to be lower in women than in men – plays a key role in the calculation of the MELD-Na score; consequently, Cholongitas et al. [2007] argue that the MELD-Na systematically underestimates womens' 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 [Lai et al., 2010, Mindikoglu et al., 2013]. 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. However, existing approaches either do not permit the incorporation of covariates (*e.g.*, Zheng and Klein [1995], Rivest and Wells [2001], de Uña-Álvarez and Veraverbeke [2013]), or make strict assumptions over the form of the marginal distributions of f_{T_E} and f_{T_C} (*e.g.*, Escarela and Carriere [2003]). These limitations do not admit 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 to correct for dependent censorship in neural network based models of survival outcomes. We present a parameteric proportional hazards model that leverages neural networks to relax assumptions on the distributional form of the marginals, 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.

2. We devise a method to learn both the model and copula parameter from data.

3. We study the challenges associated with evaluating the performance of survival models under dependent censoring. We comment provide commentary on the proprietary of common evaluation metrics under dependent censoring, and prove that the time-dependent concordance index is not proper under dependence.

2 BACKGROUND AND PRELIMINARIES

Our notation will use T_E and T_C where appropriate to refer to the time-of-event/censorship. When a time could refer to either, we will instead simply use T.

2.1 SURVIVAL ANALYSIS PRELIMINARIES

Our work will make use of the following elementary quantities defined by survival analysis. Let $f_{T|X}$, $F_{T|X}$, represent

the conditional density and cumulative distribution functions over the time of an outcome of interest (*e.g.*, event or censorship). Then, we have the following definitions.

Definition 1 (Survival Function)**.** The *survival function*, denoted $S_{T|X}$, represents the likelihood that event (or censorship) will take place after a specifed time, t.

$$S_{T|X}(t|X) \triangleq \Pr(T > t \,|\, X) = 1 - F_{T|X}(t \,|\, X) \quad (2)$$

Definition 2 (Hazard Function)**.** The *hazard function*, denoted h_T, represents the probability that the event will take place within an infinitesimal window in the future, given that it has not yet occurred.

$$h_{T|X}(t|X) \triangleq \lim_{\epsilon \to 0} \Pr(T \in [t, t+\epsilon) \,|\, T \geq t, X) = \frac{f_{T|X}(t|X)}{S_{T|X}(t|X)} \quad (3)$$

Definition 3 (Likelihood Function)**.** The likelihood function for survival data is of the following form.

$$\mathcal{L}(\mathcal{D}) = \prod_{i=1}^{N} \underbrace{\left[\int_{T_{\text{obs}}^{(i)}}^{\infty} f_{T_E,T_C|X}(T_{\text{obs}}^{(i)}, t_c|X^{(i)})\, dt_c \right]}_{\Pr\left(T_E = T_{\text{obs}}^{(i)}, T_C > T_{\text{obs}}^{(i)} | X^{(i)}\right)}^{\delta^{(i)}} \quad (4)$$

$$\underbrace{\left[\int_{T_{\text{obs}}^{(i)}}^{\infty} f_{T_E,T_C|X}(t_e, T_{\text{obs}}^{(i)}|X^{(i)})\, dt_e \right]}_{\Pr\left(T_C = T_{\text{obs}}^{(i)}, T_E > T_{\text{obs}}^{(i)} | X^{(i)}\right)}^{1-\delta^{(i)}}$$

2.2 COPULAS AND SKLAR'S THEOREM

Definition 4 (Copula [Nelsen, 2007])**.** A copula $C(u_1, ..., u_m) : [0,1]^m \to [0,1]$ is a function with the following properties.

1. Groundedness: if there exists an $i \in \{1, ..., m\}$ such that $u_i = 0$, then $C(u_1, ..., u_m) = 0$.
2. Uniform Margins: for all $i \in \{1, ..., m\}$, if $\forall j : j \neq i \Rightarrow u_j = 1$, then $C(u_1, ..., u_m) = u_i$.
3. d-Increasingness: for all $u = (u_1, ..., u_m)$, $v = (v_1, ..., v_m)$ where $u_i < v_i$ for all $i = 1, ..., m$, the following holds:

$$\sum_{l \in \{0,1\}^m} (-1)^{l_1 + ... + l_m} C(u_1^{l_1} v_1^{1-l_1}, ..., u_m^{l_m} v_m^{1-l_m}) \geq 0$$

The utility of copulas as probabilistic objects stems largely from the application Sklar's Theorem [Sklar, 1959], which demonstrates that any joint distribution can be written in terms of a copula over the quantiles of its marginal distributions.

In this work, we will place our emphasis on those copulas that model *joint survival functions*. Such copulas are known as *survival copulas*, and their own version of Sklar's theorem (Equation 5) applies.

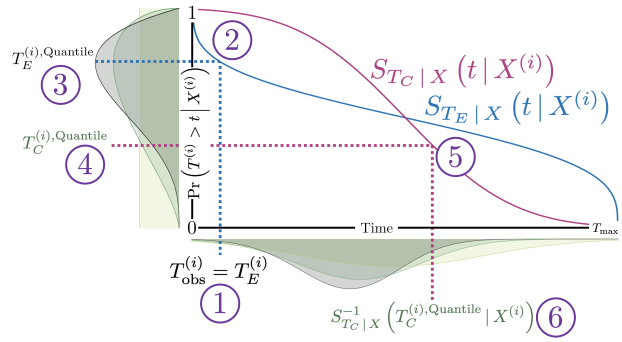

Figure 2: Visualization of how Sklar's Theorem (Survival) models quantile dependency using a copula. **(1)** The observed event time, $T_E^{(i)}$, is **(2)** mapped through the event survival function, $S_{T_E|X}$, to **(3)** obtain an *event quantile*, $T_E^{(i),\text{Quantile}}$. **(4)** A *censoring quantile* is sampled from the copula, $T_C^{(i),\text{Quantile}} \sim C_\theta(\cdot|T_E^{(i),\text{Quantile}})$; the distributions to the left of the vertical axis show the probability mass of $C_\theta(\cdot|T_E^{(i),\text{Quantile}})$ under **no**, **weak**, and **moderate** dependence. Notice how as the dependence increases, the distribution $C_\theta(\cdot|T_E^{(i),\text{Quantile}})$ concentrates mass around $T_E^{(i),\text{Quantile}}$. **(5)** The censoring quantile is then mapped through the inverse censoring survival function, $S_{T_C|X}^{-1}$, to **(6)** obtain a corresponding time-of-censorship, $T_C^{(i)}$. The distributions below the horizontal axis show the distribution of $T_C^{(i)}$ under **no**, **weak**, and **moderate** dependence.

Theorem 1 (Sklar's Theorem (Survival Copulas) [Nelsen, 2007])**.** A survival copula[1] is a copula that applies Sklar's Theorem to survival functions, as follows:

$$S_{T_1,...,T_m}(t_1, ..., t_m) = C(S_{T_1}(t_1), ..., S_{T_m}(t_m)) \quad (5)$$

A visualization of the way in which a copula induces dependency between T_E and T_C via the quantiles of $S_{T_E|X}$ and $S_{T_C|X}$, is shown in Figure 2.

We will focus on two families of copulas, the Clayton [Clayton, 1978] and Frank [Frank, 1979] families. Within these families, the copula C_θ is parameterized by a single parameter, θ, interpreted as the degree of dependence between the marginal distributions under Equation 5. A larger value of θ implies greater dependency between the marginal distributions. We additionally restrict ourselves to bivariate survival copulas, although in principle, these methods could be directly extended to accommodate an arbitrary number of competing events. Both families of copulas converge to the independence copula as θ approaches 0. Uniparametric copulas provide an efficient means of modeling the joint survival function: given that survival analysis already provides tools to model the marginal survival functions S_{T_E}, S_{T_C},

[1] The copula that relates the joint cumulative distribution $F_{X_1,...,X_m}$, with the marginal cumulative distribution functions is typically not the same as that which relates the joint survival function $S_{T_1,...,T_M}$ with the marginal survival functions, though both are valid copulas [Nelsen, 2007].

a model that couples these distribution functions via a uni-parametric copula C_θ only requires adding one additional parameter to the model.

3 RELATED WORK

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 [1995] presented the first neural-network based model of survival, by incorporating a neural network into a Cox Proportional Hazards (CoxPH) model [Cox, 1972]. Although subsequent experimentation found the Farragi-Simon model unable to outperform its linear CoxPH counterpart [Mariani et al., 1997, Xiang et al., 2000], DeepSurv [Katzman et al., 2018] leveraged modern tools from deep learning such as SELU units [Klambauer et al., 2017] and the Adam optimizer [Kingma and Ba, 2014], 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 [Lee et al., 2018] (and its extension to time-varying data, Dynamic-DeepHit [Lee et al., 2019]), Deep Survival Machines [Nagpal et al., 2021], SuMo-net [Rindt et al., 2022], Transformer-based survival models [Hu et al., 2021, Wang et al., 2022], and methods based off of Neural ODEs [Tang et al., 2022] 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.*, Kim et al. [2019], Hung et al. [2019], She et al. [2020]). Therefore, our investigation will focus on relaxing the conditional independence assumption in a parameteric proportional hazards model; we leave to future work the relaxation of the conditional independence assumption in other classes of neural network based 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 one of missing data. The standard taxonomy [Rubin, 1976, Tsiatis, 2006] of missing data 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)* [Leung et al., 1997, Lipkovich et al., 2016]. The conditional independence assumption of Equation 1 is equivalent to asserting either CCAR or CAR in the data.

MNAR data, in the general case, is non-identifiable [Nabi et al., 2020]; 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 [1975] established that, in the general case, the joint 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, and the marginals are continuous, then identifiability holds [Zheng and Klein, 1996, Carrière, 1995]. Crowder [1991] 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. [2013] 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 [1996], Carrière [1995] highlight conditions for identifiability when the form and parameter of the copula are known *a priori*. Schwarz et al. [2013] 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 highlight this as an important area for future study. Many machine learning models remain non-identified [Bona-Pellissier et al., 2021] while remaining useful as predictive and descriptive models. We consider our method a similar approach 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 Emura and Chen [2018]. To our knowledge, the first such work was that of Zheng and Klein [1995] and Rivest and Wells [2001], whose development of the nonparametric Copula-Graphic Estimator extended the Kaplan-Meier Estimator [Kaplan and Meier, 1958] to cases where the dependence between T_E and T_C takes the form of an assumed copula (both C, 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 [Escarela and Carriere, 2003][2]). Proposed semi-parametric estimators

[2]Although Escarela does not directly model dependent censoring, but rather dependent competing events, the approach can be

[Chen, 2010, Emura et al., 2017, Deresa and Van Keilegom, 2022] suffer from much 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.

4 METHODOLOGY

We now present our extension of the Weibull CoxPH model [Barrett, 2014], and discuss the problem of learning nonlinear models of survival outcomes under dependent censorship. Our approach entails modeling each outcome – event and censorship – independently with an extension of the Weibull CoxPH model, and linking them via a copula in the likelihood function during training.

4.1 MODEL 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).[3]

Assumption 2 (Proportional Hazards [Cox, 1972])**.** The hazard for each outcome (event/censorship) can be decomposed into some *baseline hazard* λ_0, dependent only on time, and some *covariate hazard* g, dependent only on the covariates X. That is, there exists some appropriate λ_0, g for which $h_{T|X}(t|X) = \lambda_0(t)\,\exp(\,g(X)\,)$.

4.2 THE WEIBULL COXPH MODEL

Let $\lambda_0(t) = \left(\frac{\nu}{\rho}\right)\left(\frac{t}{\rho}\right)^{\nu-1}$ denote the baseline hazard of the Weibull CoxPH model, and let g_ψ denote a neural network with parameters ψ mapping the covariate space $\mathcal{X}$ to the real line. Then, leveraging the proportional hazards assumption, we define our model in terms of its hazard:

$$h_{T|X}(\,t|X\,) \;=\; \left(\frac{\nu}{\rho}\right)\left(\frac{t}{\rho}\right)^{\nu-1} \exp\left(\,g_\psi(X)\,\right) \tag{6}$$

Let $\phi = \{\nu, \rho, \psi\}$ denote the complete set of model parameters, and observe that the Weibull CoxPH model is fully parametric model over these *marginal parameters* ϕ. By rearranging Equation 6, this class of models readily admits $\hat{S}_{T|X}$, the estimated survival function over each outcome, and $\hat{f}_{T|X}$, the correponding probability mass function.

directly extended to this domain.

[3]In some experiments, we weaken this assumption, and we will explicitly note where this is the case.

These two quantities will allow us to perform maximum likelihood estimation – their derivations are provided in the supplement.

$$\hat{S}_{T|X}(\,t|X\,) \;=\; \exp\left(-\left(\frac{t}{\rho}\right)^{\nu} g_\psi(X)\right) \tag{7}$$

$$\hat{f}_{T|X}(\,t|X\,) \;=\; h_{T|X}(\,t|X\,)\,\hat{S}_{T|X}(\,t|X\,) \tag{8}$$

4.3 MAXIMUM LIKELIHOOD LEARNING UNDER DEPENDENT CENSORSHIP

Let $\left\{\left(X^{(i)}, T^{(i)}_{\text{obs}}, \delta^{(i)}\right)\right\}_{i=1}^{N} \sim \mathcal{D}$ represent a dataset comprising N i.i.d. draws from some data-generating distribution $\mathcal{D}$. Let $X^{(i)} \in \mathcal{X}$ refer to a set of baseline covariates collected about each individual i. Let $T^{(i)}_{\text{obs}} \in \mathbb{R}_+$ refer to their time of last observation, taken to be the minimum of latent variables $T^{(i)}_E \in \mathbb{R}_+$, $T^{(i)}_C \in \mathbb{R}_+$, representing the event and censoring times, respectively. Finally, let $\delta^{(i)} \in \{0, 1\}$ represent an event indicator taking on the value $\mathbb{1}\left[T^{(i)}_E < T^{(i)}_C\right]$. Let C represent a survival copula. Given $\mathcal{D}$, we learn by maximizing the likelihood of the observed data.

Under conditional independence, Equation 4 factorizes and simplifies into the familiar form of the survival likelihood.

$$\mathcal{L}(\mathcal{D}) = \prod_{i=1}^{N} \left[f_{T_E|X}(T^{(i)}_{\text{obs}}|X^{(i)})S_{T_C|X}(T^{(i)}_{\text{obs}}|X^{(i)})\right]^{\delta^{(i)}} \left[f_{T_C|X}(T^{(i)}_{\text{obs}}|X^{(i)})S_{T_E|X}(T^{(i)}_{\text{obs}}|X^{(i)})\right]^{1-\delta^{(i)}} \tag{9}$$

However, when T_E, T_C are no longer conditionally independent, we can no longer rely on this clean decomposition of the log-likelihood. Instead, we make use of the following lemma.

Lemma 2 (Conditional Survival Function Under Sklar's Theorem (Survival))**.** If $S_{T_E,T_C|X}(t_e, t_c|x) = C(u_1, u_2)\Big|_{\substack{u_1=S_{T_E|X}(t_e|x)\\ u_2=S_{T_C|X}(t_c|x)}}$, then,

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

Applying Lemma 2 to Equation 4 yields the log-likelihood for survival models under dependent censorship.

$$\begin{aligned}\ell(\mathcal{D}) = \sum_{i=1}^{N} &\delta^{(i)} \log\left[f_{T_E|X}(T^{(i)}_{\text{obs}}|X^{(i)})\right] + \\ &\delta^{(i)} \log\left[\frac{\partial}{\partial u_1} C(u_1, u_2)\Big|_{\substack{u_1=S_{T_E|X}(T^{(i)}_{\text{obs}}|X^{(i)})\\ u_2=S_{T_C|X}(T^{(i)}_{\text{obs}})|X^{(i)}}}\right] + \\ &(1-\delta^{(i)}) \log\left[f_{T_C|X}(T^{(i)}_{\text{obs}}|X^{(i)})\right] + \\ &(1-\delta^{(i)}) \log\left[\frac{\partial}{\partial u_2} C(u_1, u_2)\Big|_{\substack{u_1=S_{T_E|X}(T^{(i)}_{\text{obs}}|X^{(i)})\\ u_2=S_{T_C|X}(T^{(i)}_{\text{obs}})|X^{(i)}}}\right]\end{aligned} \tag{10}$$

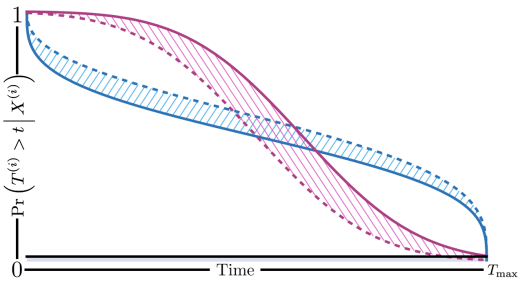

Figure 3: The *Survival*-ℓ_1 metric, $\mathcal{C}_{\text{Survival-}\ell_1}(S, \hat{S})$, for event and censoring distributions. Dashed lines represent the predicted survival curves, $\hat{S}_{T_E|X}$, and $\hat{S}_{T_C|X}$, while solid lines represent the corresponding ground-truth survival curves, $S_{T_E|X}$, and $S_{T_C|X}$. The black horizontal line represents the normalizing quantile, $Q_{\|\cdot\|}$. The area of the hatched blue region above $Q_{\|\cdot\|}$ is the value of $\mathcal{C}_{\text{Survival-}\ell_1}(S_{T_E|X}, \hat{S}_{T_E|X})$, while that of the hatched pink region is the value of $\mathcal{C}_{\text{Survival-}\ell_1}(S_{T_C|X}, \hat{S}_{T_C|X})$.

In this expression, the first term corresponds to the log likelihood of observing the event at time $T^{(i)}_{\text{obs}}$. The second term corresponds to the conditional probability of observing the censorship time after the event time, given that the event time is $T^{(i)}_{\text{obs}}$. The third and fourth terms, by symmetry, represent the same quantities for the censorship time. Despite the visual complexity of Equation 10, 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. Algorithm 1 details the optimization procedure used to jointly optimize the marginal and copula parameters. Specific implementation details and hyperparameters are discussed in the supplement.

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 $\hat{S}^{(\mathcal{M})}_{T|X}(t|X)$, $\hat{f}^{(\mathcal{M})}_{T|X}(t|X)$, for each $X^{(i)} \in \mathcal{D}$; α: learning rate for event model, censoring model, and copula parameter.

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

$\mathcal{M}_E \leftarrow$ `Instantiate`$(\mathcal{M}; \hat{\psi}^{(0)}_E)$;
$\mathcal{M}_C \leftarrow$ `Instantiate`$(\mathcal{M}; \hat{\psi}^{(0)}_C)$;
$C_\theta \leftarrow$ `Instantiate`$(C; \hat{\theta}^{(0)})$;
for $i = 1, \ldots,$ *NUM_EPOCHS* **do**
  $\mathcal{L}_i \leftarrow \ell\left[\mathcal{D}; \hat{f}^{(\mathcal{M}_E)}_{T|X}, \hat{f}^{(\mathcal{M}_C)}_{T|X}, \hat{S}^{(\mathcal{M}_E)}_{T|X}, \hat{S}^{(\mathcal{M}_C)}_{T|X}, C_{\hat{\theta}^{(i)}}\right]$;
  $\hat{\psi}^{(i)}_C \leftarrow$ `AdamUpdate`$(\mathcal{L}_{\text{EPOCH}}, \hat{\psi}_C, \alpha)$;
  $\hat{\psi}^{(i)}_E \leftarrow$ `AdamUpdate`$(\mathcal{L}_{\text{EPOCH}}, \hat{\psi}_E, \alpha)$;
  $\hat{\theta}^{(i)} \leftarrow$ `AdamUpdate`$(\mathcal{L}_{\text{EPOCH}}, \hat{\theta}, \alpha)$;
end
return $\hat{\theta}^{(i)}, \hat{\psi}^{(i)}_E, \hat{\psi}^{(i)}_C$

4.4 EVALUATION METRICS

The presumption that standard metrics such as concordance index[Harrell et al., 1982, Uno et al., 2011], time-dependent concordance (TDCI) Gerds et al. [2013], and Brier score [Brier et al., 1950] represent means to effectively evaluate models learned under dependent censoring is a dubious one. Though it is a known result that the Brier Score is not proper under dependent censoring, we prove that the TDCI is also not proper under such dependence. The proof of Theorem 3 is in the supplement.

Theorem 3. The time-dependent concordance index is not a proper scoring rule under dependent censoring.

The Survival-ℓ_1 Metric: Due to the drawbacks of conventional scoring rules, we introduce the *Survival*-ℓ_1 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), over the lifespan of the curves. In the supplement, 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 elapsed time under each survival curve. To ensure that survival curves over longer lifespans do not contribute proportionally more to the evaluation metric than those over shorter lifespans, 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.

Our *Survival*-ℓ_1 metric therefore takes the following form:

$$\mathcal{C}_{\text{Survival-}\ell_1}(S, \hat{S}) = \sum_{i=1}^{N} \frac{1}{N \times T^{(i)}_{\max}} \int_0^\infty \left| S_{T|X}(t|X^{(i)}) - \hat{S}_{T|X}(t|X^{(i)}) \right| dt \tag{11}$$

5 EXPERIMENTAL EVALUATION

Synthetic Data: 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.

Algorithm 2 provides a means of generating synthetic data under a specified copula C with Weibull CoxPH margins.

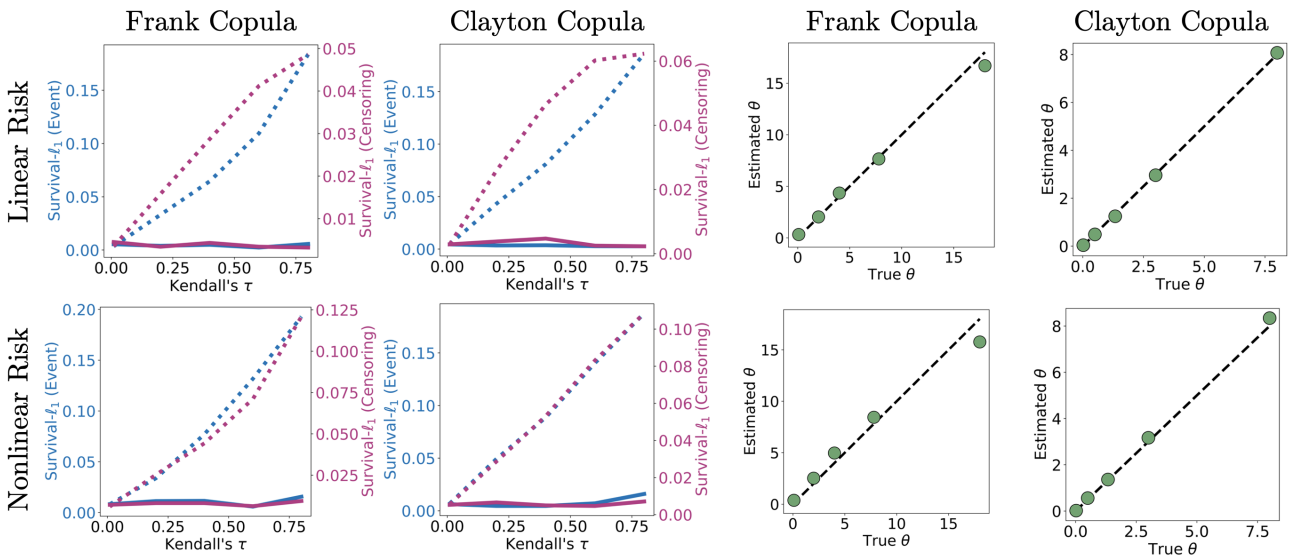

Figure 4: **Left**: Plot of the bias, $\mathcal{C}_{\text{Survival-}\ell_1}$, as a function of the dependence (Kendall's τ), for both independence-assuming and copula-based models on synthetic data. Going from left to right on the x-axis denotes stronger dependence between the survival and event time 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. The copula-based approach yields a lower line for each event, indicating a better approximation of the ground-truth survival function. **Right**: For each value of τ in the left plot, we plot the recovered copula value $\hat{\theta}$ as a function of the dependence, θ^*. The dotted line, representing $\hat{\theta} = \theta^*$, is plotted for reference. Points close to the line indicate that the learned dependence parameter was close to that of the data-generating process.

For the `Linear-Risk` experiment shown in Figure 4, we generate data according Algorithm 2 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 $\beta_E, \beta_C \in [0,1]^{10} \sim \mathcal{U}_{[0,1]}$. For the `Nonlinear-Risk` experiment, we run Algorithm 2 with $X \in \mathbb{R}^{N\times 1} \sim \mathcal{U}_{[0,1]}$, $\nu_E^* = 4, \rho_E^* = 17, \psi_E^*(X) = 2\sin(\pi X)$, $\nu_C^* = 3, \rho_C^* = 16, \psi_C^*(X) = 2\sin(\pi X + 0.1)$. Each synthetic experiment was performed on $20,000$ train, $10,000$ validation, and $10,000$ test samples.

The network g_ψ in the model we train on the `Linear-Risk` data consists of a single linear layer, while the network g_ψ in the model we train on the `Nonlinear-Risk` data consists of a three-layer fully-connected neural network with ReLU activations and hidden layers consisting of $[32, 32, 32, 1]$ dimensions, respectively.

Semi-Synthetic Data: to investigate the promise of our approach on non-synthetic data, we artificially censor a regression dataset according to a various degrees of dependence. We choose a dataset (`STEEL`) from the UCI Machine Learning Repository [Asuncion and Newman, 2007], the task of which is predicting the energy usage of a steel production plant. We induce censoring in the data according to Algorithm 4 in the supplement. We then train a linear version of our method on the artificially censored dataset and evaluate our performance via the R^2 statistic. In this experiment, we compare our approach against two baselines: a linear Weibull CoxPH model trained on the regression data *without censoring* (which should represent an upper bound on our model's performance), and a linear independence-assuming Weibull CoxPH model.

Algorithm 2: 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 $\mathcal{D}$

Our results highlight three properties of our framework. First, our model is capable of reducing the bias in the learned individual survival curve (as measured by the *Survival-*ℓ_1 metric). Second, the learning algorithm does, in many cases, recover the ground truth coefficient associated with the cop-

ula when parameterizing the prediction of the event and censoring time with neural networks. Finally, our framework opens up new avenues to learning more complex forms of dependence between event and survival time.

Reducing bias in survival outcomes: Figure 4 (left column) plots the model bias as measured by the *Survival*-ℓ_1, and how it behaves across datasets (in rows of plots). Within each plot, going from left to right on the x-axis denote stronger dependence between the survival and event time in the data generating process. The y-axis is overloaded, the scales on the left depict the magnitude of bias associated with event time and the scale on the right denote the magnitude of bias associated with censorship time.

We highlight that our approach of modeling the dependence structure between event and censorship times reduces the bias in the model's estimation of survival curves. The bias is substantially lower under our approach for all values of $\tau > 0$, and we note that the improvements are more pronounced for larger values of τ indicating that the improvements in our approach are larger as the dependence between censorship and event time is stronger. We see consistent results holding for both the `Linear-Risk` and `Nonlinear-Risk` data-generating processes, and for both the Frank and Clayton families of copula. In the special case where $\tau = 0$, we observe that our approach correctly recovers the independence copula, and learns an unbiased survival curve.

Our results on the artificially censored `STEEL` dataset suggest that our method also shows promise on non-synthetic data. Under high dependence ($\tau = 0.8$), our method achieves an R^2 of 0.508, compared to the independence-assuming model, which achieves an R^2 of 0.341.

Empirical Recovery of the Copula Parameter: How close are the recovered parameters of the copula to the true parameters used in the data-generating process? Although we do not have a formal proof of identifiability, we nevertheless study this question empirically on the two datasets in Figure 4 (right column). Here, we find that our approach is able to reliably recover a $\hat{\theta}$ that is close to θ^* across different datasets and families of copula.

Relaxating Assumption 1: Next, we showcase the flexibility of our framework via a relaxation of Assumption 1. Specifically, rather than parameterizing our model with C_θ, a single copula of an assumed functional form, we instead parameterize it with a convex combination of Clayton and Frank copulas. During optimization, we learn θ_{Frank}, θ_{Clayton}, and κ, a mixing parameter. Figure 5 shows the results of an experiment on synthetic data generated from a mixture where $\theta_{\text{Frank}} = \theta_{\text{Clayton}}$, $\kappa = 0.5$. As in the case where the functional form of C was known, the mixture model reduces bias in estimation of the event and censoring distributions.

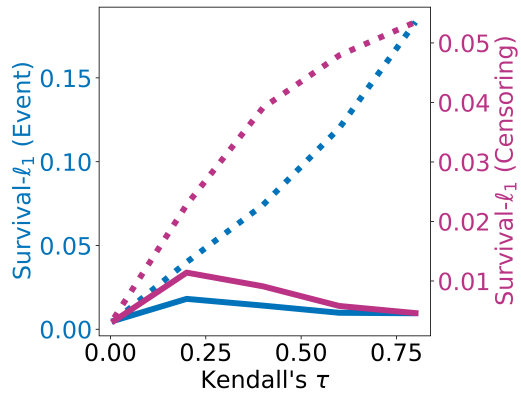

Figure 5: Plot of the bias ($\mathcal{C}_{\text{Survival-}\ell_1}$) as a function of dependence (Kendall's τ), for independence-assuming and copula-based Weibull CoxPH models on synthetic data with linear margins drawn from a convex combination of copulas. In this experiment, the model optimized a convex combination of two copulas (one Frank, one Clayton), rather than a single uniparametric copula. As in Figure 4, the dashed lines represent the bias incurred by independence-assuming models, while the solid lines represent the bias incurred by our approach. This figure highlights that our method is capable of relaxing Assumption 1 by way of a convex combination of copulas.

6 CONCLUSION

Future Work: 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 neural survival models, such as those that do not assume either proportional hazards or a Weibull baseline hazard. 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 information. The careful study of the behaviour 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.

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 our optimization technique is reliably able to recover the underlying dependence parameter in survival data across datasets of varying feature sizes.

References

Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

J Barrett. Weibull-cox proportional hazard model. *Inst. Mathematical and Molecular Biomedicine*, 2014.

Scott W Biggins et al. Serum sodium predicts mortality in patients listed for liver transplantation. *Hepatology*, 41 (1):32–39, 2005.

Joachim Bona-Pellissier et al. Parameter identifiability of a deep feedforward relu neural network. *arXiv preprint arXiv:2112.12982*, 2021.

Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Jacques F Carrière. Removing cancer when it is correlated with other causes of death. *Biometrical Journal*, 37(3): 339–350, 1995.

Yi-Hau Chen. Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):235–251, 2010.

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*, 7 (3):685–692, 2007.

David G Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978.

D Cox. Regression models and life-tables. *J Royal Statistical Society: Series B (Methodological)*, 34(2), 1972.

Martin Crowder. On the identifiability crisis in competing risks analysis. *Scandinavian Journal of Statistics*, pages 223–233, 1991.

Jacobo de Uña-Álvarez and Noël Veraverbeke. Generalized copula-graphic estimator. *Test*, 22(2):343–360, 2013.

Negera Wakgari Deresa and Ingrid Van Keilegom. Copula based cox proportional hazards models for dependent censoring. *Journal of the American Statistical Association*, (just-accepted):1–23, 2022.

Saleh Elwir and John Lake. Current status of liver allocation in the united states. *Gastroenterology & hepatology*, 12 (3):166, 2016.

J Emmerson et al. Understanding survival analysis in clinical trials. *Clinical Oncology*, 33(1):12–14, 2021.

Takeshi Emura and Yi-Hau Chen. *Analysis of survival data with dependent censoring: Copula-Based Approaches*. Springer, 2018.

Takeshi Emura et al. A joint frailty-copula model between tumour progression and death for meta-analysis. *Statistical methods in medical research*, 26(6):2649–2666, 2017.

Gabriel Escarela and Jacques F Carriere. Fitting competing risks with an assumed copula. *Statistical Methods in Medical Research*, 12(4):333–349, 2003.

David Faraggi and Richard Simon. A neural network model for survival data. *Statistics in medicine*, 14(1):73–82, 1995.

Maurice J Frank. On the simultaneous associativity off (x, y) andx+ y- f (x, y). *Aequationes mathematicae*, 19(1): 194–226, 1979.

Dan Geiger et al. d-separation: From theorems to algorithms. In *Machine Intelligence and Pattern Recognition*, volume 10, pages 139–148. Elsevier, 1990.

Thomas A Gerds, Michael W Kattan, Martin Schumacher, and Changhong Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in medicine*, 32(13):2173–2184, 2013.

Frank E Harrell et al. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

Shi Hu et al. Transformer-based deep survival analysis. In *Survival Prediction - Algorithms, Challenges and Applications*, pages 132–148. PMLR, 2021.

Andrew 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*, 124(3):487–495, 2019.

Xiaona 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*, 2019, 2019.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2011.

Ed Kaplan and P Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

Jared L Katzman et al. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):1–12, 2018.

Dong Wook Kim et al. Deep learning-based survival prediction of oral cancer patients. *Scientific reports*, 9(1):1–10, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *NeurIPS*, 30, 2017.

Erwin Kuntz and Hans-Dieter Kuntz. *Hepatology: Textbook and atlas*. Springer Science & Business Media, 2009.

Jennifer C Lai et al. Height contributes to the gender difference in wait-list mortality under the meld-based liver allocation system. *American Journal of Transplantation*, 10(12):2658–2664, 2010.

Changhee Lee et al. Deephit: A deep learning approach to survival analysis with competing risks. In *AAAI*, volume 32, 2018.

Changhee Lee et al. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE TBME*, 67(1): 122–133, 2019.

Catherine R Lesko et al. When to censor? *American journal of epidemiology*, 187(3):623–632, 2018.

Kwan-Moon Leung et al. Censoring issues in survival analysis. *Annual review of public health*, 18(1):83–104, 1997.

Ilya Lipkovich et al. Sensitivity to censored-at-random assumption in the analysis of time-to-event endpoints. *Pharmaceutical statistics*, 15(3):216–229, 2016.

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*, 44(2):167–178, 1997.

Ayse L Mindikoglu et al. Impact of estimated liver volume and liver weight on gender disparity in liver transplantation. *Liver Transplantation*, 19(1):89–95, 2013.

Razieh Nabi et al. Full law identification in graphical models of missing data: Completeness results. In *ICML*, pages 7153–7163. PMLR, 2020.

Chirag Nagpal et al. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE JBHI*, 25(8): 3163–3175, 2021.

Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

David Rindt, Robert Hu, et al. Survival regression with proper scoring rules and monotonic neural networks. In *AISTATS*, pages 1190–1205. PMLR, 2022.

Louis-Paul Rivest and Martin T Wells. A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *Journal of Multivariate Analysis*, 79(1):138–155, 2001.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Maik Schwarz et al. On the identifiability of copulas in bivariate competing risks models. *Canadian Journal of Statistics*, 41(2):291–303, 2013.

Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

Yunlang She et al. Development and validation of a deep learning model for non–small cell lung cancer survival. *JAMA network open*, 3(6):e205842–e205842, 2020.

M Sklar. Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

Weijing Tang et al. Soden: A scalable continuous-time survival model through ordinary differential equation networks. *J. Mach. Learn. Res.*, 23:34–1, 2022.

Anastasios Tsiatis. A nonidentifiability aspect of the problem of competing risks. *PNAS*, 72(1):20–22, 1975.

Anastasios A Tsiatis. Semiparametric theory and missing data. 2006.

Hajime Uno et al. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

Zifeng 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*, pages 1–9, 2022.

Anny Xiang et al. Comparison of the performance of neural network methods and cox regression for censored survival data. *Computational statistics & data analysis*, 34(2):243–257, 2000.

Ming Zheng and John P Klein. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1):127–138, 1995.

Ming Zheng and John 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*, pages 401–408. Springer, 1996.