Survival Analysis via Density Estimation

Hiroki Yanagisawa [1] **Shunta Akiyama** [1]

Abstract

This paper introduces a framework for survival analysis by reinterpreting it as a form of density estimation. Our algorithm post-processes density estimation outputs to derive survival functions, enabling the application of any density estimation model to effectively estimate survival functions. This approach broadens the toolkit for survival analysis and enhances the flexibility and applicability of existing techniques for density estimation. Our framework is versatile enough to handle various survival analysis scenarios, including competing risk models for multiple event types. It can also address dependent censoring when prior knowledge of the dependency between event time and censoring time is available in the form of a copula. In the absence of such information, our framework can estimate the upper and lower bounds of survival functions, accounting for the associated uncertainty.

1. Introduction

Survival analysis constitutes a statistical methodology extensively employed across diverse domains, including medicine, engineering, and social sciences, to analyze and predict the probability distribution of the time until the occurrence of an event of interest. In numerous real-world scenarios, the phenomenon of competing risks emerges when an individual is subject to multiple mutually exclusive events, where the occurrence of one event precludes the observation or alters the probability of the other events. The field of survival analysis with competing risks has a rich historical foundation, as extensively reviewed in survey papers (Wang et al., 2019; Wiegrebe et al., 2024).

Survival analysis with K competing risks can be conceptualized as a variant of density estimation, a subfield of machine learning aimed at estimating the probability distribution of a target variable. Given this conceptual similarity, numerous methodologies originally developed for density estimation have been adapted for survival analysis, particularly in scenarios with $K = 2$ under the conditional independence assumption. For instance, random forests (Breiman, 2001) for density estimation have been adapted into random survival forests (Ishwaran et al., 2008) for survival analysis, and modern neural network models for density estimation have been extended into models such as DeepHit (Lee et al., 2018). Furthermore, strictly proper scoring rules for density estimation (Gneiting & Raftery, 2007) have been adapted for survival analysis (Rindt et al., 2022; Yanagisawa, 2023). Calibration metrics, such as the expected calibration error (Naeini et al., 2015; Guo et al., 2017), have similarly been extended to D-calibration for survival analysis (Haider et al., 2020).

Despite the numerous extensions of density estimation methodologies for survival analysis, these adaptations encounter several limitations. First, these adaptations are typically tailored to specific methodologies on a case-by-case basis, necessitating the development of a new customized extension for each novel density estimation method. Second, most survival analysis models rely on the conditional independence assumption (or even stronger assumptions such as the proportional hazards assumption (Cox, 1972)), which may not hold in various real-world applications. This issue highlights the need for survival models that operate under weaker assumptions. Third, while Tsiatis (1975) demonstrates that the survival function cannot be identified without making any assumption, the estimation of the upper and lower bounds of the survival function has not been fully investigated under the condition where no assumptions can be made. Fourth, while the existence of a strictly proper scoring rule for survival analysis with $K = 2$ under the conditional independence assumption has been established (Rindt et al., 2022) and can be used as an evaluation metric, no strictly proper scoring rule has been established for $K > 2$.

In this paper, we propose a novel framework that reinterprets survival analysis through the lens of density estimation. By post-processing the outputs of density estimation models in the form of cumulative incidence functions, our algorithm derives survival functions, thus enabling any density estimation model to be effectively utilized for survival anal-

[1]CyberAgent, Tokyo, Japan. Correspondence to: Hiroki Yanagisawa <yanagisawa_hiroki@cyberagent.co.jp>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

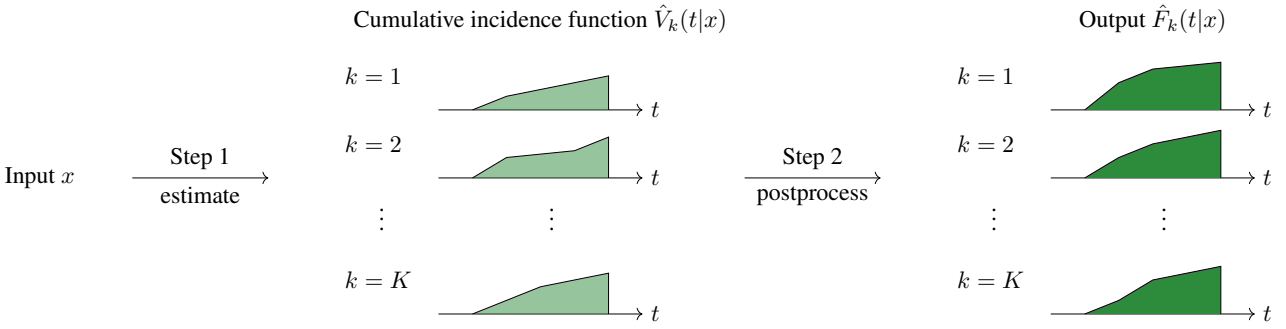


Figure 1. Two-step (TS) algorithm for survival analysis with K competing risks: it first estimates the cumulative incidence functions $\hat{V}_k(t|x)$ via density estimation and then postprocesses them to obtain the outputs $\hat{F}_k(t|x)$.

ysis, as illustrated in Fig. 1. This reinterpretation not only broadens the toolkit available for survival analysis but also enhances the flexibility and applicability of existing density estimation techniques. Our framework's versatility is evident in its capacity to handle a diverse range of survival analysis scenarios, including competing risks and dependent censoring.

The contributions of this paper are summarized as follows:

Model-agnostic framework. Our two-step algorithm can be integrated with any density estimation model, allowing for the selection of a density model based on various criteria such as prediction performance, training time, and interpretability. To demonstrate the effectiveness of our algorithm, we show that our algorithm, when combined with the LightGBM model (Ke et al., 2017), outperforms baseline models on real datasets. Furthermore, we prove that if the density estimation can achieve an arbitrarily small error ϵ (as the number of data points increases), our algorithm can estimate survival functions with a small error for $K = 2$ under plausible assumptions (see Appendices C and D).

Dependent censoring. Our approach addresses the challenge of dependent censoring, a prevalent issue in survival data where the censoring mechanism is not independent of the event time. To handle dependent censoring, a widely used assumption is that prior knowledge regarding the dependency between event time and censoring time is available in the form of a copula (Emura & Chen, 2018). Our algorithm is based on this copula-based assumption and, therefore, relies on weaker assumptions than the conditional independence assumption. Additionally, our two-step algorithm is capable of handling situations where the dependency information is provided as a parameterized copula, along with appropriate assumptions on the form of the survival function to guarantee the identifiability of the parameter.

While our framework is primarily designed to estimate the individual survival function, for the estimation of the average survival function, the Kaplan-Meier estimator (1958) is widely used when the conditional independence assumption is valid. This estimator has been extended as the copula-graphic (CG) estimator (Zheng & Klein, 1995; Carrière, 1995), which operates under the same assumption as ours, that a copula is provided as prior knowledge. Therefore, our two-step algorithm can be seen as an extension of the CG estimators to estimate the individual survival function.

Upper and lower bounds estimation. Even in scenarios where prior knowledge of the dependency is absent, our framework is capable of estimating the upper and lower bounds of individual survival functions, accounting for the uncertainty arising from the lack of knowledge about the copula. Regarding the estimation of the average survival function, Peterson (1976) presents the upper and lower bounds estimation. Our method can be seen as an extension of this approach for estimating individual survival functions.

Strictly proper scoring rule for competing risks. Given a probabilistic output, a strictly proper scoring rule is usually employed as an evaluation metric. Several proper scoring rules exist for survival analysis with $K = 2$ (Rindt et al., 2022; Yanagisawa, 2023), but no scoring rule has been proven to be proper for $K > 2$. In this paper, we introduce a new strictly proper scoring rule, termed NLL-SC, for $K \geq 2$. NLL-SC is based on the copula and can be used as an evaluation metric for survival analysis under dependent censoring, addressing the difficulties discussed by Gharari et al. (2023) in defining an appropriate evaluation metric under dependent censoring. Additionally, by utilizing NLL-SC, we construct a new monotone neural network model for K competing risks based on the copula-based assumption, named the survival copula network (SC-Net), which can be seen as an extension of the monotone neural network model for $K = 2$ under the conditional independence assumption (Rindt et al., 2022).

2. Preliminaries

In this study, we investigate survival analysis with K competing risks. Let X denote the random variable representing a feature vector whose support is $\mathcal{X}$. Let $T_1, T_2, \ldots, T_K$ be the K random variables corresponding to the event times of K distinct types of events, with each T_k supported on $\mathbb{R}_{\geq 0}$. Due to censoring, direct observation of samples $(t_1, t_2, \ldots, t_K) \sim (T_1, T_2, \ldots, T_K)$ is not feasible. However, we can observe their minimum $T = \min_k \{T_k\}_{k=1}^K$ and the corresponding index $\Delta = \arg\min_k \{T_k\}_{k=1}^K$. We note that in much of the existing literature on survival analysis, the index Δ starts at 0 (i.e., $\Delta \in \{0, 1\}$ for $K = 2$); however, in this paper, we assume Δ starts at 1 (i.e., $\Delta \in \{1, 2\}$ for $K = 2$).

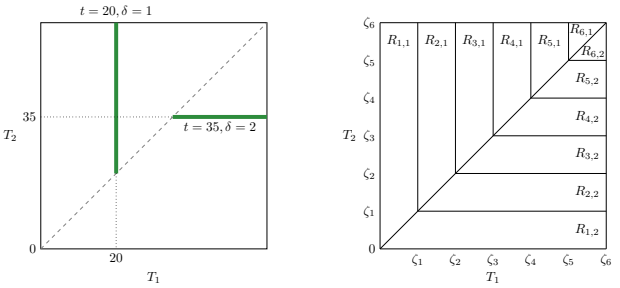
The primary objective of survival analysis is to estimate the cumulative distribution function (CDF) $F_k(t|x) = \Pr(T_k \leq t|x)$ of T_k conditioned on $x \in \mathcal{X}$ for each $k \in [K]$ from samples $(x, t, \delta) \overset{\text{i.i.d.}}{\sim} (X, T, \Delta)$, where $[K] = \{1, 2, \ldots, K\}$. In this study, we frame the problem to estimate $F_k(t|x)$ for each t in a finite set of times $\{\zeta_b\}_{b=0}^B$ such that $0 = \zeta_0 < \zeta_1 < \cdots < \zeta_B$, where ζ_B is sufficiently large to ensure $0 \leq t < \zeta_B$ for any observed time $t \sim T$. We assume that the true $F_k(t|x)$ is a continuous and monotonically increasing function of t, satisfying $F_k(\zeta_0|x) = 0$ and $F_k(\zeta_B|x) = 1$ for all $k \in [K]$ and $x \in \mathcal{X}$. This discretization approach is commonly adopted in numerous survival models (e.g., (Lee et al., 2018; Yanagisawa, 2023; Hickey et al., 2024)).

While survival analysis often focuses on estimating the survival function, defined as $S_k(t|x) = 1 - F_k(t|x)$, this study aims to estimate the CDF $F_k(t|x)$ of T_k. Additionally, we consider the estimation of the average CDF $F_k(t)$ of $F_k(t|x)$ over $x \sim X$ and the average survival function $S_k(t) = 1 - F_k(t)$.

Censored Joint Distribution (CJD) representation. Given an observation $(x, t, \delta) \sim (X, T, \Delta)$, we interpret the pair (t, δ) in a K-dimensional space. For instance, in the case where $K = 2$, the pair (t, δ) can be represented as a line segment in a two-dimensional plane, as illustrated in Fig. 2(a). In this figure, the pair $(t, \delta) = (20, 1)$ observed for $x^{(1)}$ is depicted as a vertical line segment, indicating that Event 1 is observed at time $t_1 = 20$ and $t_2 \geq t_1$. Similarly, the pair $(t, \delta) = (35, 2)$ observed for $x^{(2)}$ is represented as a horizontal line segment, indicating that Event 2 is observed at time $t_2 = 35$ and $t_1 \geq t_2$.

Given that the time horizon is discretized by the boundaries $\{\zeta_b\}_{b=0}^B$, the K-dimensional space is partitioned as illustrated in Fig. 2(b) by defining the set $R_{b,k}$ of observations $(t, \delta) \sim (T, \Delta)$ as follows:

$$R_{b,k} = \{(t, \delta) : \zeta_b < t \leq \zeta_{b+1}, \delta = k\}.$$



(a) Observations in 2D space (b) CJD divided by $R_{b,k}$

Figure 2. (a) Two observations $\{(x^{(1)}, 20, 1), (x^{(2)}, 35, 2)\}$ are illustrated as vertical and horizontal line segments in the two-dimensional space. (b) The CJD space with $B = 6$ and $K = 2$, which is divided into subregions $R_{b,k}$.

In this study, we refer to this partitioned region as the Censored Joint Distribution (CJD) representation.

Copula and survival copula. Many models for survival analysis with $K = 2$ operate under the conditional independence assumption, denoted as $T_1 \perp\!\!\!\perp T_2|X$, or even stronger assumptions such as the proportional hazards assumption, exemplified by the Cox model (1972). In this work, we adopt a more flexible assumption that the dependence structure among $T_1, T_2, \ldots, T_K$ can be modeled using a *copula*, a widely recognized method for capturing dependencies in survival analysis (Emura & Chen, 2018; Gharari et al., 2023; Zhang et al., 2024).

In probability theory and statistics, a copula is defined as a multivariate cumulative distribution function where each univariate marginal distribution is uniformly distributed over the interval $[0, 1]$. Copulas are particularly useful for characterizing the dependence structure among random variables. Formally, a copula is defined as follows (Nelsen, 2006; Gharari et al., 2023).

Definition 2.1. (Copula.) A copula is a K-dimensional function $C : [0, 1]^K \to [0, 1]$ that satisfies the following conditions:

- (i) $C(u_1, u_2, \ldots, u_{k-1}, 0, u_{k+1}, \ldots, u_K) = 0$ for every $u \in [0, 1]^K$,
- (ii) $C(1, \ldots, 1, u, 1, \ldots, 1) = u$ for every $u \in [0, 1]$,
- (iii) Given $u, v \in [0, 1]^K$ such that $u_k < v_k$ is valid for all $k \in [K]$, the following condition is satisfied:

$$\sum_{l \in \{0,1\}^K} (-1)^{l_1 + l_2 + \cdots + l_K} \cdot C(u_1^{l_1} v_1^{1-l_1}, u_2^{l_2} v_2^{1-l_2}, \ldots, u_K^{l_K} v_K^{1-l_K}) \geq 0,$$

where $l = (l_1, l_2, \ldots, l_K)$ is a length-K binary vector.

A notable example of a copula is the *independence copula*, defined as:

$$C_{\text{ind}}(u_1, u_2, \ldots, u_K) = \prod_{k=1}^{K} u_k. \tag{1}$$

Another example is the *Frank copula* for the bivariate case with a non-zero parameter θ:

$$C_{\text{Frank}}(u_1, u_2) = \frac{1}{\theta} \log \left(1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1}\right). \tag{2}$$

A significant attribute of copulas is that the joint distribution $\Pr(T_1 \leq t_1, T_2 \leq t_2, \ldots, T_K \leq t_K)$ can be uniquely represented using a copula, as stated in Sklar's theorem.

Theorem 2.2. *(Sklar's Theorem (1959).) There exists a copula C such that for all $t_1, t_2, \ldots, t_K$,*

$$\begin{aligned}&\Pr(T_1 \leq t_1, T_2 \leq t_2, \ldots, T_K \leq t_K)\\ &\quad = C(F_1(t_1), F_2(t_2), \ldots, F_K(t_K)).\end{aligned}$$

If the marginal distribution F_k is continuous for all k, then C is unique.

Copulas are instrumental in computing joint probabilities. For instance, the joint probability $\Pr(\zeta_1 < T_1 \leq \zeta_4, T_2 \leq \zeta_3)$ can be determined using a copula as follows:

$$\begin{aligned}&\Pr(\zeta_1 < T_1 \leq \zeta_4, T_2 \leq \zeta_3)\\ &\quad = C(F_1(\zeta_4), F_2(\zeta_3)) - C(F_1(\zeta_1), F_2(\zeta_3)),\end{aligned} \tag{3}$$

where $F_k(t) = \Pr(T_k \leq t)$.

In the context of survival analysis, a *survival copula* $\overline{C}$ is frequently employed, which satisfies the following equation:

$$\begin{aligned}&\Pr(T_1 > t_1, T_2 > t_2, \ldots, T_K > t_K)\\ &\quad = \overline{C}(1 - F_1(t_1), 1 - F_2(t_2), \ldots, 1 - F_K(t_K)).\end{aligned}$$

It is well-established that any survival copula $\overline{C}$ can be represented using its corresponding copula C (see, e.g., (Georges et al., 2001)). For instance, in the case where $K = 2$, the survival copula $\overline{C}$ can be expressed as:

$$\overline{C}(u_1, u_2) = u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2).$$

Additionally, it is important to note that if $C = C_{\text{ind}}$ (defined in (1)), then $\overline{C} = C_{\text{ind}}$.

3. Two-Step Algorithm

We introduce a two-step algorithm designed to estimate the CDF $F_k(\zeta_b|x)$ for survival analysis with K competing risks. As illustrated in Algorithm 1, the initial step involves estimating $V_k(t|x)$, the conditional k-th cumulative incidence function (CIF), which is formally defined as follows:

$$V_k(t|x) = \Pr(T \leq t, \Delta = k|x). \tag{4}$$

Subsequently, utilizing the estimated $\hat{V}_k(\zeta|x)$, we estimate the probability

$$r_{b,k|x} = \Pr((t, \delta) \in R_{b,k}|x)$$

within the context of the CJD representation via the following equation:

$$\hat{r}_{b,k|x} = \hat{V}_k(\zeta_b|x) - \hat{V}_k(\zeta_{b-1}|x). \tag{5}$$

Thereafter, the output distribution $\hat{F}_k(\zeta_b|x)$ is derived from the estimation $\hat{r}_{b,k|x}$, assuming that we have prior knowledge regarding the dependencies among the random variables $T_1, T_2, \ldots, T_K$ in the form of a copula C. It is noteworthy that the conditional independence assumption $(T_1 \perp\!\!\!\perp T_2|X)$ used in many survival models for $K = 2$ is equivalent to using the independence copula (1) for the copula C.

3.1. Step 1: Cumulative Incidence Function Estimation

A straightforward methodology for estimating the CIF (4) is the application of a distribution regression model, which is specifically engineered to estimate a conditional CDF $F(y|x) = \Pr(Y \leq y|x)$ given samples $(x, y) \overset{\text{i.i.d.}}{\sim} (X, Y)$, where the random variables X and Y represent feature vectors and target values from $\mathbb{R}$, respectively. Exemplary distribution regression models include those based on monotone neural networks (Chilinski & Silva, 2020) and random forests (Schlosser et al., 2019; Hothorn & Zeileis, 2021; Ćevid et al., 2022), and NGBoost (Duan et al., 2020), which is founded on gradient boosting.

An alternative approach involves directly estimating $\hat{r}_{b,k|x}$ (defined in (5)) through the utilization of a density estimation model. Here, density estimation refers to estimating $\Pr(Y = y|x)$ from samples $(x, y) \overset{\text{i.i.d.}}{\sim} (X, Y)$, where the random variables X and Y represent feature vectors and target values from a *discrete* set, respectively. Most multiclass classification models can be used for density estimation, including random forests (Breiman, 2001), gradient boosting (e.g., LightGBM (Ke et al., 2017)), and neural networks. Recent advancements in density estimation techniques are found in (Filho et al., 2023).

While a broader array of models for density estimation exists compared to distribution regression models, an advantage of employing a distribution regression model is that, if we wish to adjust the hyperparameter B, there is no need to retrain the predictive model to estimate $\hat{V}_k(t|x)$; we only need to recompute Eq. (5) to obtain $\hat{r}_{b,k|x}$.

3.2. Step 2: Postprocessing

The second step of our algorithm involves computing $\hat{F}_k(\zeta_b|x)$ utilizing the estimates $\hat{r}_{b,k|x}$ obtained in the first

step and a specified copula C. Let $\mathbf{r}_{b|x} \in [0,1]^K$ denote the vector of length K whose k-th element is $r_{b,k|x}$, and let $\mathbf{F}_{b|x} \in \mathbb{R}^K$ denote the vector of length K whose k-th element is $F_k(\zeta_b|x)$.

We begin by representing $\mathbf{r}_{b|x}$ as a function of $\mathbf{F}_{b-1|x}$, $\mathbf{F}_{b|x}$, and the copula C. For simplicity, we consider the case where $K = 2$ in this section, with generalizations for $K > 2$ detailed in Appendix B.1. By the definition of $r_{b,k|x}$, we have the following representations:

$$\begin{aligned} r_{b,1|x} &= \Pr(\zeta_{b-1} < T_1 \le \zeta_b, T_1 \le T_2|x) \\ &= q_{\{1\},b|x} - w_1\, q_{\{1,2\},b|x}, \end{aligned} \tag{6}$$

$$\begin{aligned} r_{b,2|x} &= \Pr(\zeta_{b-1} < T_2 \le \zeta_b, T_2 \le T_1|x) \\ &= q_{\{2\},b|x} - w_2\, q_{\{1,2\},b|x}, \end{aligned} \tag{7}$$

where

$$q_{\{1\},b|x} = \Pr(\zeta_{b-1} < T_1 \le \zeta_b, \zeta_{b-1} \le T_2|x), \tag{8}$$

$$q_{\{2\},b|x} = \Pr(\zeta_{b-1} < T_2 \le \zeta_b, \zeta_{b-1} \le T_1|x),$$

$$q_{\{1,2\},b|x} = \Pr(\zeta_{b-1} < T_1 \le \zeta_b, \zeta_{b-1} < T_2 \le \zeta_b|x), \tag{9}$$

and $w_1, w_2 \ge 0$ are unknown weight parameters such that $w_1 + w_2 = 1$. See Fig. 3 for an illustration of quantities (8) and (9) with $b = 2$. Unless otherwise stated, we set $w_1 = w_2 = 1/2$. Note that, if we use a sufficiently large B, the correction term (9) should be small, and therefore the choices of the weight parameters w_1 and w_2 should have minimal impact on these equations. Then, recalling that any joint distribution can be computed using a copula (as demonstrated in equation (3)), it is straightforward to represent (8)–(9) using $F_1(\zeta_{b-1}|x)$, $F_1(\zeta_b|x)$, $F_2(\zeta_{b-1}|x)$, $F_2(\zeta_b|x)$, and the copula C:

$$\begin{aligned} q_{\{1\},b|x} &= C(F_1(\zeta_b|x), 1) - C(F_1(\zeta_{b-1}|x), 1) \\ &\quad - C(F_1(\zeta_b|x), F_2(\zeta_{b-1}|x)) \\ &\quad + C(F_1(\zeta_{b-1}|x), F_2(\zeta_{b-1}|x)), \end{aligned} \tag{10}$$

$$\begin{aligned} q_{\{2\},b|x} &= C(1, F_2(\zeta_b|x)) - C(1, F_2(\zeta_{b-1}|x)) \\ &\quad - C(F_1(\zeta_{b-1}|x), F_2(\zeta_b|x)) \\ &\quad + C(F_1(\zeta_{b-1}|x), F_2(\zeta_{b-1}|x)), \end{aligned} \tag{11}$$

$$\begin{aligned} q_{\{1,2\},b|x} &= C(F_1(\zeta_b|x), F_2(\zeta_b|x)) \\ &\quad - C(F_1(\zeta_b|x), F_2(\zeta_{b-1}|x)) \\ &\quad - C(F_1(\zeta_{b-1}|x), F_2(\zeta_b|x)) \\ &\quad + C(F_1(\zeta_{b-1}|x), F_2(\zeta_{b-1}|x)). \end{aligned} \tag{12}$$

Combining (6)–(12), we can represent $r_{b,1|x}$ and $r_{b,2|x}$ using $F_1(\zeta_{b-1}|x)$, $F_1(\zeta_b|x)$, $F_2(\zeta_{b-1}|x)$, $F_2(\zeta_b|x)$, and the copula C. This means that we can write this relationship as

$$\mathbf{r}_{b|x} = g_C(\mathbf{F}_{b|x}|\mathbf{F}_{b-1|x}) \tag{13}$$

by using a function g_C that depends on C.

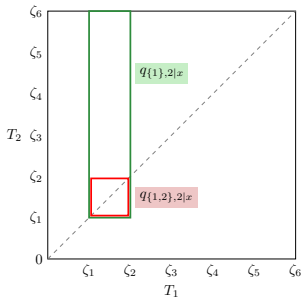


Figure 3. Illustration of $q_{\{1\},2|x}$ and $q_{\{1,2\},2|x}$.

Algorithm 1 Two-Step (TS) Algorithm

1: Estimate $\hat{V}_k(t|x)$
2: Let $\hat{r}_{b,k|x} = \hat{V}_k(\zeta_b|x) - \hat{V}_k(\zeta_{b-1}|x)$
3: Let $\hat{\mathbf{F}}_{0|x} = \mathbf{0}$
4: **for** $b = 1, 2, \ldots, B-1$ **do**
5: Calculate $\hat{\mathbf{F}}_{b|x}$ by solving Eq. (13)
6: **end for**
7: **return** $\{\hat{\mathbf{F}}_{b|x}\}_{b=1}^{B-1}$

Having established (13), we can obtain the estimation $\hat{\mathbf{F}}_{b|x}$ of $\mathbf{F}_{b|x}$ for all b by solving this equation based on the estimation $\hat{\mathbf{r}}_{b|x}$ of $\mathbf{r}_{b|x}$, as outlined in Steps 3–7 of Algorithm 1. We leverage the initial condition $\hat{\mathbf{F}}_{0|x} = \mathbf{F}_{0|x} = \mathbf{0}$ for all x, where $\mathbf{0}$ is the K-dimensional vector of zeros. At the initial step for $b = 1$, we obtain $\hat{\mathbf{F}}_{b|x}$ by solving (13). This equation is solvable because it contains K equality constraints and the unknown value is the length-K vector $\mathbf{F}_{b|x}$. We repeat this procedure for $b = 2, 3, \ldots, B-1$ to obtain $\hat{\mathbf{F}}_{b|x}$ for all b. See Appendix B.2 for more details of the algorithm.

Note that this second step of our algorithm is similar to the algorithm based on a bisection root-finding algorithm (Zheng & Klein, 1995), but their algorithm is valid only for $K = 2$ and its extension for $K > 2$ is unknown. In contrast, our algorithm is extendable for $K > 2$ as shown in Appendix B.1. We also note that as $B \to \infty$, Carrière (1995) shows another method to estimate $\hat{\mathbf{F}}_{b|x}$ by solving an equation similar to (13). We discuss this method further in Appendix B.3.

Simplified implementation of our algorithm. In our implementation, we adopted a simpler approach to solve Eq. (13) for all b simultaneously, rather than sequentially solving Eq. (13) for each b as outlined in Lines 3–7 of Algorithm 1. Specifically, by utilizing an automatic differentiation library for Python (e.g., PyTorch and Tensorflow), we estimate $\hat{\mathbf{F}}_{b|x}$ by minimizing the following objective

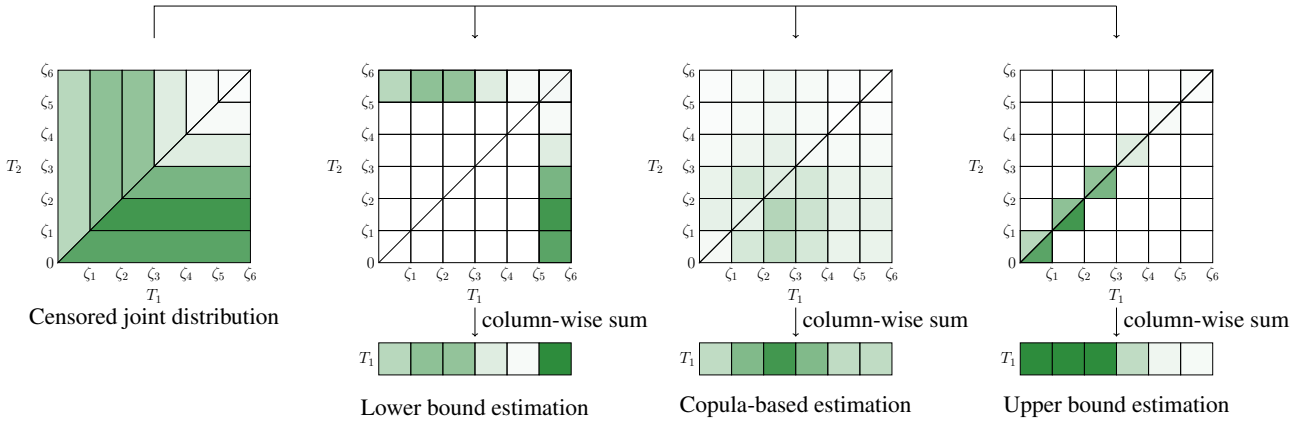


Figure 4. Illustration of the upper and lower bounds estimation. Here, region $R_{b,k}$ is divided into grids, with denser color indicating a higher probability that a data point is contained in the corresponding region. The lower bound estimation is achieved by assigning the probability mass $\hat{r}_{b,k}$ to the last time slot within region $R_{b,k}$ and then calculating the column-wise sum. Conversely, the upper bound estimation is obtained by assigning the probability mass $\hat{r}_{b,k}$ to the earliest time slot within region $R_{b,k}$ and calculating the column-wise sum.

function:

$$\sum_{b=1}^{B-1} \left(g_C(\hat{\mathbf{F}}_{b|x} | \hat{\mathbf{F}}_{b-1|x}) - \hat{\mathbf{r}}_{b|x} \right)^2 \tag{14}$$

for all b simultaneously. Note that Eq. (14) is minimized if Eq. (13) holds for all b.

3.3. Non-Identifiablity

A limitation of our two-step algorithm is the necessity for prior knowledge of the copula C, which may not be readily available in practical scenarios. One potential approach to address this issue is to employ a parameterized copula C_θ, where θ represents some parameter (e.g., the Frank copula (2) with parameter θ), and to extend the survival analysis framework to estimate the copula parameter θ in addition to the CDFs $F_k(t|x)$. However, Tsiatis (1975) demonstrated that this approach is not feasible because it is impossible to identify $F_k(t|x)$ without making some assumptions about $F_k(t|x)$.

Consequently, many researchers have explored the introduction of additional assumptions on the distribution $F_k(t|x)$ to ensure the identifiability of both the copula parameter θ and $F_k(t|x)$. For instance, Gharari et al. (2023) assume that $F_k(t|x)$ follows a Weibull distribution and present an algorithm to estimate the copula parameter θ. Other examples include the study of copula identifiability under the proportional hazards assumption in (Heckman & Honoré, 1989; Deresa & Keilegom, 2024), and the investigation of copula identifiability for other restricted classes of distributions $F_k(t|x)$ in (Czado & Keilegom, 2022; Wang, 2023; Zhang et al., 2024). Under the strongest assumption that the true distribution $F_k(t|x)$ is completely known, Schwarz et al. (2013) discuss the identifiability of Archimedean copulas and the non-identifiability of symmetric copulas.

It is important to note that our two-step algorithm is sufficiently flexible to incorporate these identifiability results. Suppose that the distribution $F_{k,\eta}(t|x)$ and the copula C_θ are parameterized by η and θ, respectively. If the identifiability of the parameters η and θ is established, then the objective function (14) should have a unique solution, where $\hat{\mathbf{F}}_{b|x}$ and C are replaced with the parameterized $\hat{\mathbf{F}}_{b,\eta|x}$ and C_θ, respectively.

4. Upper and Lower Bounds Estimation

As we discuss in Appendix A, there exist scenarios where it is infeasible to verify dependency in any manner and we cannot have any copula as a prior knowledge on the dependency. Even in such instances, our algorithm can still be employed to estimate the upper and lower bounds of $F_k(t|x)$. By definition, the following inequalities can be readily verified:

$$F_k(\zeta_b|x) \geq \Pr\left((t,\delta) \in \bigcup_{b' \leq b} R_{b',k} \middle| x \right)$$
$$F_k(\zeta_b|x) \leq \Pr\left((t,\delta) \in \bigcup_{b' \leq b, k' \in [K]} R_{b',k'} \middle| x \right),$$

which are equivalent to

$$\sum_{b' \leq b} r_{b',k|x} \leq F_k(\zeta_b|x) \leq \sum_{b' \leq b, k' \in [K]} r_{b',k'|x}. \tag{15}$$

Based on these inequalities, we estimate the upper and lower bounds of $F_k(\zeta_b|x)$ by substituting $r_{b,k|x}$ with the output $\hat{r}_{b,k|x}$ from Step 1 of our algorithm. It is important to note that these inequalities are derived without utilizing the parameters w_1 and w_2 in our two-step algorithm. As illustrated in Fig. 4, the estimation of upper and lower bounds and the second step of our algorithm to estimate $F_k(\zeta_b|x)$ can be interpreted as redistributing the probability mass within the CJD representation into fine-grained grid cells.

While these bounds for $F_k(t|x)$ are computed only for discrete values $t \in \{\zeta_b\}_{b=1}^B$, it is possible to compute these bounds for any t using a distribution regression model to estimate the CIF $V_k(t|x)$ (as defined in (4)). This approach is explained in Appendix E. These bounds can be viewed as variants of the bounds established for the average survival function in (Peterson, 1976).

It is crucial to distinguish that our upper and lower bounds differ from the concept of a *confidence interval*, which quantifies the epistemic uncertainty inherent in the prediction model (Bengs et al., 2022). Our bounds quantify uncertainty arising from the absence of prior knowledge about the true copula C. As we discuss in Appendix E, these two types of bounds can be combined to account for both sources of uncertainty.

5. Strictly Proper Scoring Rule for Competing Risks

Strictly proper scoring rules hold significant importance in the domain of statistics, particularly for the evaluation of probabilistic estimates (Gneiting & Raftery, 2007). These scoring rules ensure that the expected score is minimized when the estimated probabilities accurately reflect the true distribution of outcomes.

In this section, we establish the existence of a strictly proper scoring rule for any number of competing risks, K, given that the dependency structure is defined by a survival copula. While Rindt et al. (2022) demonstrated the existence of a strictly proper scoring rule for $K = 2$ under the assumption of conditional independence, no such rule has been established for $K > 2$ or for cases where the conditional independence assumption does not hold.

We begin by defining proper and strictly proper scoring rules as follows:

Definition 5.1. (Proper and Strictly Proper Scoring Rules.) A scoring rule $\mathcal{S}$ for the estimation $\hat{F}_k(t|x)$ of $F_k(t|x)$ is *proper* if the following inequality is satisfied:

$$\mathbb{E}[\mathcal{S}(\{\hat{F}_k(t|x)\}_{k=1}^K, (t,\delta))] \geq \mathbb{E}[\mathcal{S}(\{F_k(t|x)\}_{k=1}^K, (t,\delta))]. \tag{16}$$

A scoring rule is *strictly proper* if equality in (16) holds if and only if $\hat{F}_k(t|x)$ is equal to $F_k(t|x)$ for all k and t.

Subsequently, we demonstrate the existence of a strictly proper scoring rule for any K. According to Tsiatis (1975), the derivative $v_k(t|x)$ of the cumulative incidence function (CIF) $V_k(t|x)$ (as defined in Eq. (4)) can be represented using $F_k(t|x)$ and a survival copula $\overline{C}$:

$$\begin{aligned} v_k(t|x) &= \frac{\mathrm{d}}{\mathrm{d}t} V_k(t|x) \\ &= -\frac{\partial}{\partial t_k} \overline{C}(1 - F_1(t_1|x), \\ &\qquad 1 - F_2(t_2|x), \\ &\qquad \dots, \\ &\qquad 1 - F_K(t_K|x)) \Big|_{t_1=t_2=\dots=t_K=t}. \end{aligned}$$

By using this equation, we propose a scoring rule NLL-SC, which stands for Negative Log-Likelihood based on Survival Copula.

Assumption 5.2. The Kullback-Leibler (KL) divergence between $v_k(t|x)$ and its estimate $\hat{v}_k(t|x)$ satisfies $D_{\mathrm{KL}}(v_x||\hat{v}_x) < \infty$ for all k, where v_x and $\hat{v}_x$ denote the probability distribution over (k,t) of $v_k(t|x)$ and $\hat{v}_k(t|x)$, respectively.

Theorem 5.3. *(A Strictly Proper Scoring Rule for Competing Risks.) If Assumption 5.2 holds, the following scoring rule $\mathcal{S}_{\mathrm{NLL-SC}}$ is strictly proper:*

$$\mathcal{S}_{\mathrm{NLL-SC}}(\{\hat{F}_k(t|x)\}_{k=1}^K, (t,\delta)) = -\mathbb{1}_{\delta=k} \log \hat{v}_k(t|x).$$

Proof. Since Assumption 5.2 ensures the existence of the KL divergence, we have

$$\begin{aligned} &\mathbb{E}[\mathcal{S}_{\mathrm{NLL-SC}}(\{\hat{F}_k(t|x)\}_{k=1}^K, (t,\delta))] \\ &- \mathbb{E}[\mathcal{S}_{\mathrm{NLL-SC}}(\{F_k(t|x)\}_{k=1}^K, (t,\delta))] \\ &\quad = \sum_{k=1}^K \int_0^\infty v_k(t|x)(\log v_k(t|x) - \log \hat{v}_k(t|x))\mathrm{d}t \\ &\quad = D_{\mathrm{KL}}(v_x||\hat{v}_x) \\ &\quad \geq 0. \end{aligned}$$

Equality in the last inequality holds if and only if $v_k(t|x) = \hat{v}_k(t|x)$ holds for all k and t, which is equivalent to that $F_k(t|x) = \hat{F}_k(t|x)$ holds for all k and t. Hence the scoring rule $\mathcal{S}_{\mathrm{NLL-SC}}$ is strictly proper. □

Neural network model based on NLL-SC. Utilizing the scoring rule NLL-SC, we propose a new monotone neural network model, the survival copula network (SC-Net). Similar to the monotone neural network model for $K = 2$ proposed by Rindt et al. (2022), our SC-Net employs a monotone neural network to represent a CDF (Chilinski & Silva, 2020), and we use the NLL-SC as its loss function.

Table 1. Evaluation metrics

Task	Metric Name	Assumption	Metric Type
Density estimation	CJD-Brier	-	Discrimination
	CJD-Logarithmic	-	Discrimination
	CJD-KS	-	Calibration
Survival analysis	NLL-SC	Copula	Discrimination
	Cen-log	Independence	Discrimination
	D-calibration	Independence	Calibration
	KM-calibration	Independence	Calibration

We utilized the smooth min-max neural network (Igel, 2024) as the monotone neural network, though other monotone neural networks, such as those proposed in (Yanagisawa et al., 2022; Kim & Lee, 2024), could also be used. Note that our SC-Net with $K = 2$ can be viewed as incorporating prior knowledge of the ground truth copula into the DC-Survival model (Zhang et al., 2024), although the primary objective of DCSurvival appears to be the identification of parameters within an Archimedean copula.

6. Experiments

We conducted a series of experimental evaluations to assess the performance of our proposed two-step algorithm. We conducted the experimental procedures on a virtual machine possessing a single CPU devoid of any GPU, equipped with a memory of 64 GB, and operating on CentOS Stream 9. The software implementation was achieved using Python 3.11.6 and PyTorch 2.1.2. The datasets employed for this purpose were the Dialysis and oldmort datasets, sourced from the Python package `SurvSet` (Drysdale, 2022).

Models. In the first step of our algorithm, we utilized five different models. Specifically, the TS-Brier and TS-Log models employed neural networks for density estimation, utilizing the Brier and Logarithmic scores (Gneiting & Raftery, 2007) as their respective loss functions. The TS-LGB, TS-RF, and TS-DRF models utilized the LightGBM (LGB) model (Ke et al., 2017), the random forest (RF) model available in the Python package `sklearn`, and the distribution regression forest (DRF) (Ćevid et al., 2022), respectively. The prefix TS denotes the Two-Step algorithm, with each model implementing a distinct first-step model but sharing a common second-step algorithm (specifically, the simplified implementation as detailed in Sec. 3.2). The hyperparameter was set to $B = 32$. Additionally, we employed our SC-Net model proposed in Sec. 5.

For comparative purposes, we included the Cox model (Cox, 1972), the random survival forest (RSF) (Ishwaran et al., 2008), and the DeepHit model (Lee et al., 2018), a neural network model. In the DeepHit model, we set the parameter $\alpha = 0$ to ensure its loss function was a proper scoring rule as recommended by Yanagisawa (2023).

Evaluation metrics. We used three metrics to evaluate the estimates of the CJD representation $\hat{r}_{b,k|x}$ and four metrics to assess the estimated distribution $\hat{F}_k(t|x)$ as summarized in Table 1. For the CJD representation, we used the Brier and Logarithmic scores (Gneiting & Raftery, 2007) as discrimination metrics. Additionally, we applied the sum of the Kolmogorov-Smirnov calibration error (Gupta et al., 2021) as a calibration metric. This metric, based on the Kolmogorov-Smirnov test (Kolmogorov, 1933; Smirnov, 1939), is defined as follows:

$$\sum_{k=1}^{K} \max_{0\le\sigma\le 1} \left| h_{k,\sigma} - \tilde{h}_{k,\sigma} \right|,$$

where

$$h_{k,\sigma} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\hat{f}_k(x^{(i)})\le\sigma} \cdot \mathbb{1}_{y^{(i)}=k}$$

and

$$\tilde{h}_{k,\sigma} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\hat{f}_k(x^{(i)})\le\sigma} \cdot \hat{f}_k(x^{(i)}).$$

Here, $\hat{f}_k(x^{(i)})$ represents the probability of the feature vector $x^{(i)}$ being classified in class k, and each $\hat{f}_k(x)$ is equal to its corresponding $\hat{r}_{b,k|x}$. For models that output only the distribution $\hat{F}_k(t|x)$, the CJD representation $\hat{r}_{b,k|x}$ was estimated using Eq. (13) with the output and the independence copula C_{ind} (as defined in (1)).

For the estimated distribution $\hat{F}_k(t|x)$, we utilized the simplified variant of the censored logarithmic score (referred to as *cen-log*) (Yanagisawa, 2023) and our strictly proper scoring rule, NLL-SC, as evaluation metrics. Additionally, D-calibration (Haider et al., 2020) and KM-calibration (Yanagisawa, 2023) were employed as calibration metrics.

Results. Figure 5 illustrates the results for the Dialysis and oldmort datasets. The results indicate that while the DeepHit model demonstrated superior performance compared to the Cox and RSF models, our TS-LGB model

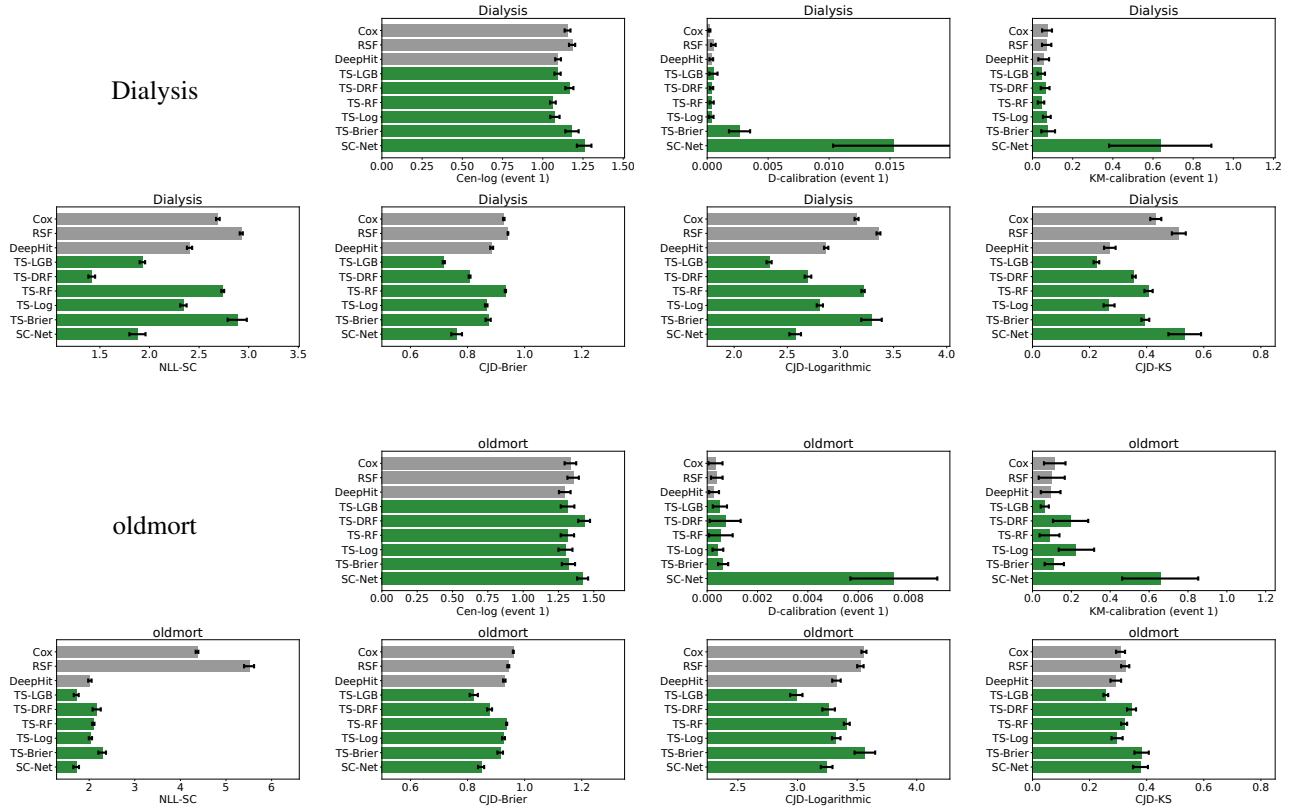


Figure 5. Performance comparison on the Dialysis and oldmort datasets with $B = 32$ (lower is better). While DeepHit performs the best among the baseline models, our TS-LGB model shows comparable or better performances than DeepHit.

exhibited comparable or superior performance relative to DeepHit. It is also noteworthy that neural network-based models, including DeepHit, typically require longer training times compared to tree-based models such as RSF and TS-LGB. Consequently, the TS-LGB model emerged as the most effective model for the Dialysis and oldmort datasets. Additional evaluation results using other datasets for $K = 2$ and scenarios involving competing risks with $K = 3$, in comparison with models such as Deep Survival Machines (DSM) (Nagpal et al., 2021), DeSurv (Danks & Yau, 2022), and Neural Fine-Gray (NeuralFG) (Jeanselme et al., 2023), are detailed in Appendix F. Furthermore, the appendix includes additional evaluation results encompassing the estimation of upper and lower bounds, as well as an ablation study on the hyperparameter B.

Source Codes. The implementations of our models are accessible at **https://github.com/CyberAgentAILab/cenreg**.

7. Conclusion

We demonstrated a reduction from survival analysis to density estimation, which allows the application of any density estimation model to survival analysis. This algorithm operates under the assumption of having prior knowledge of the copula C, which is a weaker assumption compared to the conditional independence assumption. This approach is consistent with Tsiatis's non-identifiability result (1975), and we have shown that our algorithm can also be utilized to estimate the upper and lower bounds of individual survival functions.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Bengs, V., Hüllermeier, E., and Waegeman, W. Pitfalls of epistemic uncertainty quantification through loss minimisation. In *Advances in Neural Information Processing Systems*, pp. 29205–29216, 2022.

Bilodeau, B., Foster, D. J., and Roy, D. M. Minimax rates for conditional density estimation via empirical entropy. *The Annals of Statistics*, 51(2):762–790, 2023.

Breiman, L. Random forests. *Machine Learning*, 45:5–32, 2001.

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

Ćevid, D., Michel, L., Näf, J., Bühlmann, P., and Meinshausen, N. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23(333):1–79, 2022.

Chilinski, P. and Silva, R. Neural likelihoods via cumulative distribution functions. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 420–429, 2020.

Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

Czado, C. and Keilegom, I. V. Dependent censoring based on parametric copulas. *Biometrika*, 110(3):721–738, 2022.

Danks, D. and Yau, C. Derivative-based neural modelling of cumulative distribution functions for survival analysis. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 7240–7256, 2022.

Defazio, A., Yang, X. A., Mehta, H., Mishchenko, K., Khaled, A., and Cutkosky, A. The road less scheduled. In *Advances in Neural Information Processing Systems*, 2024.

Deresa, N. W. and Keilegom, I. V. Copula based Cox proportional hazards models for dependent censoring. *Journal of the American Statistical Association*, 119(546):1044–1054, 2024.

Drysdale, E. SurvSet: An open-source time-to-event dataset repository. Technical report, arXiv:2203.03094, 2022.

Duan, T., Avati, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A., and Schuler, A. NGBoost: Natural gradient boosting for probabilistic prediction. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 2690–2700, 2020.

Emura, T. and Chen, Y.-H. *Analysis of Survival Data with Dependent Censoring: Copula-Based Approaches*. Springer, 2018.

Filho, T. S., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., and Flach, P. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112:3211–3260, 2023.

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

Georges, P., Lamy, A.-G., Nicolas, E., Quibel, G., and Roncalli, T. Multivariate survival modelling: A unified approach with copulas. Technical report, SSRN, 2001.

Gharari, A. H. F., Cooper, M., Greiner, R., and Krishnan, R. G. Copula-based deep survival models for dependent censoring. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 669–680, 2023.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Greenwood, M. A report on the natural duration of cancer. In *Reports on Public Health and Medical Subjects. Ministry of Health*. London: H.M.S.O., 1926.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017.

Gupta, K., Rahimi, A., Ajanthan, T., Mensink, T., Sminchisescu, C., and Hartley, R. Calibration of neural networks using splines. In *ICLR 2021*, 2021.

Haider, H., Hoehn, B., Davis, S., and Greiner, R. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 21(1), 2020.

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

Hickey, J., Henao, R., Wojdyla, D., Pencina, M., and Engelhard, M. Adaptive discretization for event prediction (ADEPT). In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pp. 1351–1359, 2024.

Hothorn, T. and Zeileis, A. Predictive distribution modeling using transformation forests. *Journal of Computational and Graphical Statistics*, 30(4):1181–1196, 2021.

Igel, C. Smooth min-max monotonic networks. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 20908–20923, 2024.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

Jeanselme, V., Yoon, C. H., Tom, B., and Barrett, J. Neural Fine-Gray: Monotonic neural networks for competing risks. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 379–392, 2023.

Kannel, W. B. and McGee, D. L. Diabetes and cardiovascular disease: the Framingham study. *JAMA*, 241(19): 2035–2038, 1979.

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

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 3149–3157, 2017.

Kim, H. and Lee, J.-S. Scalable monotonic neural networks. In *The Twelfth International Conference on Learning Representations*, 2024.

Kolmogorov, A. Sulla determinazione empírica di uma legge di distribuzione, 1933.

Lee, C., Zame, W., Yoon, J., and van der Schaar, M. DeepHit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using Bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), 2015.

Nagpal, C., Li, X., and Dubrawski, A. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE Journal of Biomedical and Health Informatics*, 25 (8):3163–3175, 2021.

Nelsen, R. B. *An Introduction to Copulas*. Springer New York, NY, 2006.

Niles-Weed, J. and Berthet, Q. Minimax estimation of smooth densities in Wasserstein distance. *The Annals of Statistics*, 50(3):1519–1540, 2022.

Peterson, A. V. Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks. *Proceedings of the National Academy of Sciences*, 73(1):11–13, 1976.

Rindt, D., Hu, R., Steinsaltz, D., and Sejdinovic, D. Survival regression with proper scoring rules and monotonic neural networks. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 1190–1205, 2022.

Sart, M. Estimating the conditional density by histogram type estimators and model selection. *ESAIM: Probability and Statistics*, 21:34–55, 2017.

Schlosser, L., Hothorn, T., Stauffer, R., and Zeileis, A. Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *The Annals of Applied Statistics*, 13(3):1564–1589, 2019.

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

Sklar, A. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris*, 8:229–231, 1959.

Smirnov, N. On the estimation of the discrepancy between empirical curves of distribution for two independent samples, 1939.

Therneau, T. M. and Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model*. Springer Science & Business Media, 2000.

Tsiatis, A. A nonidentifiability aspect of the problem of competing risks. *Proc. Natl. Acad. Sci. USA*, 72(1):20–22, 1975.

Vallender, S. S. Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974.

Wang, A. The identifiability of copula models for dependent competing risks data with exponentially distributed margins. *Statistica Sinica*, 33(2):983–1001, 2023.

Wang, P., Li, Y., and Reddy, C. K. Machine learning for survival analysis: A survey. *ACM Comput. Surv.*, 51(6), 2019.

Wiegrebe, S., Kopper, P., Sonabend, R., Bischl, B., and Bender, A. Deep learning for survival analysis: A review. *Artificial Intelligence Review*, 57, 2024.

Yanagisawa, H. Proper scoring rules for survival analysis. In *ICML 2023*, 2023.

Yanagisawa, H., Miyaguchi, K., and Katsuki, T. Hierarchical lattice layer for partially monotone neural networks. In *Advances in Neural Information Processing Systems*, pp. 11092–11103, 2022.

Zhang, W., Ling, C. K., and Zhang, X. Deep copula-based survival analysis for dependent censoring with identifiability guarantees. In *AAAI 2024*, 2024.

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

A. On Applications of Proposed Methods

Survival analysis is a crucial tool in various fields such as medicine, engineering, and social sciences, where the time until an event of interest occurs is studied. The applicability of our proposed methods can be classified into three distinct types based on the nature of the dependency between the event time and the censoring time.

- (Conditional Independence.) The first class of applications encompasses scenarios where the conditional independence assumption between the event time and the censoring time is valid or highly likely to hold. This situation typically arises in cases of administrative censoring, where data points are censored solely due to the limited window of observation times. For instance, in clinical trials, patients might be censored at the end of the study period regardless of whether the event of interest has occurred. In such cases, the conditional independence assumption simplifies the analysis, allowing the use of survival models based on the conditional independence assumption including our two-step algorithm.
- (Verifiable Dependency.) The second class of applications includes scenarios where there is a dependency between the event time and the censoring time, but this dependency can be verified, albeit at a significant cost, for a small subset of data points. An example of this situation is found in medical studies where the primary event of interest is patient mortality, and censoring occurs when patients are discharged from the hospital. In such cases, it might be feasible to investigate the true event time for a small fraction of patients, thereby assessing the dependency between the event and censoring times as a form of copula. This estimated copula can then be used to model the dependency in the entire dataset, allowing for more accurate estimation of survival functions.
- (Unverifiable Dependency.) The third class of applications involves situations where the dependency between the event time and the censoring time cannot be determined, even with extensive resources. In such scenarios, identifying the survival function is inherently challenging due to the unknown nature of the dependency. We note that, even without any knowledge of the dependency, our method can estimate the upper and lower bounds to account for the uncertainty in the dependency structure. Additionally, if we have some confidence that the dependency can be represented by a parameterized copula with some range of the parameters (e.g., the Frank copula with parameters $-5 \leq \theta \leq 5$), we can also estimate the survival functions using the copula information to narrow down the bounds.

B. Step 2 of Proposed Algorithm

B.1. Generalization for competing risks

We generalize the second step of our algorithm presented in Sec. 3 for $K > 2$. For notational simplicity, we omit x in this section.

For a subset $\mathbf{I} \subseteq [K]$, let $Q_{\mathbf{I},b} = \{(t_1, t_2, \ldots, t_K) : \wedge_{k' \in \mathbf{I}}(\zeta_{b-1} < t_{k'} \leq \zeta_b) \text{ and } \wedge_{k' \notin \mathbf{I}} (\zeta_{b-1} \leq t_{k'})\}$. Then we can compute the probability $q_{\mathbf{I},b} = \Pr((t_1, t_2, \ldots, t_K) \in Q_{\mathbf{I},b})$ by using the inclusion-exclusion principle:

$$q_{\mathbf{I},b} = \sum_{j=0}^{K} (-1)^{K-j} \sum_{\mathbf{J}:\mathbf{J}\subseteq[K],|\mathbf{J}|=j} c(\mathbf{I}, \mathbf{J}, b),$$

where $c(\mathbf{I}, \mathbf{J}, b) = C(p_{\mathbf{I},\mathbf{J},b,1}, p_{\mathbf{I},\mathbf{J},b,2}, \ldots, p_{\mathbf{I},\mathbf{J},b,K})$ and

$$p_{\mathbf{I},\mathbf{J},b,k} = \begin{cases} 1 & \text{if } k \in \mathbf{J} \setminus \mathbf{I}, \\ F_k(\zeta_b) & \text{if } k \in \mathbf{J} \cap \mathbf{I}, \\ F_k(\zeta_{b-1}) & \text{if } k \notin \mathbf{J}. \end{cases}$$

Note that, if $K = 2$, the quantity $q_{\mathbf{I},b}$ here is equal to the quantities computed in equations (10), (11), and (12).

As generalizations of equations (6) and (7) for $K > 2$, we represent $r_{b,k}$ using $q_{\mathbf{I},b}$ as

$$r_{b,k} = q_{\{k\},b} - \underbrace{\sum_{i=2}^{K} (-1)^i \sum_{\mathbf{H}:k\in\mathbf{H}\subseteq[K],|\mathbf{H}|=i} w_{\mathbf{H}} \cdot q_{\mathbf{H},b}}_{\text{Correction term}},$$

Algorithm 2 Algorithm to solve equations

1: Initialize $\hat{\mathbf{F}}_{b|x} = \hat{\mathbf{F}}_{b-1|x}$
2: **while** $\hat{\mathbf{F}}_{b|x}$ is not converged **do**
3: **for** $k \in \{1, 2, \ldots, K\}$ **do**
4: Increase the k-th element of $\hat{\mathbf{F}}_{b|x}$ so that the k-th equation of $\hat{\mathbf{r}}_{b|x} = g_C(\hat{\mathbf{F}}_{b|x}|\hat{\mathbf{F}}_{b-1|x})$ is satisfied (while other k'-th element ($k \neq k'$) of $\hat{\mathbf{F}}_{b|x}$ is fixed)
5: **end for**
6: **end while**
7: **return** $\hat{\mathbf{F}}_{b|x}$

where $w_{\mathbf{H}} \geq 0$ is a weight parameter and we assume that $w_{\mathbf{H}} = 1/|\mathbf{H}|$. Note that, if $K = 2$, this equation is equal to equations (6) and (7).

Since $q_{\mathbf{I},b}$ are functions of $\mathbf{F}_b$, $\mathbf{F}_{b-1}$, and copula C, we can write this relationship as

$$\mathbf{r}_b = g_C(\mathbf{F}_b|\mathbf{F}_{b-1}),$$

by using a function g_C that depends on C. Having established this equation, we can estimate $\hat{\mathbf{F}}_b$ by using Algorithm 1.

B.2. Solving Equations

In this section, we present an algorithm to solve Eq. (13). Algorithm 2 shows a pseudo-code to solve this equation, and it capitalizes on the following property:

Property B.1. Assuming that $\hat{\mathbf{F}}_{b-1|x}$ is fixed:

(1) The k-th element of the length-K vector $g_C(\hat{\mathbf{F}}_{b|x}|\hat{\mathbf{F}}_{b-1|x})$ is *monotonically increasing* with respect to the k-th element of $\hat{\mathbf{F}}_{b|x}$.

(2) The $k'(\neq k)$-th element of the length-K vector $g_C(\hat{\mathbf{F}}_{b|x}|\hat{\mathbf{F}}_{b-1|x})$ is *monotonically decreasing* with respect to the k'-th element of $\hat{\mathbf{F}}_{b|x}$.

First, we demonstrate that the following inequality always holds during the execution of Algorithm 2:

$$\hat{\mathbf{r}}_{b|x} \geq g_C(\hat{\mathbf{F}}_{b|x}|\hat{\mathbf{F}}_{b-1|x}). \tag{17}$$

At line 1 of Algorithm 2, $\hat{\mathbf{F}}_{b|x}$ is initialized with $\hat{\mathbf{F}}_{b-1|x}$. Hence, by the definition of Eq. (13), we have $g_C(\hat{\mathbf{F}}_{b|x}|\hat{\mathbf{F}}_{b-1|x}) = \mathbf{0}$, which means that inequality (17) holds. At line 4 of this algorithm, we can increase the k-th element of $\hat{\mathbf{F}}_{b|x}$ to satisfy $\hat{\mathbf{r}}_{b|x} = g_C(\hat{\mathbf{F}}_{b|x}|\hat{\mathbf{F}}_{b-1|x})$ due to Property B.1(1), and this increment does not violate inequality (17) due to Property B.1(2). Since each element in the length-K vector $\hat{\mathbf{F}}_{b|x}$ does not decrease during the execution of this algorithm, we can find the solution to equation (13) by repeating the while-loop (lines 2–6) until the convergence of $\hat{\mathbf{F}}_{b|x}$.

Parameterization. When we formulate equation (13), we assume that $F_k(\zeta_b|x)$ is represented by using some parameters. In our implementation, we used the parameterization by using the softmax function

$$F_k(\zeta_b|x) = \frac{\sum_{b'=1}^{b} \exp(\alpha_{b',k,x})}{\sum_{b'=1}^{B} \exp(\alpha_{b',k,x})}$$

with B parameters $\alpha_{b,k,x}$ for each k and x. The softmax function was chosen to ensure that any function $F_k(\zeta|x)$ can be expressed. Note that, any other parameterization can be used to represent $F_k(\zeta_b|x)$ in our algorithm.

B.3. Alternative Algorithm for Step 2

We briefly explain the algorithm presented in (Carrière, 1995). This algorithm exploits the following equation:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}V_k(t) &= -\left.\frac{\partial}{\partial t_k}\overline{C}(1-F_1(t_1), 1-F_2(t_2), \dots, 1-F_K(t_K))\right|_{t_1=t_2=\cdots=t_K=t} \\ &= -\left.\frac{\partial}{\partial t_k}\overline{C}(S_1(t_1), S_2(t_2), \dots, S_K(t_K))\right|_{t_1=t_2=\cdots=t_K=t} \\ &= -\left.\frac{\partial \overline{C}(u_1, u_2, \dots, u_K)}{\partial u_k}\right|_{u_1=S_1(t), u_2=S_2(t), \dots, u_K=S_K(t)} \left.\frac{\mathrm{d}}{\mathrm{d}t}S_k(t_k)\right|_{t_k=t},\end{aligned}$$

where $\overline{C}$ is the survival copula corresponding to copula C. Assuming that B is sufficiently large, his algorithm solves this equation with respect to $S_k(t)$ on $t \in \{\zeta_b\}_{b=0}^B$ by using an estimate $\hat{V}_k(t)$ of $V_k(t)$ and these approximations for all k:

$$\begin{aligned}u_k &\approx \frac{S_k(\zeta_b) + S_k(\zeta_{b+1})}{2}, \\ \left.\frac{\mathrm{d}}{\mathrm{d}t}S_k(t)\right|_{t=\zeta_b} &\approx \frac{S_k(\zeta_{b+1}) - S_k(\zeta_b)}{\zeta_{b+1} - \zeta_b}, \\ \left.\frac{\mathrm{d}}{\mathrm{d}t}\hat{V}_k(t)\right|_{t=\zeta_b} &\approx \frac{\hat{V}_k(\zeta_{b+1}) - \hat{V}_k(\zeta_b)}{\zeta_{b+1} - \zeta_b}.\end{aligned}$$

This algorithm is designed to estimate the average survival function $S_k(t)$ and thereby $F_k(t)$, but we can modify the algorithm to obtain $F_k(t|x)$ conditional on x if a conditional estimate $\hat{V}_k(t|x)$ is available.

C. Theoretical Analysis

In this section, we theoretically verify that the two-step algorithm outputs solutions with sufficiently small errors. We consider the case $K = 2$ for simplicity, and we assume $\zeta_b = \frac{b}{B}\zeta_B$. As discussed in the preceding section, various models can be implemented to estimate the CIF in the first step of our algorithm. Therefore, we evaluate errors affected by Step 2, solving (13), under the assumption that the models employed in the first step accurately approximate the true probabilities such that

$$|\hat{r}_{b,k} - r_{b,k}| \le \epsilon \tag{18}$$

holds for all $b = 1, \dots, B$ and $k = 1, 2$. Note that while how small a value we can take as ϵ in (18) depends on the choice of the model in Step 1, we can apply the results exhibited in this section. We provide examples of achieving (18) in Appendix D. To formally state our theoretical results, we introduce the following assumption:

Assumption C.1. We assume the following conditions:

(1) (True probability is not biased.) There exists a global constant $c_0 > 0$ such that for every $b = 1, \dots, B$ and $k = 1, 2$, $F_k(\zeta_b|x) - F_k(\zeta_{b-1}|x) = \Pr(\zeta_{b-1} < T_k \le \zeta_b) \le \frac{c_0}{B}$ holds.

(2) (Copula.) We assume that the copula C is of class $\mathcal{C}^2$ and satisfies

$$\begin{aligned}\ell &\coloneqq \inf_{(u,v)\in[0,1]^2} \frac{\partial^2}{\partial u \partial v}C(u,v) > 0, \\ L &\coloneqq \sup_{(u,v)\in[0,1]^2} \max\left\{\frac{\partial^2}{\partial u^2}C(u,v), \frac{\partial^2}{\partial u \partial v}C(u,v), \frac{\partial^2}{\partial v^2}C(u,v)\right\} < +\infty.\end{aligned}$$

(3) (All t_k are equally observed.) There exist constants $c_1 > 0$ depend on ℓ, L, and $\tau > 0$ such that the following condition holds: Let $\delta_0 > 0$ be a constant determined by ℓ and L and $b_0 \coloneqq \max_b\left\{b \mid \forall k, F_k(\zeta_b|x) \le 1 - \frac{\delta_0}{\tau \log B}\right\}$. Then, $\min_k F_k(\zeta_{b_0}|x) \ge 1 - \frac{c_1}{\log B}$.

We make some remarks on the assumption. The first condition is required to bound the error by the choice of w_1 and w_2; if the probability concentrates on a squared region partitioned by suboptimal w_1 and w_2, significant errors are inevitable.

This condition is satisfied if F_k is Lipschitz-continuous with $c_0\zeta_B$ serving as the Lipschitz constant. The second condition manages the sensitivity of the estimation relative to the true distribution and noise. For example, if $\frac{\partial^2 C}{\partial u \partial v} \ll 1$, indicating that C exhibits minimal variation as u and v change, substantial adjustments to the estimation are necessary to accommodate for noise and achieve (13). This condition is typically met for the independence copula $C(u, v) = uv$ with any $\ell < 1$ and $L > 1$. The third condition appears to be technical. As will be demonstrated in subsequent analyses, errors between $\hat{F}_k$ and F_k can only be effectively bounded for $b \leq b_0$. As b approaches B, and consequently $r_{b,k}$ diminishes, the impact of ϵ intensifies. Condition (3) excludes scenarios where a part of t_ks is concentrated in the region $b > b_0$. In other words, all t_ks are equally observed in the region $b \leq b_0$.

Let $W_1(\cdot, \cdot)$ be the Wasserstein distance[1]. Then, we provide the statement about the W_1 distance between the estimated and true probabilities. We consider the extension of $\hat{F}_k(\zeta_b|x)$ to a CDF on $[0, \zeta_B]$ by $\hat{F}_k(t|x) \coloneqq \hat{F}_k(\zeta_b|x)$, where $\zeta_b < t \leq \zeta_{b+1}$.

Theorem C.2. *Suppose that Assumption C.1 holds. Then, there exists a constant $c_\epsilon > 0$ depending c_0, ℓ and L such that if $\epsilon \leq \frac{c_\epsilon}{B}$, the following inequality holds:*

$$W_1\big(\hat{\mu}_{k|x}, \mu_{k|x}\big) \lesssim \zeta_B \left(B^{1+\tau}\epsilon + c_1 \cdot \frac{B - b_0}{B \log B} \right), \tag{19}$$

where $\hat{\mu}_{k|x}$ and $\mu_{k|x}$ are probability measures whose CDFs are given by $\hat{F}_k(\cdot|x)$ and $F_k(\cdot|x)$, respectively.

The proof is deffered to the following subsection. Suppose that the condition $\epsilon = o(B^{1+\tau})$ holds. Then, we obtain an upper bound as $W_1\big(\hat{\mu}_{k|x}, \mu_{k|x}\big) = \zeta_B \cdot o(1)$. Thus, we can ensure that as $B \to +\infty$ and the sample size increases as we can take sufficiently small ϵ, the output of Step 2 converges to the ground truth distribution in terms of the W_1 distance.

We provide some comments on Theorem C.2. We can observe a trade-off in (19) based on the choice of B: while the second term decreases as B increases, $B^{1+\tau}$ and ϵ in the first term should increase. Consequently, an optimal choice of B should be considered under appropriate assumptions that determine ϵ and b_0, such as the model utilized in Step 1 and the properties of $\mathbf{F}_{\cdot|x}$. It is also significant to examine that the derived bound achieves a statistical min-max lower bound exhibited in (Niles-Weed & Berthet, 2022; Bilodeau et al., 2023), for example. We reserve these considerations for future research endeavors.

C.1. Proof of Theorem C.2

This subsection provides proofs of Theorem C.2. Here, for notational simplicity, we abbreviate the conditional variable x. For example, we denote $F_k(\zeta_b)$ instead of $F_k(\zeta_b|x)$. We assume $w_1 = w_2 = \frac{1}{2}$ just for simplicity. We can extend our analysis to arbitrary $w_1, w_2 \in (0, 1)$. Moreover, we assume that Step 2 exactly solves (13), i.e., $\hat{F}_k$ exactly satisfies the equation (13).

First, we evaluate the error between the outputs of Step 2 and the true probability.

Proposition C.3 (Estimation error when solving (13))**.** *Suppose that Assumption C.1 holds. Let $\hat{\mathbf{F}}_b$ be the output of Algorithm 1. Then, for every $b = 1, \ldots, B$ and $k = 1, 2$, there exists a constant $c_\epsilon > 0$ depending c_0, ℓ and L such that under the condition $\epsilon \leq \frac{c_\epsilon}{B}$, for sufficiently large B and every b satisfying $\max_{k\in\{1,2\}} F_k(\zeta_b) \leq 1 - \delta$ with a positive constant δ satisfying $\frac{8c_0 L}{\tau \ell \log B} \leq \delta$ with $\tau \in (0, \frac{1}{2})$,*

$$|\hat{F}_k(\zeta_b) - F_k(\zeta_b)| \leq \left[\epsilon + \frac{16L\delta^{-2}}{\ell^2} \left(\epsilon + L \cdot \frac{c_0}{B} \right)^2 \right] \cdot \left(\frac{1}{\ell} + \frac{B}{4c_0 L} \right) \left(1 - \frac{8c_0 L}{B\ell\delta} \right)^{-b+1}. \tag{20}$$

Proof. Let us denote $\Delta_{b,k} \coloneqq \hat{F}_k(\zeta_b) - F_k(\zeta_b)$. Instead of (20), we aim to obtain a tighter bound

$$|\Delta_{b,k}| \leq \left[\epsilon + \frac{16L\delta^{-2}}{\ell^2} \left(\epsilon + L \cdot \frac{c_0}{B} \right)^2 \right] \cdot \left[\left(\frac{1}{\ell} + \frac{B}{4c_0 L} \right) \left(1 - \frac{8c_0 L}{B\ell\delta} \right)^{-b+1} - \frac{B}{4c_0 L} \right]. \tag{21}$$

[1] $W_1(\mu, \nu) \coloneqq \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^2} |x - y| \mathrm{d}\pi(x, y)$, where $\Pi(\mu, \nu)$ denotes the set of all couplings of two probability measures μ and ν on $\mathbb{R}$.

We give its proof by induction on b. For the case $b = 1$, we have

$$C(\hat{F}_1(\zeta_1), 1) = \hat{r}_{1,1} + w_1 \cdot C(\hat{F}_1(\zeta_1), \hat{F}_2(\zeta_1))$$

and

$$C(F_1(\zeta_1), 1) = r_{1,1} + w^*_{1,1} \cdot C(F_1(\zeta_1), F_2(\zeta_1))$$

with $w^*_{1,1} \in [0, 1]$. By taking the difference of the both sides, we obtain

$$C(\hat{F}_1(\zeta_1), 1) - C(F_1(\zeta_1), 1) = \hat{r}_{1,1} - r_{1,1} + w_1 \cdot C(\hat{F}_1(\zeta_1), \hat{F}_2(\zeta_1)) - w^*_{1,1} \cdot C(F_1(\zeta_1), F_2(\zeta_1)). \tag{22}$$

First, we evaluate the term $C(\hat{F}_1(\zeta_1), \hat{F}_2(\zeta_1))$. Rearranging terms gives

$$\begin{aligned} C(\hat{F}_1(\zeta_1), 1) - w_1 \cdot C(\hat{F}_1(\zeta_1), \hat{F}_2(\zeta_1)) &= \hat{r}_{1,1} - r_{1,1} + C(F_1(\zeta_1), 1) - w^*_{1,1} \cdot C(F_1(\zeta_1), F_2(\zeta_1)) \\ &\leq \epsilon + C(F_1(\zeta_1), 1) \\ &\leq \epsilon + L \cdot \frac{c_0}{B}, \end{aligned}$$

where the first inequality follows from $\hat{r}_{1,1} - r_{1,1} \leq \epsilon$ by (18) and $w^*_{1,1} \cdot C(F_1(\zeta_1), F_2(\zeta_1)) \geq 0$, and the last inequality is derived from $F_1(\zeta_1) \leq \frac{c_0}{B}$ by Assumption C.1-(1) and

$$C(F_1(\zeta_1), 1) = \int_0^{F_1(\zeta_1)} \int_0^1 \underbrace{\frac{\partial^2}{\partial u \partial v} C(u, v)}_{\leq L} \mathrm{d}v \mathrm{d}u \leq L \cdot F_1(\zeta_1) \leq L \cdot \frac{c_0}{B}.$$

Since $C(\hat{F}_1(\zeta_1), \hat{F}_2(\zeta_1)) \leq C(\hat{F}_1(\zeta_1), 1)$, we obtain

$$C(\hat{F}_1(\zeta_1), 1) - w_1 \cdot C(\hat{F}_1(\zeta_1), 1) \leq C(\hat{F}_1(\zeta_1), 1) - w_1 \cdot C(\hat{F}_1(\zeta_1), \hat{F}_2(\zeta_1)) \leq \epsilon + L \cdot \frac{c_0}{B}.$$

By using $1 - w_1 = \frac{1}{2}$, we obtain

$$C(\hat{F}_1(\zeta_1), 1) \leq 2\Big(\epsilon + L \cdot \frac{c_0}{B}\Big).$$

Moreover, we have

$$C(\hat{F}_1(\zeta_1), 1) = \int_0^{F_1(\zeta_1)} \int_0^1 \underbrace{\frac{\partial^2}{\partial u \partial v} C(u, v)}_{\geq \ell} \mathrm{d}v \mathrm{d}u \geq \ell \hat{F}_1(\zeta_1),$$

and hence,

$$\hat{F}_1(\zeta_1) \leq \frac{2}{\ell}\Big(\epsilon + L \cdot \frac{c_0}{B}\Big).$$

A similar argument gives

$$\hat{F}_2(\zeta_1) \leq \frac{2}{\ell}\Big(\epsilon + L \cdot \frac{c_0}{B}\Big).$$

By combining these bounds, we obtain

$$C(\hat{F}_1(\zeta_1), \hat{F}_2(\zeta_1)) = \int_0^{\hat{F}_1(\zeta_1)} \int_0^{\hat{F}_2(\zeta_1)} \underbrace{\frac{\partial^2}{\partial u \partial v} C(u, v)}_{\leq L} \mathrm{d}v \mathrm{d}u \leq L \cdot \hat{F}_1(\zeta_1) \hat{F}_2(\zeta_1) \leq \frac{4L}{\ell^2}\Big(\epsilon + L \cdot \frac{c_0}{B}\Big)^2. \tag{23}$$

Thus we get the bound on the term $C(\hat{F}_1(\zeta_1), \hat{F}_2(\zeta_1))$.

Moreover, we have

$$C(F_1(\zeta_1), F_2(\zeta_1)) = \int_0^{F_1(\zeta_1)} \int_0^{F_2(\zeta_1)} \underbrace{\frac{\partial^2}{\partial u \partial v} C(u,v)}_{\leq L} \mathrm{d}v\mathrm{d}u \leq L \cdot F_1(\zeta_1)F_2(\zeta_1) \leq L \cdot \left(\frac{c_0}{B}\right)^2, \tag{24}$$

where we use Assumption C.1-(1) for the last inequality.

Then, by taking the absolute value of the both sides of (22), we have

$$\begin{aligned}
\left|C(\hat{F}_1(\zeta_1), 1) - C(F_1(\zeta_1), 1)\right| &= \left|\hat{r}_{1,1} - r_{1,1} + w_1 \cdot C(\hat{F}_1(\zeta_1), \hat{F}_2(\zeta_1)) - w^*_{1,1} \cdot C(F_1(\zeta_1), F_2(\zeta_1))\right| \\
&\leq |\hat{r}_{1,1} - r_{1,1}| + \left|w_1 \cdot C(\hat{F}_1(\zeta_1), \hat{F}_2(\zeta_1)) - w^*_{1,1} \cdot C(F_1(\zeta_1), F_2(\zeta_1))\right| \\
&\leq \epsilon + \max\left\{C(\hat{F}_1(\zeta_1), \hat{F}_2(\zeta_1)), C(F_1(\zeta_1), F_2(\zeta_1))\right\} \\
&\leq \epsilon + \max\left\{\frac{4L}{\ell^2}\left(\epsilon + L \cdot \frac{c_0}{B}\right)^2, L \cdot \left(\frac{c_0}{B}\right)^2\right\} \\
&\leq \epsilon + \frac{4L}{\ell^2}\left(\epsilon + L \cdot \frac{c_0}{B}\right)^2,
\end{aligned}$$

where we use the triangle inequality for the first inequality, $0 \leq w_1, w^*_{1,1} \leq 1$ for the second one, and (23), (24) for the third one. Since $\left|C(\hat{F}_1(\zeta_1), 1) - C(F_1(\zeta_1), 1)\right| \geq \ell\left|\hat{F}_1(\zeta_1) - F_1(\zeta_1)\right| = \ell|\Delta_{1,1}|$, we obtain

$$|\Delta_{1,1}| \leq \frac{1}{\ell}\left[\epsilon + \frac{4L}{\ell^2}\left(\epsilon + L \cdot \frac{c_0}{B}\right)^2\right] \leq \frac{1}{\ell}\left[\epsilon + \frac{16L\delta^{-2}}{\ell^2}\left(\epsilon + L \cdot \frac{c_0}{B}\right)^2\right],$$

where the last term coincides to the right hand side of (21) with $b = 1$.

We can obtain the bound for $k = 2$ by utilizing the same argument. Thus we get (21) for $b = 1$.

Assume that (21) holds for $b = b'$, i.e.,

$$|\Delta_{b',k}| \leq \left[\epsilon + \frac{16L\delta^{-2}}{\ell^2}\left(\epsilon + L \cdot \frac{c_0}{B}\right)^2\right] \cdot \left[\left(\frac{1}{\ell} + \frac{B}{4c_0L}\right)\left(1 - \frac{8c_0L}{B\ell\delta}\right)^{-b'+1} - \frac{B}{4c_0L}\right]$$

holds for $k = 1, 2$. We consider the case $b = b' + 1$. This bound gives that by taking $\epsilon \leq \frac{c_\epsilon}{B}$ with a sufficiently small c_ϵ,

$$\begin{aligned}
|\Delta_{b',k}| &\leq \left[\epsilon + \frac{16L\delta^{-2}}{\ell^2}\left(\epsilon + L \cdot \frac{c_0}{B}\right)^2\right] \cdot \left(\frac{1}{\ell} + \frac{B}{4c_0L}\right)\left(1 - \frac{8c_0L}{B\ell\delta}\right)^{-b'+1} \\
&\leq \left[\epsilon + \frac{16L\delta^{-2}}{\ell^2}\left(\epsilon + L \cdot \frac{c_0}{B}\right)^2\right] \cdot \left(\frac{1}{\ell} + \frac{B}{4c_0L}\right)\left(1 - \frac{8c_0L}{B\ell\delta}\right)^{-B} \\
&\leq \left[\epsilon + \frac{16L\delta^{-2}}{\ell^2}\left(\epsilon + L \cdot \frac{c_0}{B}\right)^2\right] \cdot \left(\frac{1}{\ell} + \frac{B}{4c_0L}\right)\exp\left(\frac{8c_0L}{\ell\delta}\right) \\
&\leq \left[\epsilon + \frac{\ell \log B}{2c_0^2}\left(\epsilon + L \cdot \frac{c_0}{B}\right)^2\right] \cdot \left(\frac{1}{\ell} + \frac{B}{4c_0L}\right)B^\tau \qquad (25) \\
&\lesssim \frac{1}{B^{1-\tau}}, \qquad (26)
\end{aligned}$$

where we use $1 - x \leq e^{-x}$ in the third inequality and the definition of δ in the fourth inequality. This and $F_k(\zeta_{b'}) < 1 - \delta$ gives $\hat{F}_k(\zeta_{b'}) < 1 - \frac{\delta}{2}$ for sufficiently large B.

We remind

$$\hat{r}_{b'+1,1} = \hat{q}_{\{1\},b'} - w_1 \cdot \hat{q}_{\{1,2\},b'}, \tag{27}$$

$$r_{b'+1,1} = q_{\{1\},b'} - w^*_{b'+1,1} \cdot q_{\{1,2\},b'}, \tag{28}$$

Now, we consider integral representation of $\hat{q}_{\{1\},b'}$ and $\hat{q}_{\{1,2\},b'}$ as

$$\begin{aligned}
\hat{q}_{\{1\},b'} &= C(\hat{F}_1(\zeta_{b'+1}),1) - C(\hat{F}_1(\zeta_{b'}),1) - C(\hat{F}_1(\zeta_{b'+1}),\hat{F}_2(\zeta_{b'})) + C(\hat{F}_1(\zeta_{b'}),\hat{F}_2(\zeta_{b'})) \\
&= \int_{\hat{F}_1(\zeta_{b'})}^{\hat{F}_1(\zeta_{b'+1})} \int_{\hat{F}_2(\zeta'_b)}^{1} \frac{\partial^2}{\partial u \partial v} C(u,v) \mathrm{d}v \mathrm{d}u, \qquad (29) \\
\hat{q}_{\{1,2\},b'} &= C(\hat{F}_1(\zeta_{b'+1}),\hat{F}_2(\zeta_{b'+1})) - C(\hat{F}_1(\zeta_{b'+1}),\hat{F}_2(\zeta_{b'})) - C(\hat{F}_1(\zeta_{b'}),\hat{F}_2(\zeta_{b'+1})) + C(\hat{F}_1(\zeta_{b'}),\hat{F}_2(\zeta_{b'})) \\
&= \int_{\hat{F}_1(\zeta_{b'})}^{\hat{F}_1(\zeta_{b'+1})} \int_{\hat{F}_2(\zeta'_b)}^{\hat{F}_2(\zeta_{b'+1})} \frac{\partial^2}{\partial u \partial v} C(u,v) \mathrm{d}v \mathrm{d}u.
\end{aligned}$$

The same expression holds for $q_{\{1\},b'}$ and $q_{\{1,2\},b'}$ by replacing $\hat{F}_1$ and $\hat{F}_2$ with F_1 and F_2. By taking the difference between (27) and (28), we obtain

$$\hat{r}_{b'+1,1} - r_{b'+1,1} = \hat{q}_{\{1\},b'} - \hat{q}_{\{1\},b'} - w_1 \cdot \hat{q}_{\{1,2\},b'} + w^*_{1,1} \cdot q_{\{1,2\},b'}. \qquad (30)$$

Similar to the case $b = 1$, we first evaluate the term $\hat{q}_{\{1,2\},b'}$ (note that $\hat{q}_{\{1,2\},b'} = C(\hat{F}_1(\zeta_1),\hat{F}_2(\zeta_1))$). By rearranging terms, we have

$$\begin{aligned}
\hat{q}_{\{1\},b'} - w_1 \cdot \hat{q}_{\{1,2\},b'} &= \hat{r}_{b'+1,1} - r_{b'+1,1} + q_{\{1\},b'} - w^*_{1,1} \cdot q_{\{1,2\},b'} \\
&\le \epsilon + q_{\{1\},b'} \\
&= \epsilon + \int_{F_1(\zeta_{b'})}^{F_1(\zeta_{b'+1})} \int_{F_2(\zeta'_b)}^{1} \underbrace{\frac{\partial^2}{\partial u \partial v} C(u,v)}_{\le L} \mathrm{d}v \mathrm{d}u \\
&\le \epsilon + L \cdot (F_1(\zeta_{b'+1}) - F_1(\zeta_{b'}))(1 - F_2(\zeta'_b)) \\
&\le \epsilon + L \cdot \frac{c_0}{B}, \qquad (31)
\end{aligned}$$

where the first inequality follows from $\hat{r}_{b'+1,1} - r_{b'+1,1} \le \epsilon$ by (18) and $q_{\{1,2\},b'} \ge 0$, and the last inequality, follows from Assumption C.1-(1) and $1 - F_2(\zeta_b) < 1$. Moreover, the left hand side is lower bounded by

$$\begin{aligned}
\hat{q}_{\{1\},b'} - w_1 \cdot \hat{q}_{\{1,2\},b'} &\ge (1 - w_1)\hat{q}_{\{1\},b'} \\
&= \frac{1}{2} \int_{\hat{F}_1(\zeta_{b'})}^{\hat{F}_1(\zeta_{b'+1})} \int_{\hat{F}_2(\zeta'_b)}^{1} \underbrace{\frac{\partial^2}{\partial u \partial v} C(u,v)}_{\ge \ell} \mathrm{d}v \mathrm{d}u \\
&\ge \frac{1}{2} \cdot \ell\Big(\hat{F}_1(\zeta_{b'+1}) - \hat{F}_1(\zeta_{b'})\Big)(1 - \hat{F}_2(\zeta_{b'})),
\end{aligned}$$

where we use $1 - w_1 = \frac{1}{2}$ and (29) for the equality. Combining this with (31), we have

$$\hat{F}_1(\zeta_{b'+1}) - \hat{F}_1(\zeta_{b'}) \le \frac{2}{\ell(1 - \hat{F}_2(\zeta_{b'}))}\Big(\epsilon + L \cdot \frac{c_0}{B}\Big). \qquad (32)$$

This and the triangle inequality give

$$\begin{aligned}
|\Delta_{b'+1,1}| &= \Big|\Big(\hat{F}_1(\zeta_{b'+1}) - \hat{F}_1(\zeta_{b'})\Big) + \Big(\hat{F}_1(\zeta_{b'}) - F_1(\zeta_{b'})\Big) + (F_1(\zeta_{b'}) - F_1(\zeta_{b'+1}))\Big| \\
&\le \Big|\hat{F}_1(\zeta_{b'+1}) - \hat{F}_1(\zeta_{b'})\Big| + \Big|\hat{F}_1(\zeta_{b'}) - F_1(\zeta_{b'})\Big| + |F_1(\zeta_{b'}) - F_1(\zeta_{b'+1})| \\
&\le \frac{2}{\ell(1 - \hat{F}_2(\zeta_{b'}))}\Big(\epsilon + L \cdot \frac{c_0}{B}\Big) + |\Delta_{b',1}| + \frac{c_0}{B},
\end{aligned}$$

which we will use in the latter of this proof.

The same bound as (32) holds for $k = 2$, which gives

$$\Big(\hat{F}_1(\zeta_{b'+1}) - \hat{F}_1(\zeta_{b'})\Big)\Big(\hat{F}_2(\zeta_{b'+1}) - \hat{F}_2(\zeta_{b'})\Big) \le \frac{4}{\ell^2\Big(1 - \hat{F}_1(\zeta_{b'})\Big)\Big(1 - \hat{F}_2(\zeta_{b'})\Big)}\Big(\epsilon + L \cdot \frac{c_0}{B}\Big)^2.$$

Thus, we obtain

$$\begin{aligned}\hat{q}_{\{1,2\},b'} &= \int_{\hat{F}_1(\zeta_{b'})}^{\hat{F}_1(\zeta_{b'+1})} \int_{\hat{F}_2(\zeta'_b)}^{\hat{F}_2(\zeta_{b'+1})} \underbrace{\frac{\partial^2}{\partial u \partial v} C(u,v)}_{\leq L} \mathrm{d}v\mathrm{d}u \\ &\leq L \cdot \Big(\hat{F}_1(\zeta_{b'+1}) - \hat{F}_1(\zeta_{b'})\Big)\Big(\hat{F}_2(\zeta_{b'+1}) - \hat{F}_2(\zeta_{b'})\Big) \\ &\leq \frac{4L}{\ell^2\Big(1-\hat{F}_1(\zeta_{b'})\Big)\Big(1-\hat{F}_2(\zeta_{b'})\Big)}\Big(\epsilon + L\cdot\frac{c_0}{B}\Big)^2.\end{aligned} \tag{33}$$

Moreover, we have

$$\begin{aligned}q_{\{1,2\},b'|x} &= \int_{F_1(\zeta_{b'})}^{F_1(\zeta_{b'+1})} \int_{F_2(\zeta'_b)}^{F_2(\zeta_{b'+1})} \underbrace{\frac{\partial^2}{\partial u \partial v} C(u,v)}_{\leq L} \mathrm{d}v\mathrm{d}u \\ &\leq L \cdot (F_1(\zeta_{b'+1}) - F_1(\zeta_{b'}))(F_2(\zeta_{b'+1}) - F_2(\zeta_{b'})) \\ &\leq L \cdot \Big(\frac{c_0}{B}\Big)^2,\end{aligned} \tag{34}$$

where we use Assumption C.1-(1) for the last inequality.

Then, by taking the absolute value of the both sides of (30), we have

$$\begin{aligned}\big|\hat{q}_{\{1\},b'|x} - q_{\{1\},b'|x}\big| &= \big|\hat{r}_{b'+1,1} - r_{b'+1,1} + w_1 \cdot \hat{q}_{\{1,2\},b'|x} - w^*_{b'+1,1} \cdot q_{\{1,2\},b'|x}\big| \\ &\leq |\hat{r}_{b'+1,1} - r_{b'+1,1}| + \big|w_1 \cdot \hat{q}_{\{1,2\},b'|x} - w^*_{b'+1,1} \cdot q_{\{1,2\},b'|x}\big| \\ &\leq \epsilon + \max\big\{\hat{q}_{\{1,2\},b'|x}, q_{\{1,2\},b'|x}\big\} \\ &\leq \epsilon + \max\left\{\frac{4L}{\ell^2\Big(1-\hat{F}_1(\zeta_{b'})\Big)\Big(1-\hat{F}_2(\zeta_{b'})\Big)}\Big(\epsilon + L\cdot\frac{c_0}{B}\Big)^2, L\cdot\Big(\frac{c_0}{B}\Big)^2\right\} \\ &\leq \epsilon + \frac{4L}{\ell^2\Big(1-\hat{F}_1(\zeta_{b'})\Big)\Big(1-\hat{F}_2(\zeta_{b'})\Big)}\Big(\epsilon + L\cdot\frac{c_0}{B}\Big)^2,\end{aligned} \tag{35}$$

where we use the triangle inequality for the first inequality, $0 \leq w_1, w^*_{b'+1,1} \leq 1$ for the second one, and (33) and (34) for the third one.

Then, we evaluate the left hand side. We have

$$\big|\hat{q}_{\{1\},b'} - q_{\{1\},b'}\big| = \left|\int_{\hat{F}_1(\zeta_{b'})}^{\hat{F}_1(\zeta_{b'+1})} \int_{\hat{F}_2(\zeta'_b)}^{1} \frac{\partial^2}{\partial u \partial v} C(u,v)\mathrm{d}v\mathrm{d}u - \int_{F_1(\zeta_{b'})}^{F_1(\zeta_{b'+1})} \int_{F_2(\zeta'_b)}^{1} \frac{\partial^2}{\partial u \partial v} C(u,v)\mathrm{d}v\mathrm{d}u\right|$$

and

$$
\begin{aligned}
&\int_{F_1(\zeta_{b'})}^{F_1(\zeta_{b'+1})}\int_{F_2(\zeta'_b)}^{1}\frac{\partial^2}{\partial u\partial v}C(u,v)\mathrm{d}v\mathrm{d}u-\int_{\hat{F}_1(\zeta_{b'})}^{\hat{F}_1(\zeta_{b'+1})}\int_{\hat{F}_2(\zeta'_b)}^{1}\frac{\partial^2}{\partial u\partial v}C(u,v)\mathrm{d}v\mathrm{d}u\\
&=\int_{F_1(\zeta_{b'})}^{F_1(\zeta_{b'+1})}\frac{\partial}{\partial u}C(u,1)\mathrm{d}u-\int_{\hat{F}_1(\zeta_{b'})}^{\hat{F}_1(\zeta_{b'+1})}\frac{\partial}{\partial u}C(u,1)\mathrm{d}u\\
&\quad-\int_{F_1(\zeta_{b'})}^{F_1(\zeta_{b'+1})}\frac{\partial}{\partial u}C(u,F_2(\zeta_{b'}))\mathrm{d}u+\int_{\hat{F}_1(\zeta_{b'})}^{\hat{F}_1(\zeta_{b'+1})}\frac{\partial}{\partial u}C(u,\hat{F}_2(\zeta_{b'}))\mathrm{d}u\\
&=\int_{\hat{F}_1(\zeta_{b'+1})}^{F_1(\zeta_{b'+1})}\frac{\partial}{\partial u}C(u,1)\mathrm{d}u-\int_{\hat{F}_1(\zeta_{b'})}^{F_1(\zeta_{b'})}\frac{\partial}{\partial u}C(u,1)\mathrm{d}u\\
&\quad-\int_{\hat{F}_1(\zeta_{b'+1})}^{\hat{F}_1(\zeta_{b'})}\frac{\partial}{\partial u}C(u,F_2(\zeta_{b'}))\mathrm{d}u+\int_{F_1(\zeta_{b'+1})}^{F_1(\zeta_{b'})}\frac{\partial}{\partial u}C(u,F_2(\zeta_{b'}))\mathrm{d}u\\
&\quad+\int_{\hat{F}_1(\zeta_{b'})}^{\hat{F}_1(\zeta_{b'+1})}\left[\frac{\partial}{\partial u}C(u,\hat{F}_2(\zeta_{b'}))-\frac{\partial}{\partial u}C(u,F_2(\zeta_{b'}))\right]\mathrm{d}u\\
&=\int_{\hat{F}_1(\zeta_{b'+1})}^{F_1(\zeta_{b'+1})}\frac{\partial}{\partial u}C(u,1)\mathrm{d}u-\int_{\hat{F}_1(\zeta_{b'})}^{F_1(\zeta_{b'})}\frac{\partial}{\partial u}C(u,1)\mathrm{d}u\\
&\quad-\int_{\hat{F}_1(\zeta_{b'+1})}^{F_1(\zeta_{b'+1})}\frac{\partial}{\partial u}C(u,F_2(\zeta_{b'}))\mathrm{d}u+\int_{\hat{F}_1(\zeta_{b'})}^{F_1(\zeta_{b'})}\frac{\partial}{\partial u}C(u,F_2(\zeta_{b'}))\mathrm{d}u\\
&\quad+\int_{\hat{F}_1(\zeta_{b'})}^{\hat{F}_1(\zeta_{b'+1})}\left[\frac{\partial}{\partial u}C(u,\hat{F}_2(\zeta_{b'}))-\frac{\partial}{\partial u}C(u,F_2(\zeta_{b'}))\right]\mathrm{d}u.
\end{aligned}
\tag{36}
$$

By the mean value theorem for integrals, there exist constants

$$
u_{b',1},u_{b',2}\in\left[\min\left\{\hat{F}_1(\zeta_{b'}),F_1(\zeta_{b'})\right\},\max\left\{\hat{F}_1(\zeta_{b'}),F_1(\zeta_{b'})\right\}\right]
$$

and

$$
u_{b'+1,1},u_{b'+1,2}\in\left[\min\left\{\hat{F}_1(\zeta_{b'+1}),F_1(\zeta_{b'+1})\right\},\max\left\{\hat{F}_1(\zeta_{b'+1}),F_1(\zeta_{b'+1})\right\}\right]
$$

such that

$$
\begin{aligned}
(36)=&-\Delta_{b'+1,1}\cdot\frac{\partial}{\partial u}C(u_{b'+1,1},1)+\Delta_{b',1}\frac{\partial}{\partial u}C(u_{b',1},1)+\Delta_{b'+1,1}\cdot\frac{\partial}{\partial u}C(u_{b'+1,2},F_2(\zeta_{b'+1}))\\
&-\Delta_{b'+1,1}\cdot\frac{\partial}{\partial u}C(u_{b',2},F_2(\zeta_{b'}))+\int_{\hat{F}_1(\zeta_{b'})}^{\hat{F}_1(\zeta_{b'+1})}\left[\frac{\partial}{\partial u}C(u,\hat{F}_2(\zeta_{b'}))-\frac{\partial}{\partial u}C(u,F_2(\zeta_{b'}))\right]\mathrm{d}u.
\end{aligned}
$$

We denote

$$
\begin{aligned}
(\mathrm{I})=&\Delta_{b'+1,1}\cdot\frac{\partial}{\partial u}C(u_{b'+1,1},1)-\Delta_{b',1}\frac{\partial}{\partial u}C(u_{b',1},1)\\
&-\Delta_{b'+1,1}\cdot\frac{\partial}{\partial u}C(u_{b'+1,2},F_2(\zeta_{b'+1}))+\Delta_{b'+1,1}\cdot\frac{\partial}{\partial u}C(u_{b',2},F_2(\zeta_{b'})),\\
(\mathrm{II})=&\int_{\hat{F}_1(\zeta_{b'})}^{\hat{F}_1(\zeta_{b'+1})}\left[\frac{\partial}{\partial u}C(u,\hat{F}_2(\zeta_{b'}))-\frac{\partial}{\partial u}C(u,F_2(\zeta_{b'}))\right]\mathrm{d}u.
\end{aligned}
$$

We evaluate the each term. First, we have

$$\begin{aligned}
\text{(I)} =&(\Delta_{b'+1,1}-\Delta_{b',1})\frac{\partial}{\partial u}C(u_{b',1},1)+\Delta_{b'+1,1}\left[\frac{\partial}{\partial u}C(u_{b'+1,1},1)-\frac{\partial}{\partial u}C(u_{b',1},1)\right]\\
&-(\Delta_{b'+1,1}-\Delta_{b',1})\frac{\partial}{\partial u}C(u_{b',2},F_2(\zeta_{b'}))-\Delta_{b'+1,1}\left[\frac{\partial}{\partial u}C(u_{b'+1,2},F_2(\zeta_{b'+1}))-\frac{\partial}{\partial u}C(u_{b',2},F_2(\zeta_{b'}))\right]\\
=&(\Delta_{b'+1,1}-\Delta_{b',1})\left[\frac{\partial}{\partial u}C(u_{b',1},1)-\frac{\partial}{\partial u}C(u_{b',2},F_2(\zeta_{b'}))\right]+\Delta_{b'+1,1}\left[\frac{\partial}{\partial u}C(u_{b'+1,1},1)-\frac{\partial}{\partial u}C(u_{b',1},1)\right]\\
&+\Delta_{b'+1,1}\left[\frac{\partial}{\partial u}C(u_{b'+1,2},F_2(\zeta_{b'+1}))-\frac{\partial}{\partial u}C(u_{b',2},F_2(\zeta_{b'}))\right].
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
|(\text{I})| \geq&\left|(\Delta_{b'+1,1}-\Delta_{b',1})\left[\frac{\partial}{\partial u}C(u_{b',1},1)-\frac{\partial}{\partial u}C(u_{b',2},F_2(\zeta_{b'}))\right]\right|\\
&-\left|\Delta_{b'+1,1}\left[\frac{\partial}{\partial u}C(u_{b'+1,1},1)-\frac{\partial}{\partial u}C(u_{b',1},1)\right]\right|-\left|\Delta_{b'+1,1}\left[\frac{\partial}{\partial u}C(u_{b'+1,2},F_2(\zeta_{b'+1}))-\frac{\partial}{\partial u}C(u_{b',2},F_2(\zeta_{b'}))\right]\right|,
\end{aligned}$$

where we use the triangle inequality.

Moreover, we can bound the terms in the above inequality by

$$\left|\frac{\partial}{\partial u}C(u_{b'+1,1},1)-\frac{\partial}{\partial u}C(u_{b',1},1)\right|=\left|\int_{u_{b',1}}^{u_{b'+1,1}}\underbrace{\frac{\partial^2}{\partial u^2}C(u,1)}_{\leq L}\mathrm{d}u\right|\leq L(u_{b'+1,1}-u_{b',1})\leq L\Big(|\Delta_{b'+1,1}|+\frac{c_0}{B}\Big),$$

and

$$\begin{aligned}
&\left|\frac{\partial}{\partial u}C(u_{b'+1,2},F_2(\zeta_{b'+1}))-\frac{\partial}{\partial u}C(u_{b',2},F_2(\zeta_{b'}))\right|\\
&\leq\left|\frac{\partial}{\partial u}C(u_{b'+1,2},F_2(\zeta_{b'+1}))-\frac{\partial}{\partial u}C(u_{b'+1,2},F_2(\zeta_{b'}))\right|+\left|\frac{\partial}{\partial u}C(u_{b'+1,2},F_2(\zeta_{b'}))-\frac{\partial}{\partial u}C(u_{b',2},F_2(\zeta_{b'}))\right|\\
&=\left|\int_{F_2(\zeta_{b'})}^{F_2(\zeta_{b'+1})}\underbrace{\frac{\partial^2}{\partial v^2}C(u_{b'+1,2},v)}_{\leq L}\mathrm{d}v\right|+\left|\int_{u_{b'+1,1}}^{u_{b',2}}\underbrace{\frac{\partial^2}{\partial u^2}C(u,F_2(\zeta_{b'}))}_{\leq L}\mathrm{d}u\right|\\
&\leq L\cdot\frac{c_0}{B}+L\Big(\frac{c_0}{B}+|\Delta_{b'+1,1}|\Big)=L\left(|\Delta_{b'+1,1}|+\frac{2c_0}{B}\right),
\end{aligned}$$

where we use Assumption C.1-(1) for the first term and $|u_{b',2}-u_{b'+1,2}|\leq\Delta_{b'+1,1}+\frac{c_0}{B}$ for the second term. Moreover, we have

$$|(\text{II})|=\left|\int_{\hat{F}_1(\zeta_{b'})}^{\hat{F}_1(\zeta_{b'+1})}\int_{F_2(\zeta_{b'})}^{\hat{F}_2(\zeta_{b'})}\underbrace{\frac{\partial^2}{\partial u\partial v}C(u,v)}_{\leq L}\mathrm{d}v\mathrm{d}u\right|\leq L\frac{2}{\ell\delta}\Big(\epsilon+L\cdot\frac{c_0}{B}\Big)|\Delta_{b',2}|\leq\frac{4c_0L^2}{B\ell\delta}|\Delta_{b',2}|,$$

where we use $\epsilon\leq\frac{c_0}{B}$ in the last inequality.

Then, (35) gives

$$\begin{aligned}
&|\Delta_{b'+1,1}-\Delta_{b',1}|\cdot\left|\frac{\partial}{\partial u}C(u_{b',1},1)-\frac{\partial}{\partial u}C(u_{b',2},F_2(\zeta_{b'}))\right|\\
&\leq\epsilon+\frac{4L}{\ell^2\Big(1-\hat{F}_1(\zeta_{b'})\Big)\Big(1-\hat{F}_2(\zeta_{b'})\Big)}\Big(\epsilon+L\cdot\frac{c_0}{B}\Big)^2+\frac{4c_0L^2}{B\ell\delta}|\Delta_{b',2}|+L|\Delta_{b'+1,1}|\left(|\Delta_{b'+1,1}|+\frac{3c_0}{B}\right).
\end{aligned}$$

Moreover, by the triangle inequality, for a sufficiently large B satisfying $L \cdot \frac{c_0}{B} \le \frac{\ell\delta}{2}$, we have

$$\begin{aligned}
&\left|\frac{\partial}{\partial u}C(u_{b',1},1) - \frac{\partial}{\partial u}C(u_{b',2},F_2(\zeta_{b'}))\right| \\
&\ge \left|\frac{\partial}{\partial u}C(u_{b',2},1) - \frac{\partial}{\partial u}C(u_{b',2},F_2(\zeta_{b'}))\right| - \left|\frac{\partial}{\partial u}C(u_{b',1},1) - \frac{\partial}{\partial u}C(u_{b',2},1)\right| \\
&= \left|\int_{F_2(\zeta_{b'})}^{1} \underbrace{\frac{\partial^2}{\partial v^2}C(u_{b',2},v)}_{\ge \ell}\,\mathrm{d}v\right| - \left|\int_{u_{b',1}}^{u_{b',2}} \underbrace{\frac{\partial^2}{\partial u^2}C(u,1)}_{\le L}\,\mathrm{d}u\right| \\
&\ge \ell(1-F_2(\zeta_{b'})) - L|u_{b',2}-u_{b',1}| \ge \ell\delta - L\cdot\frac{c_0}{B} \ge \frac{\ell\delta}{2},
\end{aligned}$$

where the third inequality follows from $F_2(\zeta_b) \le 1-\delta$ and

$$|u_{b',2}-u_{b',1}| \le \max\Big\{\hat{F}_1(\zeta_{b'}),F_1(\zeta_{b'})\Big\} - \min\Big\{\hat{F}_1(\zeta_{b'}),F_1(\zeta_{b'})\Big\} = |\Delta_{b',k}| \le \frac{c_0}{B}$$

by (26).

Then, we have

$$\begin{aligned}
&|\Delta_{b'+1,1}-\Delta_{b',1}| \\
&\le \frac{2}{\ell\delta}\left[\epsilon + \frac{4L}{\ell^2\Big(1-\hat{F}_1(\zeta_{b'})\Big)\Big(1-\hat{F}_2(\zeta_{b'})\Big)}\Big(\epsilon + L\cdot\frac{c_0}{B}\Big)^2 + \frac{4c_0L^2}{B\ell\delta}|\Delta_{b',2}| + L|\Delta_{b'+1,1}|\Big(|\Delta_{b'+1,1}| + \frac{3c_0}{B}\Big)\right] \\
&\le \frac{2}{\ell\delta}\left[\epsilon + \frac{4L}{\ell^2\Big(1-\hat{F}_1(\zeta_{b'})\Big)\Big(1-\hat{F}_2(\zeta_{b'})\Big)}\Big(\epsilon + L\cdot\frac{c_0}{B}\Big)^2 + \frac{4c_0L^2}{B\ell\delta}|\Delta_{b',2}|\right] + \frac{2L}{\ell\delta}|\Delta_{b'+1,1}|\Big(|\Delta_{b'+1,1}| + \frac{3c_0}{B}\Big).
\end{aligned}$$

We consider two cases (i) $\Delta_{b'+1,1} \le \frac{c_0}{B}$ and (ii) $\Delta_{b'+1,1} > \frac{c_0}{B}$. If (i) holds, we have

$$|\Delta_{b'+1,1}-\Delta_{b',1}| \le \frac{2}{\ell\delta}\left[\epsilon + \frac{4L}{\ell^2\Big(1-\hat{F}_1(\zeta_{b'})\Big)\Big(1-\hat{F}_2(\zeta_{b'})\Big)}\Big(\epsilon + L\cdot\frac{c_0}{B}\Big)^2 + \frac{4c_0L^2}{B\ell\delta}|\Delta_{b',2}|\right] + \frac{8c_0L}{B\ell\delta}|\Delta_{b'+1,1}|.$$

By using the triangle inequality again, we obtain

$$|\Delta_{b'+1,1}| \le |\Delta_{b',1}| + \frac{2}{\ell\delta}\left[\epsilon + \frac{4L}{\ell^2\Big(1-\hat{F}_1(\zeta_{b'})\Big)\Big(1-\hat{F}_2(\zeta_{b'})\Big)}\Big(\epsilon + L\cdot\frac{c_0}{B}\Big)^2 + \frac{4c_0L^2}{B\ell\delta}|\Delta_{b',2}|\right] + \frac{8c_0L}{B\ell\delta}|\Delta_{b'+1,1}|.$$

Finally, by rearranging the above inequality and using the induction hypothesis, we have

$$\begin{aligned}
\Big(1-\frac{8c_0L}{B\ell\delta}\Big)|\Delta_{b'+1,1}| &\le \left[\epsilon + \frac{16L\delta^{-2}}{\ell^2}\Big(\epsilon + L\cdot\frac{c_0}{B}\Big)^2\right]\cdot\left[\Big(\frac{1}{\ell}+\frac{B}{4c_0L}\Big)\Big(1-\frac{8c_0L}{B\ell\delta}\Big)^{-b'+1} - \frac{B}{4c_0L}\right] \\
&\quad + \frac{2}{\ell\delta}\left[\epsilon + \frac{4L}{\ell^2\Big(1-\hat{F}_1(\zeta_{b'})\Big)\Big(1-\hat{F}_2(\zeta_{b'})\Big)}\Big(\epsilon + L\cdot\frac{c_0}{B}\Big)^2\right] \\
&\le \left[\epsilon + \frac{16L\delta^{-2}}{\ell^2}\Big(\epsilon + L\cdot\frac{c_0}{B}\Big)^2\right]\cdot\left[\Big(\frac{1}{\ell}+\frac{B}{4c_0L}\Big)\Big(1-\frac{8c_0L}{B\ell\delta}\Big)^{-b'+1} - \frac{B}{4c_0L}\right] \\
&\quad + \frac{2}{\ell\delta}\left[\epsilon + \frac{16L\delta^{-2}}{\ell^2}\Big(\epsilon + L\cdot\frac{c_0}{B}\Big)^2\right],
\end{aligned}$$

where we use $F_k(t) < 1 - \frac{\delta}{2}$ for the last inequality, and hence,

$$\begin{aligned}
|\Delta_{b'+1,1}| &\le \left[\epsilon + \frac{16L\delta^{-2}}{\ell^2}\left(\epsilon + L\cdot\frac{c_0}{B}\right)^2\right]\cdot\left[\left(\frac{1}{\ell}+\frac{B}{4c_0L}\right)\left(1-\frac{8c_0L}{B\ell\delta}\right)^{-b'+1} - \frac{B}{4c_0L}\right]\left(1-\frac{8c_0L}{B\ell\delta}\right)^{-1} \\
&\quad + \left(1-\frac{8c_0L}{B\ell\delta}\right)^{-1}\cdot\frac{2}{\ell\delta}\left[\epsilon + \frac{16L\delta^{-2}}{\ell^2}\left(\epsilon + L\cdot\frac{c_0}{B}\right)^2\right] \\
&= \left[\epsilon + \frac{16L\delta^{-2}}{\ell^2}\left(\epsilon + L\cdot\frac{c_0}{B}\right)^2\right]\cdot\left[\left(\frac{1}{\ell}+\frac{B}{4c_0L}\right)\left(1-\frac{8c_0L}{B\ell\delta}\right)^{-(b'+1)+1} - \frac{B}{4c_0L}\right],
\end{aligned}$$

which ensures (21) for $b = b' + 1$ and $k = 1$. Since the same argument holds for $k = 2$, we obtain the conclusion for the case (i).

If (ii) holds, we have

$$|\Delta_{b'+1,1} - \Delta_{b',1}| \le \frac{2}{\ell\delta}\left[\epsilon + \frac{4L}{\ell^2\left(1-\hat{F}_1(\zeta_{b'})\right)\left(1-\hat{F}_2(\zeta_{b'})\right)}\left(\epsilon + L\cdot\frac{c_0}{B}\right)^2 + \frac{4c_0L^2}{B\ell\delta}|\Delta_{b',2}|\right] + \frac{8L}{\ell\delta}|\Delta_{b'+1,1}|^2.$$

Then, by using (26), we have $|\Delta_{b'+1,1}|^2 \lesssim B^{-2(1-\tau)}$. This and triangle inequality gives

$$\begin{aligned}
|\Delta_{b'+1,1}| &\le |\Delta_{b',1}| + \frac{2}{\ell\delta}\left[\epsilon + \frac{4L}{\ell^2\left(1-\hat{F}_1(\zeta_{b'})\right)\left(1-\hat{F}_2(\zeta_{b'})\right)}\left(\epsilon + L\cdot\frac{c_0}{B}\right)^2 + \frac{3c_0L^2}{B\ell\delta}|\Delta_{b',2}|\right] + \frac{8L}{\ell\delta}|\Delta_{b'+1,1}|^2 \\
&\le |\Delta_{b',1}| + \left(1-\frac{8c_0L}{B\ell\delta}\right)^{-1}\frac{2}{\ell\delta}\left[\epsilon + \frac{16L}{\ell^2\left(1-\hat{F}_1(\zeta_{b'})\right)\left(1-\hat{F}_2(\zeta_{b'})\right)}\left(\epsilon + L\cdot\frac{c_0}{B}\right)^2 + \frac{4c_0L^2}{B\ell\delta}|\Delta_{b',2}|\right]
\end{aligned}$$

with taking a sufficiently large B. This gives the conclusion for the case (ii). □

To obtain the bound with respect to W_1, we utilize the following lemma:

Lemma C.4 (Vallender (1974))**.** *Let μ and ν be probability measures on $\mathbb{R}$ whose CDFs are defined by F and G. Then,*

$$\int_{-\infty}^{+\infty}|F(t) - G(t)|\mathrm{d}t = W_1(\mu,\nu).$$

Then, we move to the proof of Theorem C.2.

Proof of Theorem C.2. Set δ_0 in Assumption C.1-(3) by $\delta_0 \coloneqq \frac{8c_0L}{\ell}$. By using Lemma C.3, we obtain

$$\begin{aligned}
W_1(\hat{F}_k, F_k) &= \int_{-\infty}^{+\infty}\left|\hat{F}_k(t) - F_k(t)\right|\mathrm{d}t \\
&= \int_0^T\left|\hat{F}_k(t) - F_k(t)\right|\mathrm{d}t \\
&= \underbrace{\sum_{b=1}^{b_0}\int_{\zeta_{b-1}}^{\zeta_b}\left|\hat{F}_k(t) - F_k(t)\right|\mathrm{d}t}_{\text{(I)}} + \underbrace{\int_{\zeta_{b_0}}^{\zeta_B}\left|\hat{F}_k(t) - F_k(t)\right|\mathrm{d}t}_{\text{(II)}},
\end{aligned} \tag{37}$$

where b_0 is defined by Assumption C.1-(3). Then, we bound the each term.

We denote $\Delta_{b,k} \coloneqq \hat{F}_k(\zeta_b) - F_k(\zeta_b)$ again. By using the bound (25), we have

$$|\Delta_{b,k}| \lesssim \left(\epsilon + \frac{\log B}{B^2}\right)B^{1+\tau}$$

Then, the first term can be bounded by

$$\begin{aligned}
(\mathrm{I}) &\leq \sum_{b=1}^{b_0} \int_{\zeta_{b-1}}^{\zeta_b} \max\left\{\left|\hat{F}_k(t) - F_k(\zeta_{b-1})\right|, \left|\hat{F}_k(t) - F_k(\zeta_b)\right|\right\} \mathrm{d}t \\
&\leq \sum_{b=1}^{b_0} \int_{\zeta_{b-1}}^{\zeta_b} \max\left\{|\Delta_{b,k}|, \frac{c_0}{B}\right\} \mathrm{d}t \\
&\leq \sum_{b=1}^{b_0} \int_{\zeta_{b-1}}^{\zeta_b} \left(|\Delta_{b,k}| + \frac{c_0}{B}\right) \mathrm{d}t \\
&\lesssim \sum_{b=1}^{b_0} \left(\epsilon + \frac{\log B}{B^2}\right) B^{1+\tau} \cdot \frac{\zeta_B}{B} \\
&\leq \sum_{b=1}^{B} \left(\epsilon + \frac{\log B}{B^2}\right) B^{1+\tau} \cdot \frac{\zeta_B}{B} = \left(\epsilon + \frac{\log B}{B^2}\right) B^{1+\tau} \zeta_B. \qquad (38)
\end{aligned}$$

For the second term, since $1 - F_k(\zeta_{b_0'}) \gtrsim \log^{-1} B$ by Assumption C.1-(3), we have $\left|\hat{F}_k(t) - F_k(t)\right| \lesssim \log^{-1} B$ for $t \geq \zeta_{b_0}$ with sufficiently large B. This implies

$$(\mathrm{II}) \lesssim \int_{\zeta_{b_0}}^{\zeta_B} \log^{-1} B \mathrm{d}t \leq \frac{B - b_0}{B \log B} \zeta_B. \qquad (39)$$

By substituting the bounds (38) and (39) into (37), we obtain the conclusion. □

D. Examples of Bounds on Step 1 Error

In this section, we present an example that satisfies (18). As mentioned in Sections 3 and C, any estimation method for the probability distribution can be employed, and theoretical results pertaining to those models can be leveraged to guarantee (18). Among the various methods, we introduce results derived from the Distributional Random Forest (DRF) (Ćevid et al., 2022) and histogram type estimators (Sart, 2017).

D.1. Distributional Random Forest

DRF constructs random forests designed to estimate the conditional distribution of multivariate responses. It achieves this by splitting the data using a distributional metric, specifically the maximal mean discrepancy (MMD), with the goal of maximizing the differences in distributions between child nodes. DRF then estimates targets, such as the CIF in this study, by employing a weight function that reflects how frequently the training data points end up in the same leaf as the test point across different trees.

Ćevid et al. (2022) impose the following assumptions:

(P1) *(Data Sampling.)* Instead of the traditional bootstrap sampling with replacement, commonly used in forest-based methods, a subsampling approach is employed. For each tree, a random subset of size s_n is selected from n training data points. It is assumed that s_n approaches infinity as n increases, with the rate specified below.

(P2) *(Honesty.)* The data used to construct each tree is split into two parts: one part is used for determining the splits, and the other is used for populating the leaves and thus for estimating the response.

(P3) *(α-Regularity.)* Each split leaves at least a fraction $0 < \alpha \leq 0.2$ of the available training sample on each side. Additionally, trees are grown until each leaf contains between κ and $2\kappa - 1$ observations, where $\kappa \in \mathbb{N}$ is a fixed tuning parameter.

(P4) *(Symmetry.)* The (randomized) output of a tree does not depend on the ordering of the training samples.

(P5) *(Random-Split.)* At every split point, the probability that the split occurs along the feature X_j is bounded below by π/p, for some $\pi > 0$ and for all $j = 1, \ldots, p$.

Note that each of these conditions can be verified by inspecting the constructed forest. The following proposition is direct consequence from Corollary 5 of (Ćevid et al., 2022).

Proposition D.1. *Under the assumptions (P1)-(P5), it holds that*

$$\hat{r}_{b,k|x} \xrightarrow{p} r_{b,k|x}$$

for any b and k as the sample size goes to infinity[2].

The proposition above ensures the probabilistic convergence of the estimation. Specifically, for arbitrary values of $\epsilon > 0$ and $\delta > 0$, there exists a sample size threshold $n_{\epsilon,\delta} > 0$ such that if the sample size exceeds $n_{\epsilon,\delta}$, then (18) holds with a probability of at least $1 - \delta$.

D.2. Histogram Type Estimators

Sart (2017) proposes histogram type conditional density estimators on $\mathcal{X} \times \mathcal{Y}$ by

$$\widehat{\Pr}(y|x) \coloneqq \hat{s}(x,y) = \sum_{K \in m} \frac{\sum_{i=1}^{n} \mathbb{1}_K(x_i, y_i)}{\sum_{i=1}^{n} (\delta_{x_i} \otimes \mu)(K)} \mathbb{1}_K(x,y),$$

where m is a partition of $\mathcal{X} \times \mathcal{Y}$, μ is a reference measure of the conditional density, and δ_x is the Dirac measure at $x \in \mathcal{X}$. In the context of survival analysis, we can choose $\mathcal{Y} = [0,1]^K$ equipped with the Lebesgue measure μ and m as a set of regions defined by $(\zeta_{b-1} < t_k \leq \delta_b, \delta = k)$ for $b \in [B], k \in [K]$.

Let ν be a measure defined on $\mathcal{X}$ and

$$h(f,g) = \int_{\mathcal{X} \times \mathcal{Y}} \left(\sqrt{f(x,y)} - \sqrt{g(x,y)} \right)^2 \mathrm{d}\nu(x)\, \mathrm{d}\mu(y)$$

be the Hellinger distance. Then, Sart (2017) provides the following result:

Proposition D.2 (Proposition 2.6 of (Sart, 2017))**.** *Let s be a true conditional density. Then, there exists global constants C_1, $C_2 > 0$ such that for any $\xi > 0$,*

$$\Pr\left[h^2(s, \hat{s}) \leq \inf_{v \in V_m} h^2(s,v) + C_1 \frac{|m|}{n} + C_2 \xi \right] \geq 1 - e^{-n\xi},$$

where

$$V_m \coloneqq \left\{ \sum_{K \in m} a_K \mathbb{1}_K, \forall K \in m, a_K \geq 0 \right\}.$$

See (Sart, 2017) for specific examples of deriving the term $\inf_{v \in V_m} h^2(s,v)$ under the conditions on s and $\mathcal{X}$. The bound on the Hellinger distance implies the bound on the total variation distance $\mathrm{TV}(f,g) = \int_{\mathcal{X} \times \mathcal{Y}} |f(x,y) - g(x,y)|\, \mathrm{d}\nu(x)\, \mathrm{d}\mu(y)$ as

$$\mathrm{TV}(f,g) \leq h(f,g),$$

which follows from the inequality between the L_1-norm and the L_2-norm. Thus, by utilizing Proposition D.2, we obtain (18) by taking ϵ as the total variation distance between $\hat{r}$ and r.

E. Upper and Lower Bounds Based on Cumulative Incidence Function

If the CIF $V_k(t|x)$ (as defined in (4)) is available, the upper and lower bounds of $F_k(t|x)$ can be easily computed by

$$\begin{aligned} & \Pr(T \leq t, \delta = k|x) \leq F_k(t|x) \leq \Pr(T \leq t|x) \\ \Leftrightarrow \quad & V_k(t|x) \leq F_k(t|x) \leq \sum_k V_k(t|x). \end{aligned}$$

[2] $\xrightarrow{p}$ denotes the probability convergence.

Table 2. Real datasets used in our experiments

Name	K	N	# categorical	# numuerical	censored	max. time
dataDIVAT1	2	5943	3	2	83.6%	6225
oldmort	2	6495	5	2	69.7%	20
Dialysis	2	6805	2	2	76.4%	44
flchain	2	7874	4	6	72.5%	5215
support2	2	9105	11	24	31.9%	2029
prostateSurvival	2	14294	3	0	94.4%	119
PBC	3	312	5	12	45.8%	15
Framingham	3	4434	10	9	56.2%	8767

By averaging over $x \sim X$, we can derive the same upper and lower bounds as in (Peterson, 1976):

$$\mathop{\mathbb{E}}_{x\sim X}[V_k(t|x)] \leq F_k(t) \leq \mathop{\mathbb{E}}_{x\sim X}\left[\sum_k V_k(t|x)\right]. \tag{40}$$

Given a dataset $\mathcal{D} = \{(x^{(i)}, t^{(i)}, \delta^{(i)})\}_{i=1}^N$, we can compute the empirical estimates of the expectations in (40) as follows:

$$\mathop{\mathbb{E}}_{x\sim X}[V_k(t|x)] \approx \frac{1}{N} \sum_{(x^{(i)},t^{(i)},\delta^{(i)})\in\mathcal{D}} \mathbb{1}_{t^{(i)}\leq t, \delta^{(i)}=k}$$

$$\mathop{\mathbb{E}}_{x\sim X}\left[\sum_k V_k(t|x)\right] \approx \frac{1}{N} \sum_{(x^{(i)},t^{(i)},\delta^{(i)})\in\mathcal{D}} \mathbb{1}_{t^{(i)}\leq t}.$$

Note that these values are equivalent to empirical CDFs, and they may not correspond to the actual bounds if the number of data points N is insufficient. In such cases, the confidence intervals of these empirical CDFs should also be computed using methods such as Greenwood's method (1926).

F. Additional Experiments

Datasets. We used eight datasets, summarized in Table 2, where N denotes the number of data points, and the fourth and fifth columns indicate the numbers of categorical and numerical features in the feature vectors, respectively. The six datasets with $K = 2$ were obtained from the Python package `SurvSet` (Drysdale, 2022). The Framingham (Kannel & McGee, 1979) and PBC (Therneau & Grambsch, 2000) datasets with $K = 3$ were ones used in (Jeanselme et al., 2023).

All datasets were randomly split into training (65%), validation (15%), and testing (20%) sets. The results reported in this section are the mean and standard deviation over five random splits. We divided the time horizon $[0, t_{\max}]$ into $B-1$ evenly spaced boundaries and added an additional time slot to represent times greater than $t_{\max}$, where $t_{\max}$ is the maximum observed time within the dataset. This setup means that we used B time slots divided by $\{\zeta_b\}_{b=0}^B$. We set $B = 32$ unless otherwise stated.

Models and hyperparameters. We used a multi-layer perceptron (MLP) with three hidden layers as a neural network model. The dropout layer was employed with a dropout rate of 0.5, and the ReLU function was utilized as the activation layer. The softmax function served as the output layer. The neural network was trained using the AdamWScheduleFree optimizer (Defazio et al., 2024) with early stopping. For each dataset, we performed a hyperparameter search to determine the number of neurons in the hidden layers and the learning rate of the optimizer: the number of neurons was chosen from the set $\{4, 8, 16, 32, 64, 128, 256\}$, and the learning rate was chosen from the set $\{0.00001, 0.0001, 0.001, 0.01, 0.1, 1.0, 10.0\}$.

For TS-RF and TS-DRF models, hyperparameter searches were performed on these parameters: `n_estimators` was chosen from the integers between 100 and 1000, `max_depth` was chosen from the integers between 10 and 50, `min_samples_split` was chosen from the integers between 2 and 64, `min_samples_leaf` was chosen from the integers between 1 and 32, `max_features` was chosen from `sqrt` or `log2`, `criterion` was chosen from `log_loss`,

`gini`, or `entropy`, `splitting_rule` was chosen from `CART` or `FourierMMD`, `num_features` was chosen from the integers between 1 and 100, `sample_fraction` was chosen from the numbers between 0.1 and 0.5, `min_node_size` was chosen from the integers between 1 and 10, and `alpha` was chosen from the numbers between 0.01 and 0.3.

For TS-LGB model, hyperparameter search was performed on these parameters: `n_estimators` was chosen from the integers between 30 and 300, `max_depth` was chosen from the integers between 1 and 64, `min_child_samples` was chosen from the integers between 1 and 32, `num_leaves` was chosen from the integers between 2 and 64, `criterion` was chosen from `log_loss`, `gini`, or `entropy`, `learning_rate` was chosen from the numbers between 0.001 and 0.01, `lambda_l1` was chosen from the numbers between 10^{-8} and 1.0, and `lambda_l2` was chosen from the numbers between 10^{-8}.

Prediction performance on the other datasets. To complement the results in Sec. 6 for the Dialysis and oldmort datasets, we conducted experiments using the dataDIVAT1, flchain, prostateSurvival, and support datasets with $K = 2$. The outcomes of these experiments are presented in Fig. 6. Unlike the Dialysis and oldmort datasets, no single model consistently outperformed the others across different metrics and datasets.

We also evaluated the prediction performance using the Framingham and PBC datasets with $K = 3$. As baseline methods, we compared our models with those utilized in Jeanselme et al. (2023). Specifically, we compared against: DeepHit (Lee et al., 2018), Deep Survival Machines (DSM) (Nagpal et al., 2021), DeSurv (Danks & Yau, 2022), and Neural Fine-Gray (NeuralFG) (Jeanselme et al., 2023), which is a neural network model extending the Fine-Gray model (Fine & Gray, 1999). For these models, we used the implementations available at `https://github.com/Jeanselme/NeuralFineGray/` under MIT license, and performed hyperparameter searches based on the guidelines provided in the source code. We compared our models with DeepHit, DSM, DeSurv, and NeuralFG models using the independence copula, and the results are displayed in Fig. 7. These results demonstrate that our two-step algorithm is competitive with these baseline models.

Ablation study on hyperparameter B. We conducted an ablation study on the hyperparameter B in our two-step algorithm. This study aimed to evaluate the prediction performance on the cen-log metric using the TS-LGB model with the parameters $w_1 \in \{0, 0.5, 1\}$ and $B \in \{4, 8, 16, 32, 64\}$, where w_1 is the parameter for the primary event of interest. In this study, we also implemented the method proposed in (Carrière, 1995), which is labeled as 'middle'.

Figure 8 shows the results, where the prediction performances are normalized relative to those with $w_1 = 0.5$ for each B. The results indicate that prediction performances varied significantly depending on the choice of the parameter w_1 on several datasets when B was small. However, these differences diminished for $B \geq 32$. Regarding the method proposed in (Carrière, 1995), while it is valid only if $B \to \infty$ in theory, somewhat surprisingly, the results demonstrate that it performed comparable to our algorithm even for small B in practice.

Upper and lower bounds estimation. We illustrate survival functions along with bounds for the six datasets with $K = 2$ in Fig. 9. The six graphs on the left depict average survival functions. In these graphs, we employed the Kaplan-Meier (KM) estimator (1958) and the copula-graphic (CG) estimator (Zheng & Klein, 1995) using the Frank copula (2) with parameters $\theta = -5$ and $\theta = 5$. Recall that the CG estimator is a generalization of the KM estimator. The shaded regions indicate the bounds enclosed by the upper and lower limits (Peterson, 1976). The twelve graphs in the center and right display individual survival functions. To estimate these functions, we used the TS-LGB model combined with the independence copula and the Frank copula with parameters $\theta = -5$ and $\theta = 5$. The shaded regions represent the bounds enclosed by the upper and lower limits given in Inequality (15).

The figures displaying the average survival functions indicate that uncertainty due to the unknown copula increases as time progresses. The figures displaying the individual survival functions also show similar uncertainty, but the degree of uncertainty varies by individual. In particular, the right figures for the flchain and support2 datasets showed small uncertainty, even without prior knowledge of the copula. We also see that the estimated uncertainty due to the unknown copula can be narrowed by using the estimated survival functions based on the Frank copula.

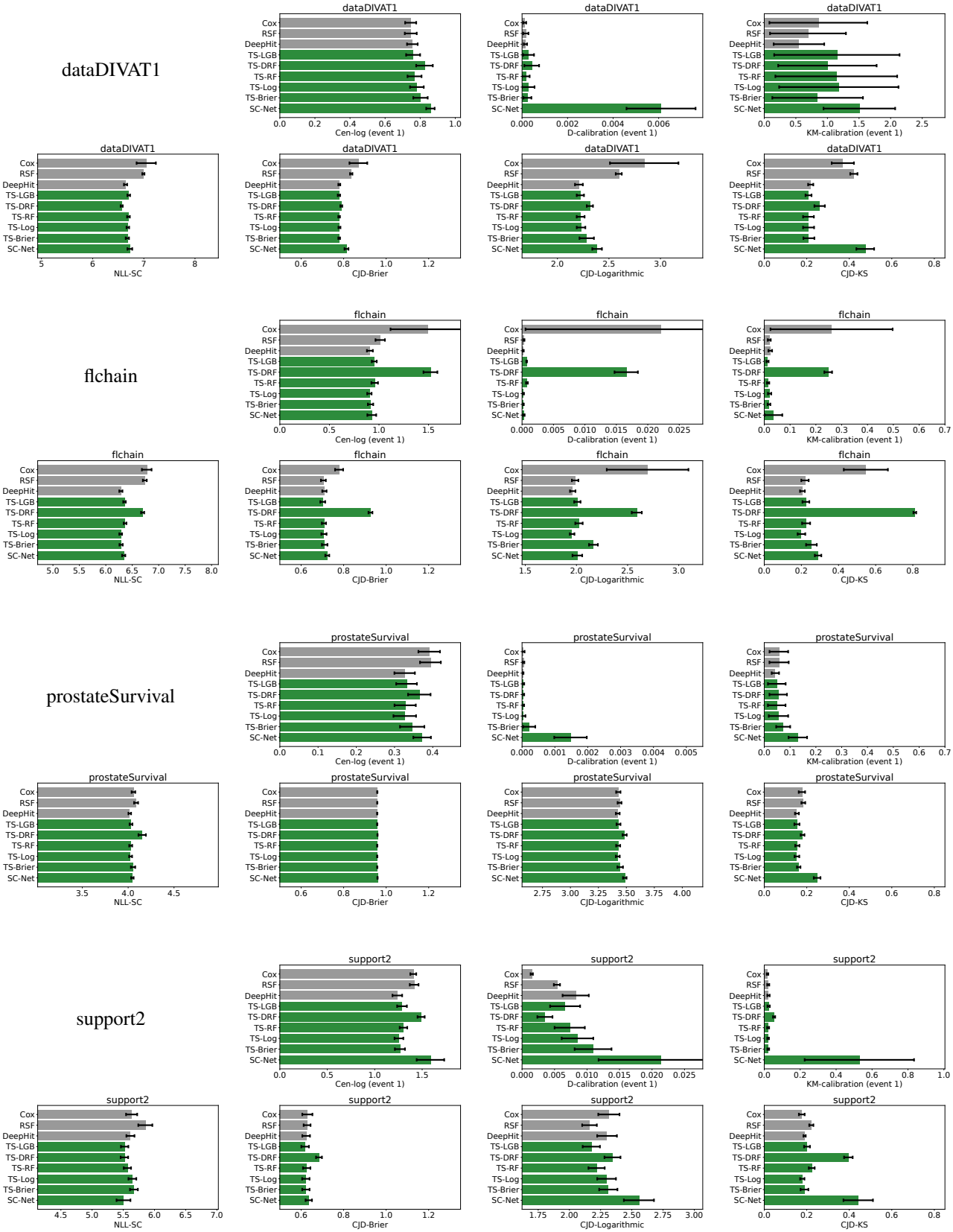


Figure 6. Prediction performance comparison on dataDIVAT1, flchain, prostateSurvival, and support2 datasets with various metrics (lower is better).

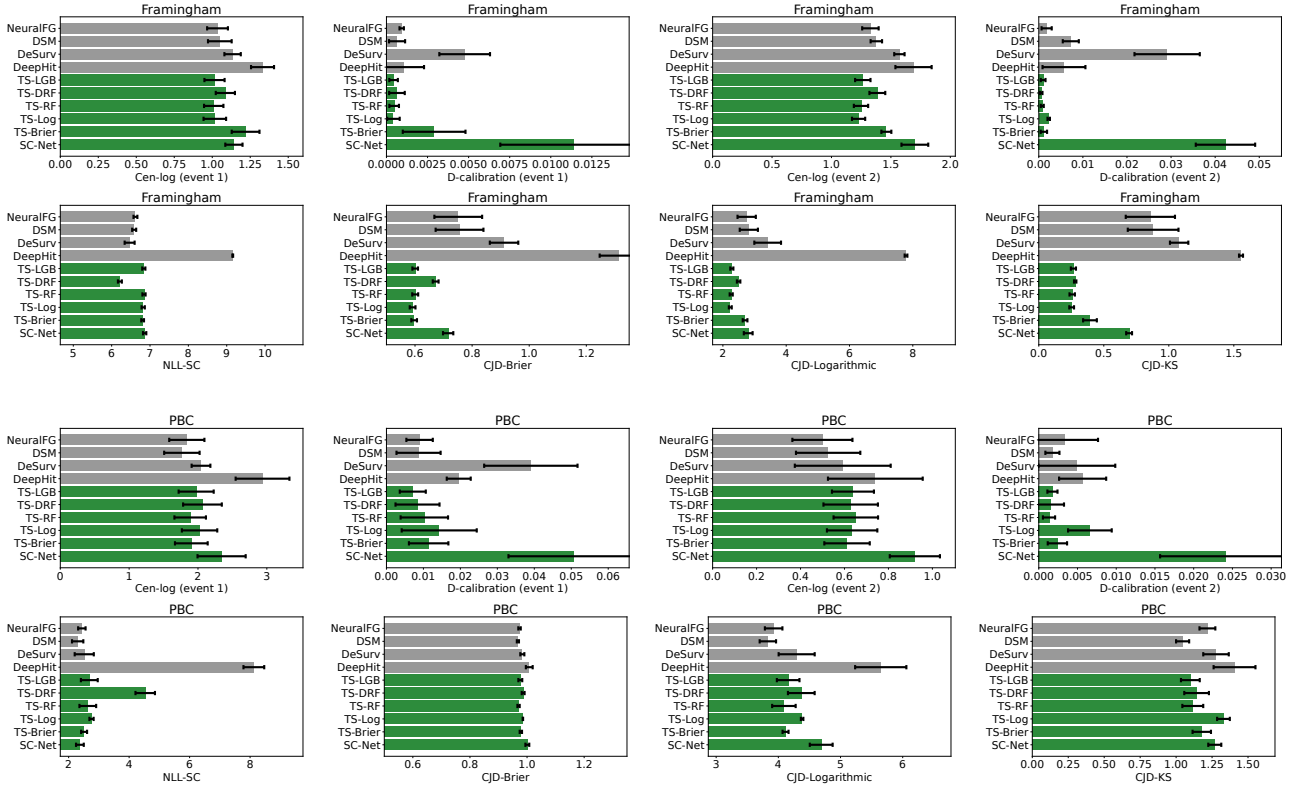


Figure 7. Prediction performance comparison on Framingham and PBC datasets with various metrics (lower is better).

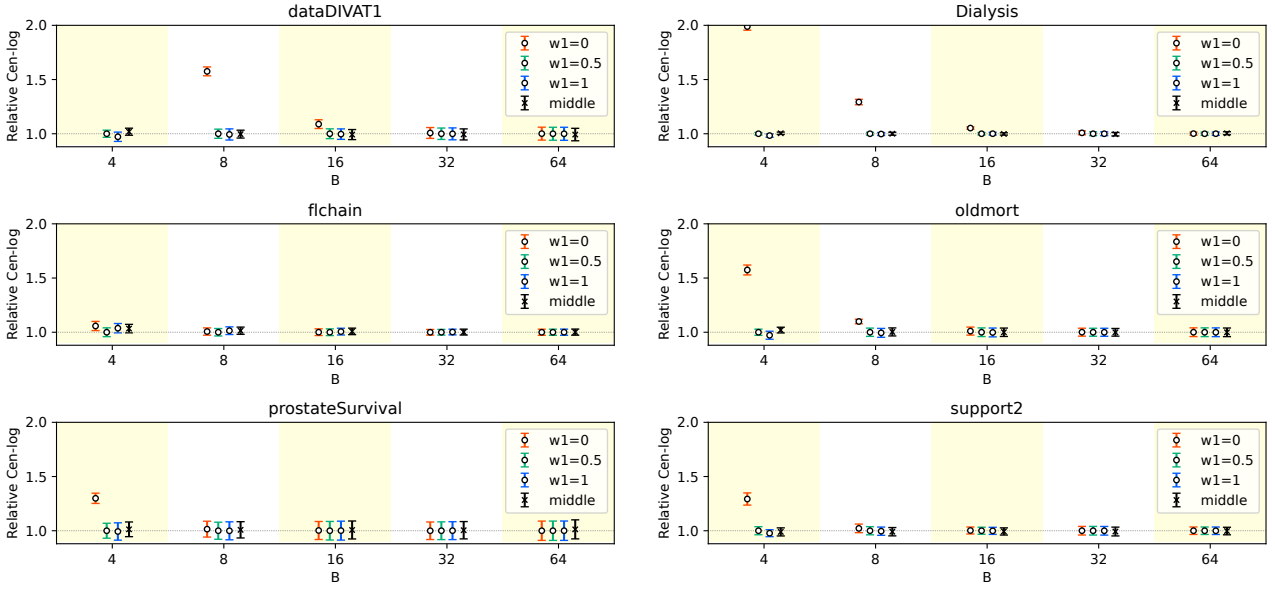


Figure 8. Ablation study on hyperparameter B on the six datasets (lower is better).

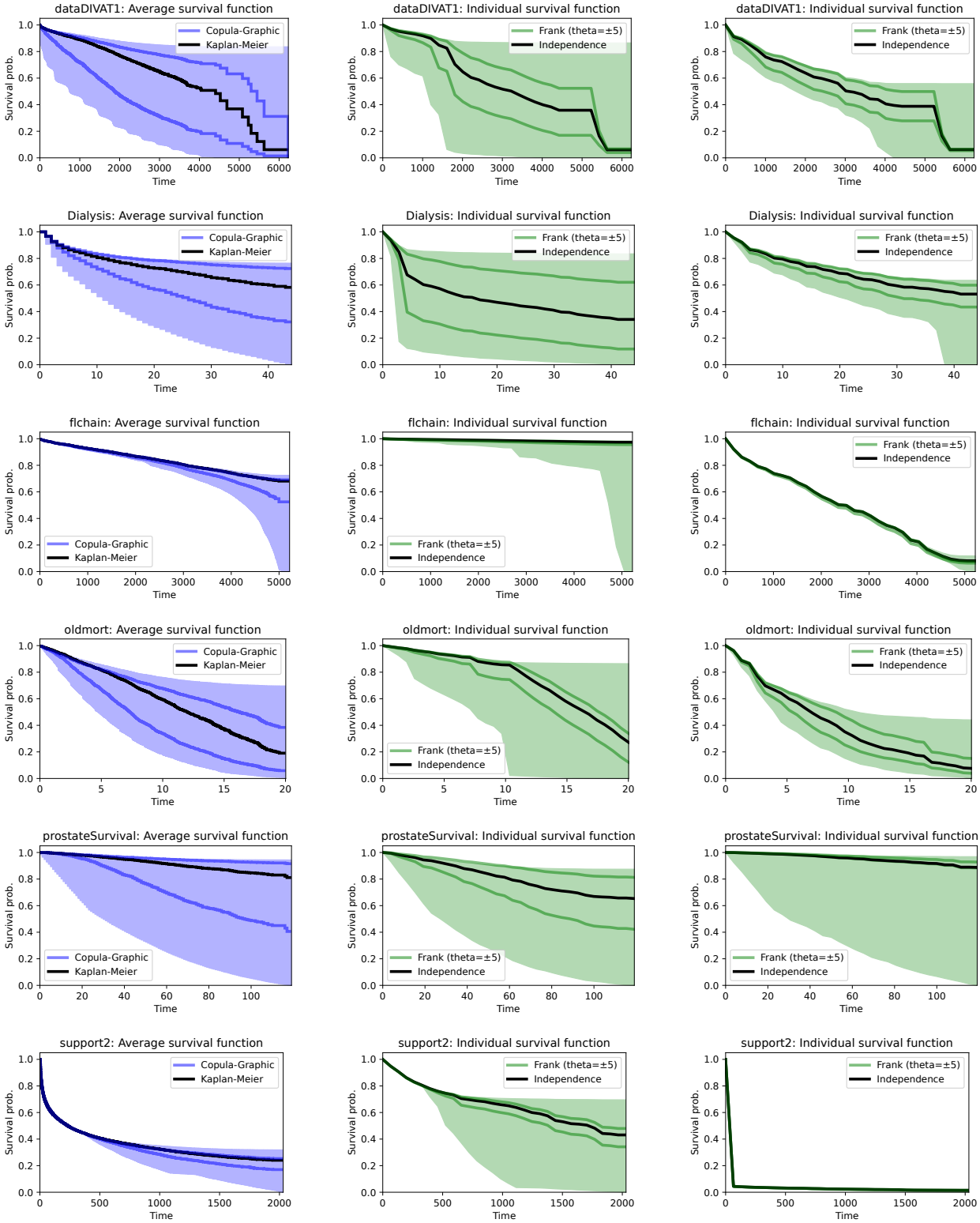


Figure 9. Estimated survival functions with upper and lower bounds for the six datasets with $K = 2$. The left graphs show average survival functions, while the graphs in the center and the right show arbitrary chosen individual survival functions. In these graphs, the shaded region corresponds to the upper and lower bounds of survival functions accounting for the uncertainty arising from the lack of knowledge about the copula.