
Overcoming Dependent Censoring in the Evaluation of Survival Models

Christian Marius Lillelund^{1,2}

Shi-ang Qi^{1,3,4}

Russell Greiner^{1,3}

¹Department of Computing Science, University of Alberta, Edmonton, Canada

²Department of Electrical and Computer Engineering, Aarhus University, Aarhus, Denmark

³Alberta Machine Intelligence Institute, Edmonton, Canada

⁴Vector Institute, Toronto, Canada

Abstract

Dependent censoring occurs when the event time and censoring time are not conditionally independent given the observed covariates. This complicates survival model evaluation because widely used metrics, such as the Brier score, typically handle right-censoring using inverse probability of censoring weighting (IPCW). Unfortunately, IPCW is valid only when the estimated censoring distribution is independent of the event time. We propose a dependent Brier score based on an Archimedean copula and the Copula-Graphic estimator, and establish consistency and asymptotic normality of its margin-time estimator. To evaluate the metric, we introduce a semi-synthetic framework that creates realistic dependent censoring while preserving the original covariate structure and known event times. Across 12 datasets, the proposed metric reduces estimation error by 12-16% on average relative to IPCW. Source code is available at <https://github.com/thecml/DependentEVAL>.

1 INTRODUCTION

Survival analysis differs from standard regression because outcomes may be *censored*: for some individuals, we only know that the event time exceeds a certain specified value. A fundamental assumption underlying most methods and evaluation metrics is *independent censoring*, meaning that finding someone censored at time c does not convey information about their true survival time e , beyond $e > c$ [Kleinbaum and Klein, 2012]. Censoring is considered *dependent* whenever the censoring time C and the event time E are not conditionally independent given the observed covariates X – i.e., $E \not\perp C \mid X$ (see Appendix A for details). While dependent censoring has been relatively well studied in the context of *model estimation*, with several methods

proposed to address it [Foomani et al., 2023, Zhang et al., 2024, Liu et al., 2025], its impact on *model evaluation* remains largely unaddressed. In particular, the integrated Brier score (IBS) [Graf et al., 1999] is typically implemented via inverse probability of censoring weighting (IPCW) based on the Kaplan-Meier (KM) [Kaplan and Meier, 1958] estimator or a Cox proportional hazards (CoxPH) model [Cox, 1972]. These approaches assume independence between event and censoring times. When this assumption is violated, the IPCW weights are misspecified, potentially biasing performance estimates and distorting comparisons between survival models.

Figure 1a illustrates how the independent censoring assumption fails when a confounding covariate – such as “Tumor grade” – is not observed. Because tumor grade affects both relapse (E) and dropout (C), advanced tumor grades lead to shorter event and censoring times, inducing dependence. Standard KM-based evaluation relies on independent censoring, assuming that censored individuals are like those remaining at risk. As shown in Figure 1b, when a patient i is censored before the event for reasons related to the event, the KM estimator may overestimate survival beyond c_i . Because this estimated survival distribution is used to construct the IPCW weights, the distortion produces misspecified weights and, potentially, biased performance estimates.

Contribution. We introduce a dependent Brier score and its integrated version, IBS-Dep, for evaluating survival models under dependent censoring. Our approach models the joint distribution of event and censoring times and uses the Copula-Graphic (CG) estimator to impute (marginal) event times for censored instances. Our theory establishes consistency and asymptotic normality of the CG-based margin-time estimator; consequently, IBS-Dep is consistent for the corresponding margin-time surrogate Brier risk. We evaluate how well this surrogate approximates the oracle IBS based on uncensored event times in a semi-synthetic framework, and find that IBS-Dep reduces estimation error compared to IBS using IPCW. To our knowledge, this is the first metric for survival analysis that supports dependent censoring.

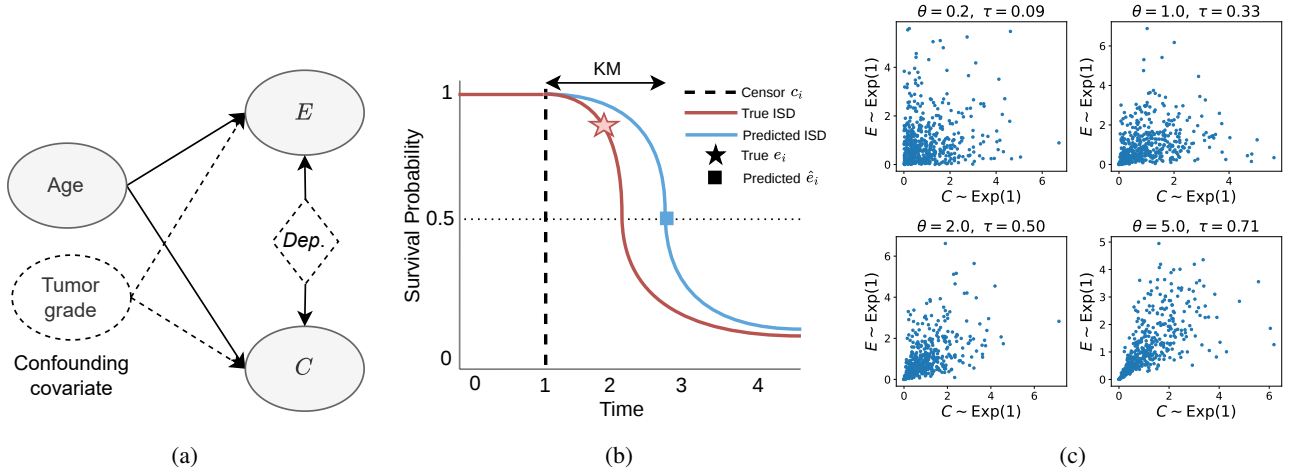


Figure 1: (a) Residual dependence between event time E (cancer relapse) and censoring time C (dropout) denoted by the diamond “Dep.”, due to an unobserved confounding covariate – here “Tumor grade”. (b) Evaluation under dependent censoring, where a patient i is censored at time c_i just prior to the event, e_i , due to reasons related to the event (but unobserved). This induces bias in the predicted individual survival distribution and the predicted event time $\hat{e}_i$ produced by the KM estimator, because the censored patient is assumed to be similar to those still at risk. Note that the true and KM-estimated survival curves coincide on the interval $[0, c_i]$ – here $c_i = 1$ – but diverge beyond due to dependent censoring. (c) Scatter plots of (e_i, c_i) , $i = 1, \dots, 500$, simulated from a Clayton copula [Clayton, 1978] with parameter θ (reported via the corresponding Kendall’s τ) and standard exponential marginals. Samples are generated by drawing $(u_{1i}, u_{2i}) \sim C_\theta$ and applying the inverse exponential CDF. The dependence structure is induced by the Archimedean generator φ_θ , where $C_\theta(u_1, u_2) = \varphi_\theta^{-1}(\varphi_\theta(u_1) + \varphi_\theta(u_2))$. This is the same generator function that appears in the CG estimator in Definition 2.2.

2 BACKGROUND AND RELATED WORK

2.1 SURVIVAL DATA

We consider a survival dataset $\mathcal{D} = \{(\mathbf{x}_i, t_i, \delta_i)\}_{i=1}^N$, where for each instance i , $\mathbf{x}_i \in \mathbb{R}^d$ denotes the covariates, $t_i \in \mathbb{R}_+$ denotes the observed time (either event or censoring), and $\delta_i \in \{0, 1\}$ denotes the event indicator. Each instance has latent event and censoring times, e_i and c_i , with $t_i = \min(e_i, c_i)$ and $\delta_i = \mathbf{1}\{e_i \leq c_i\}$. At the population-level, let E and C denote non-negative random variables for the true event and censoring times, with cumulative distribution functions $F_E(t) = \Pr(E \leq t)$ and $F_C(t) = \Pr(C \leq t)$ and survival functions $S_E(t) = 1 - F_E(t)$ and $S_C(t) = 1 - F_C(t)$. A common nonparametric estimator of $S_E(t)$ is the KM estimator.

Definition 2.1. Let $u_1 < \dots < u_J$ denote the distinct observed event times, and let $t_{\max} = \max_{1 \leq i \leq N} t_i$. For each event time u_j , define $d_j = \sum_{i=1}^N \mathbf{1}[t_i = u_j, \delta_i = 1]$ as the number of events at time u_j , and $n_j = \sum_{i=1}^N \mathbf{1}[t_i \geq u_j]$ as the number of instances at risk immediately before u_j . The Kaplan-Meier (KM) estimator is

$$\hat{S}_E(t) = \prod_{u_j \leq t} \left(1 - \frac{d_j}{n_j}\right), \quad 0 \leq t \leq t_{\max}. \quad (2.1)$$

The empty product is defined as one, and $\hat{S}_E(t)$ is undefined for $t > t_{\max}$. See Appendix D.1 for the derivation.

2.2 COPULA-BASED ESTIMATION

Zheng and Klein [1995] generalized the KM estimator to the CG estimator, which relaxes its independent censoring assumption by modeling dependence through a copula. A copula links two marginal distributions into a joint distribution and can thereby capture dependence between E and C . Rivest and Wells [2001] derived the first explicit form of the CG estimator under an *assumed copula* from the Archimedean family (e.g., the Clayton [Clayton, 1978] and Frank [Frank, 1979] copulas), which specifies dependence through a single parameter θ via a generator function φ_θ [Emura and Chen, 2018] (see details in Appendix B). Figure 1c shows sampling from a Clayton copula, which can model *lower-tail dependence* – that is, a tendency for early events (small values of E) to occur alongside early censoring (small values of C). Now, assuming an Archimedean copula, we can estimate $\hat{S}_E(t)$ using the CG estimator.

Definition 2.2. Let $u_1 < \dots < u_J$ denote the distinct observed event times, and let $t_{\max} = \max_{1 \leq i \leq N} t_i$. The Copula-Graphic (CG) estimator is

$$\hat{S}_E(t) = \varphi_\theta^{-1} \left(\sum_{u_j \leq t} \left[\varphi_\theta \left(\frac{n_j - d_j}{N} \right) - \varphi_\theta \left(\frac{n_j}{N} \right) \right] \right). \quad (2.2)$$

For $0 \leq t \leq t_{\max}$, where φ_θ is an Archimedean copula generator, $d_j = \sum_{i=1}^N \mathbf{1}[t_i = u_j, \delta_i = 1]$ is the number of

events at u_j , and $n_j = \sum_{i=1}^N \mathbb{1}[t_i \geq u_j]$ is the number at risk immediately before u_j . The empty sum is zero, and $\hat{S}_E(t)$ is undefined for $t > t_{\max}$ [Emura and Chen, 2018]. For independent copula $\varphi_{\text{ind}}(u) = -\log u$, the CG estimator reduces to the KM estimator. See Appendix D.2 for the derivation.

2.3 IPCW-WEIGHTED BRIER SCORE

The Brier score (BS) measures the squared error between the predicted survival probability and the observed binary outcome at a time t^* [Brier, 1950], which for uncensored instances, is:

$$\text{BS}(\mathcal{D}; \hat{S}_E(t^* | \cdot)) = \frac{1}{N} \sum_i (\mathbb{1}[t_i > t^*] - \hat{S}_E(t^* | \mathbf{x}_i))^2. \quad (2.3)$$

To incorporate censoring, Graf et al. [1999] used inverse probability of censoring weights (IPCW), where $\hat{S}_C(t)$ is typically estimated using the marginal KM estimator in standard implementations [Qi et al., 2023b]¹:

$$w_i(t^*) = \frac{\delta_i \mathbb{1}[t_i \leq t^*]}{\hat{S}_C(t_i)} + \frac{\mathbb{1}[t_i > t^*]}{\hat{S}_C(t^*)}, \quad (2.4)$$

where the first term applies to events before t^* and the second to those surviving beyond. Individuals censored before t^* receive weight zero and therefore do not contribute to the sum. Then, the BS-IPCW metric can be expressed as:

$$\begin{aligned} \text{BS}_{\text{IPCW}}(\mathcal{D}; \hat{S}_E(t^* | \cdot)) \\ = \frac{1}{N} \sum_{i=1}^N w_i(t^*) (\mathbb{1}[t_i > t^*] - \hat{S}_E(t^* | \mathbf{x}_i))^2. \end{aligned} \quad (2.5)$$

Often, practitioners want to summarize Equation 2.5 over a time interval. Therefore, Graf et al. [1999] proposed the integrated Brier score (IBS), defined as the uniform time average of the Brier score over a prespecified interval $[t_1, t_2]$:

$$\text{IBS}(\mathcal{D}; \hat{S}_E, t_1, t_2) = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \text{BS}_{\text{IPCW}}(\hat{S}_E(\tau | \cdot)) d\tau. \quad (2.6)$$

A common choice is $[t_1, t_2] = [0, t_{\max}]$, where $t_{\max}$ is the maximum observed time in the dataset.

The IPCW-weighted Brier score depends on estimating the censoring distribution. When $\hat{S}_C(t)$ is obtained via the KM, its consistency relies on the independent censoring assumption. Under dependent censoring, KM can be biased and inconsistent [Emura and Chen, 2018, Campigotto and Weller, 2014]. Conditional IPCW estimators [Gerds and Schumacher, 2006] attempt to address this by estimating $\hat{S}_C(t | X)$, but two limitations remain: (1) *model misspecification*, which may give unreliable weights, and (2) *residual dependence*, where unobserved factors induce dependence not captured by the conditional model.

¹More generally, it may use a conditional estimator $\hat{S}_C(t | X)$.

3 THE DEPENDENT BRIER SCORE

3.1 DEFINITION

We propose a Brier score that relaxes independent censoring by modeling dependence via an Archimedean copula. Our method consists of two parts: (1) the CG estimator to impute *margin time* estimates for censored instances, and (2) an *uncertainty weighting* mechanism.

The margin time is the conditional expectation of the event time given that the event time is greater than the censoring time. Given that an instance is censored at time t_i , we can calculate its margin time using [Haider et al., 2020]:

$$e^{\text{margin}}(S_E, t_i) = \mathbb{E}[E | E > t_i] = t_i + \frac{\int_{t_i}^{\infty} S_E(t) dt}{S_E(t_i)}. \quad (3.1)$$

Because S_E is unknown, we estimate the margin time using $e^{\text{margin}}(\hat{S}_E^{\text{CG}}, t_i)$,² where $\hat{S}_E^{\text{CG}}$ is obtained from the CG estimator (Definition 2.2) under an Archimedean copula with parameter θ . The resulting margin time serves as an imputed event time for each censored instance. This construction requires specification of a copula family and its parameter and accounts for marginal dependence between E and C through their joint copula model.

To account for uncertainty in these imputed event times, we adaptively down-weight censored instances according to how early censoring occurs. For an instance censored at t_i , we use its estimated progress through the event-time distribution, $\hat{F}_E^{\text{CG}}(t_i) = 1 - \hat{S}_E^{\text{CG}}(t_i)$, as an uncertainty weight. We therefore define $w_i = 1$ for uncensored instances and $w_i = 1 - \hat{S}_E^{\text{CG}}(t_i)$ for censored instances. Early censoring yields a smaller w_i , whereas later censoring generally yields a larger weight. This leads to the proposed Brier score, which is free from IPCW and targets a margin-time surrogate of the Brier risk:

$$\begin{aligned} \text{BS}_{\text{Dep}}(\mathcal{D}; \hat{S}_E(t^* | \cdot)) \\ = \frac{1}{\sum_{i=1}^N w_i} \sum_{i=1}^N w_i (\mathbb{1}[\tilde{e}_i > t^*] - \hat{S}_E(t^* | \mathbf{x}_i))^2, \end{aligned} \quad (3.2)$$

where

$$\tilde{e}_i = \begin{cases} e_i, & \delta_i = 1, \\ e^{\text{margin}}(\hat{S}_E^{\text{CG}}, t_i), & \delta_i = 0. \end{cases}$$

Similarly, we define the integrated dependent Brier score by averaging Equation 3.2 over $[t_1, t_2]$:

$$\text{IBS}_{\text{Dep}}(\mathcal{D}; \hat{S}_E, t_1, t_2) = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \text{BS}_{\text{Dep}}(\hat{S}_E(\tau | \cdot)) d\tau. \quad (3.3)$$

²For brevity, we write $e^{\text{margin}}(t_i)$ when the survival estimator is fixed.

Section 3.2 establishes consistency of the CG-based margin-time estimator for its copula-induced target, which implies consistency of IBS-Dep for the associated surrogate Brier risk.

3.2 THEORETICAL ANALYSIS

We now develop a concise large-sample theory for the CG-based margin-time functional under dependent censoring [Rivest and Wells, 2001]. The theory concerns the imputed margin times used by IBS-Dep, rather than the oracle Brier score based on unobserved event times. For ease of understanding, we sketch the main ideas and implications here, and defer the formal definitions, regularity conditions, and detailed proofs to Appendix D.

First, we show that the margin time admits a direct probabilistic interpretation: it equals the conditional mean remaining event time beyond a landmark,

$$e^{\text{margin}}(c) = \mathbb{E}[E \mid E > c],$$

provided $S_E(c) > 0$ and the corresponding tail integral is finite (see Appendix D.3.1). This justifies margin time as a clinically interpretable summary of the post- c prognosis.

Second, we establish *consistency* of the plug-in estimator $\hat{e}^{\text{margin}}(c)$ for a large-sample target $e^{*,\text{margin}}(c)$ induced by the (possibly misspecified) Archimedean copula model considered by Rivest and Wells [2001]. Under uniform consistency of $\hat{S}_E$ for S_E^* on a finite horizon $[0, t_{\max})$ and mild regularity ($S_E^*(c) > 0$ and negligible tail contribution beyond $t_{\max}$), a deterministic perturbation bound shows that small uniform errors in $\hat{S}_E$ translate into small error in $\hat{e}^{\text{margin}}(c)$. When the copula model is correctly specified, $S_E^* = S_E$ and hence $e^{*,\text{margin}}(c) = e^{\text{margin}}(c)$ (see details in Appendix D.3.2).

Finally, we derive a *first-order (delta-method) expansion* and an asymptotic normal limit. Writing the margin time as a functional $\mathcal{T}_c(S)$, we obtain an exact linearization

$$\mathcal{T}_c(S + h) - \mathcal{T}_c(S) = \dot{\mathcal{T}}_{c;S}(h) + R_{c;S}(h),$$

with an explicit remainder $R_{c;S}(h)$ that is second order in the perturbation (see Appendix D.3.3). Combining this expansion with the functional central limit theorem (CLT) of Rivest and Wells [2001, Theorem 2] for $\sqrt{n}(\hat{S}_E - S_E^*)$ yields the asymptotic linear representation (converges in distribution)

$$\sqrt{n}(\hat{e}^{\text{margin}}(c) - e^{*,\text{margin}}(c)) \implies \frac{1}{S_E^*(c)} \int_c^{t_{\max}} \mathbb{G}_E(u) du - \frac{\int_c^{t_{\max}} S_E^*(u) du}{(S_E^*(c))^2} \mathbb{G}_E(c),$$

where $\mathbb{G}_E(c)$ stands for the Gaussian process. If the right-hand side is non-degenerate, this implies an $N^{-1/2}$ convergence rate and enables principled uncertainty quantification for margin time via the induced influence form.

4 SYNTHETIC EXPERIMENTS

4.1 EXPERIMENTAL SETUP

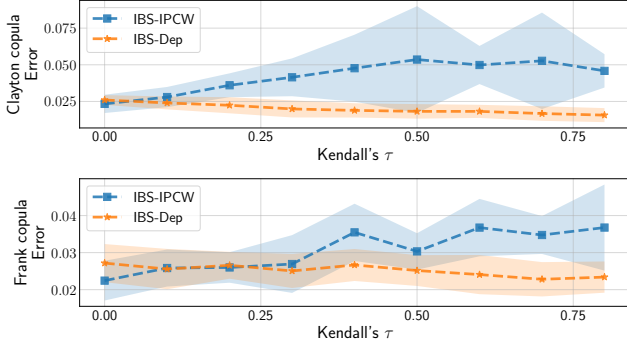
To validate our dependent Brier score, we simulate synthetic survival data with known ground-truth event times. Covariates $x_i \sim \text{Unif}([0, 1]^d)$ ($d = 10$) are fixed per seed. Event and censoring times follow covariate-dependent Weibull proportional hazards models with linear risk functions $g(x) = x^\top \beta$ and fixed shape/scale parameters (see Appendix E.1). Dependence is parameterized by Kendall's τ : when $\tau = 0$, event and censoring times are independent; for $\tau > 0$, dependence is introduced via a copula, with family and parameter assumed known to the dependent metric. For each of 10 random seeds, we generate $N = 10,000$ and use fixed training (70%) and test (30%) splits. The event times and fitted CoxPH survival model are held fixed across censoring settings, so changes in evaluation error reflect the censoring mechanism rather than changes in the prediction model. Oracle IBS is computed on the test set using the fully observed event times prior to censoring.

4.2 RESULTS AND TAKEAWAYS

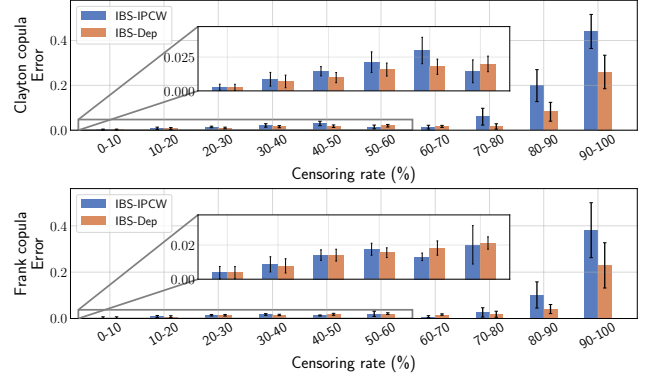
Increasing dependence. Figure 2 (a) shows the estimation error as a function of Kendall's τ at fixed censoring ($\approx 50\%$). IBS-IPCW becomes increasingly unstable as τ grows, while our IBS-Dep remains comparatively stable and achieves lower (or comparable) error. This reflects the breakdown of independent censoring: as dependence increases, IPCW becomes biased and exhibits higher variance because marginal inverse weights no longer correct the induced selection. The effect is stronger under the Clayton copula due to its lower-tail dependence, which increases the probability that early (high-risk) events coincide with early censoring and thereby more severely distorts the observable risk set. As a result, the empirical IBS-IPCW deviates increasingly from the oracle IBS computed from the fully observed data, whereas IBS-Dep tracks the oracle much more closely.

Increasing censoring. Figure 2 (b) shows error as a function of the censoring rate with $\tau = 0.5$. Both metrics deteriorate under heavy censoring, but IBS-IPCW degrades much more rapidly, particularly beyond $\sim 80\%$ censoring. As censoring increases, IPCW relies on fewer uncensored observations with larger inverse weights, inflating variance and bias under dependence. In contrast, IBS-Dep shows slower error growth, since it adjusts contributions using the joint (E, C) structure rather than relying solely on extreme reweighting.

Misspecification of C_θ . Figure 3 evaluates IBS-Dep when the assumed copula differs from the data-generating model. We consider three settings: (1) *Wrong family*, where Clayton data are evaluated with Frank, or vice versa, at the same Kendall's τ ; (2) *Wrong dependence*, where the family is correct but data generated with $\tau = 0.5$ are evaluated using



(a) Increasing dependence (Censoring = 50%).



(b) Increasing censoring (Kendall's $\tau = 0.5$).

Figure 2: Mean ($\pm$ SD) absolute estimation error of IBS-IPCW and the proposed IBS-Dep under a *known* copula, averaged over 10 experiments. (a) Error as a function of the dependence between E and C with censoring held constant. (b) Error as a function of the censoring rate with Kendall's τ held constant.

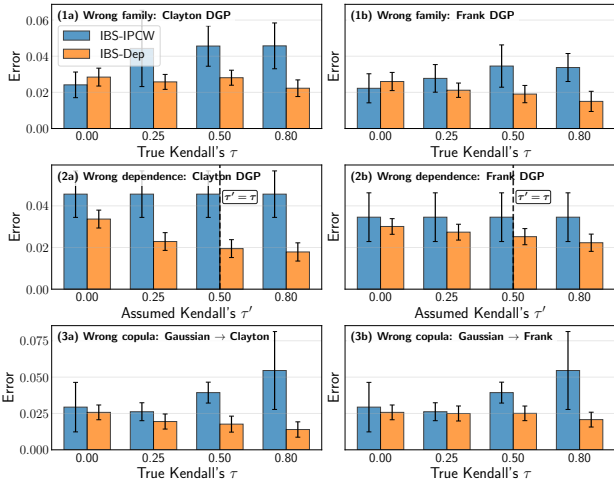


Figure 3: Mean ($\pm$ SD) absolute estimation error of IBS-IPCW and IBS-Dep under copula misspecification, averaged over 10 experiments. The rows show (respectively) wrong family, wrong dependence strength, and wrong copula experiments. In the first two rows, columns separate Clayton and Frank data-generating copulas; in the final row, the Gaussian data-generating copula is evaluated using Clayton (resp., Frank) copulas. Censoring is held fixed at 50%.

varying τ' ; and (3) *Wrong copula*, where non-Archimedean Gaussian-copula data are evaluated using Clayton or Frank at the same Kendall's τ . Across settings, IBS-Dep remains less sensitive to dependent censoring than IBS-IPCW, even with an imperfect assumed copula. However, the smallest error does not necessarily occur at the true τ : IBS-Dep is slightly upward biased, and stronger assumed dependence can reduce this bias. We therefore interpret these results as a robustness analysis rather than a parameter-recovery diagnostic.

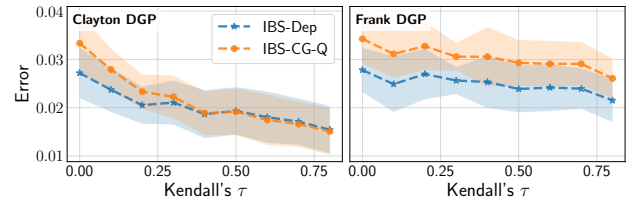


Figure 4: Mean ($\pm$ SD) absolute estimation error of IBS-Dep and IBS-CG-Q under correctly specified copulas, shown separately for Clayton and Frank and averaged over 10 experiments. Censoring is held fixed at 50%.

Integrating over censored event-time uncertainty. IBS-Dep imputes each censored individual with a single CG-based margin time. To test whether this point estimate discards useful uncertainty, we also evaluate IBS-CG-Q, an ablation that integrates over the conditional event-time distribution after censoring. For evaluation time t , define

$$\hat{q}_i^{\text{CG}}(t) = \begin{cases} \mathbb{1}[t_i > t], & \delta_i = 1, \\ \Pr_{\text{CG}}(E > t \mid E > t_i), & \delta_i = 0. \end{cases}$$

where $\Pr_{\text{CG}}$ denotes the conditional survival probability estimated by the CG estimator. The CG-Q Brier score replaces the binary survival status by $\hat{q}_i^{\text{CG}}(t)$:

$$\text{BS}_{\text{CG-Q}}(\mathcal{D}; \hat{S}_E, t) = \frac{1}{N} \sum_{i=1}^N \left[\hat{q}_i^{\text{CG}}(t) \left(1 - \hat{S}_E(t \mid \mathbf{x}_i) \right)^2 + (1 - \hat{q}_i^{\text{CG}}(t)) \hat{S}_E(t \mid \mathbf{x}_i)^2 \right]. \quad (4.1)$$

The integrated version, IBS-CG-Q, is obtained by averaging Equation 4.1 over a prespecified interval $[t_1, t_2]$, analogously to IBS-Dep. Figure 4 compares IBS-Dep and IBS-CG-Q under correctly specified copulas. Overall, IBS-Dep gives slightly lower error than its conditional counterpart, al-

though the two become nearly indistinguishable for stronger Clayton dependence. Both variants are upward biased across all dependence levels, with the largest bias at weak dependence and particularly under the Frank copula for IBS-CG-Q. As Kendall’s τ increases, the upward bias decreases, so the apparent improvement at stronger dependence reflects reduced bias rather than an intrinsically easier censoring problem.

5 SEMI-SYNTHETIC EXPERIMENTS

5.1 EXPERIMENTAL SETUP

To evaluate IBS-Dep on real data, we require ground-truth event times, which are unavailable due to censoring. Following Qi et al. [2023a], we construct semi-synthetic datasets (see Figure 5) in which event and censoring mechanisms are learned from real data and used to resample outcomes while preserving the covariate structure. Dependent censoring is induced via targeted feature selection, creating settings where censoring depends on omitted information. The resulting dataset $\mathcal{D}''$ provides known event times for oracle evaluation while keeping realistic marginal behavior.

Datasets (Step 1 in Figure 5): We apply the pipeline to 12 datasets: WHAS, METABRIC, Churn, GBSG, NACD, FLCHAIN, SUPPORT, Employee, MIMIC-IV, and 3 from SEER (death from brain, liver and stomach cancer). Table 1 summarizes the datasets; details are in Appendix E.2.

Table 1: Overview of the raw datasets. [†]Number of features before (after) one-hot encoding.

Dataset	#Instances	#Features [†]	%Censored
WHAS	500	14 (14)	57.0%
GBSG	686	8 (8)	56.4%
METABRIC	1,902	9 (9)	42.1%
Churn	1,958	12 (25)	52.4%
NACD	2,396	48 (48)	36.4%
FLCHAIN	7,871	9 (31)	72.5%
SUPPORT	8,873	14 (27)	32.0%
Employee	14,999	8 (17)	76.2%
MIMIC-IV	38,520	91 (103)	66.7%
SEER-brain	73,703	10 (10)	40.1%
SEER-liver	82,841	14 (14)	37.6%
SEER-stomach	100,360	14 (14)	43.4%

Feature selection (Step 2 in Figure 5): To induce dependent censoring, we use three strategies: (1) *Original*, keeping all features; (2) *Top-k*, selecting the most predictive features based on permutation importance [Breiman, 2001] using Harrell’s C-index [Harrell Jr. et al., 1996]; and (3) *Random-%*, selecting a random subset of the original features.

Copula model (Step 6 in Figure 5): To estimate dependence under censoring, we train Clayton and Frank copulas separately with linear Weibull proportional hazards mod-

els for the event-time and censoring-time marginals. In this semi-synthetic setting, where the true copula parameter is unknown, the dependence parameter θ is thus estimated from the data. Model parameters are learned by minimizing the penalized negative log-likelihood (NLL) under dependent censoring on the training set:

$$\begin{aligned} \ell_{\text{pen}}(\theta; \mathcal{D}) &= -\frac{1}{N} \sum_{i=1}^N \left\{ \delta_i \left[\log f_E(t_i | \mathbf{x}_i) + \log \partial_{u_1} C_\theta(u_1, u_2) \right] \right. \\ &\quad \left. + (1 - \delta_i) \left[\log f_C(t_i | \mathbf{x}_i) + \log \partial_{u_2} C_\theta(u_1, u_2) \right] \right\} \\ &\quad + \lambda \theta^2. \end{aligned} \tag{5.1}$$

Here, $u_1 = S_E(t_i | \mathbf{x}_i)$ and $u_2 = S_C(t_i | \mathbf{x}_i)$, where C_θ is an Archimedean copula and $\lambda = 0.01$ is the regularization coefficient. The full derivation of the unpenalized likelihood is in Appendix C. We select the candidate model $\hat{C}_\theta$ with the lowest validation penalized NLL. We optimize Equation 5.1 using Adam [Kingma and Ba, 2017], with separate learning rates for the Weibull marginal parameters and the copula model. The validation penalized NLL is also used for early stopping. Algorithm 1 summarizes the procedure. Only the selected copula family and parameter are used when computing IBS-Dep. See Appendix E.4 for further implementation details.

Survival learners (Step 7 in Figure 5): We compare predictions from five survival learners: CoxPH [Cox, 1972], Gradient Boosting Survival Analysis (GBSA) [Ridgeway, 1999], Random Survival Forests (RSF) [Ishwaran et al., 2008], DeepSurv [Katzman et al., 2018] and Multi-Task Logistic Regression (MTLR) [Yu et al., 2011]. These learners are trained to predict survival curves $\hat{S}_E(t | \mathbf{x}_i)$ for each instance i in the test set. See Appendix E.3 for implementation details and chosen hyperparameters.

We repeat the evaluation pipeline in Figure 5 over 10 independent experiments. In each experiment, the semi-synthetic dataset $\mathcal{D}''$ is split into stratified training (70%), validation (10%), and test (20%) sets (seeds 0–9). Missing values are imputed using training-set statistics (mean for continuous, mode for categorical variables). Numerical features are z -score normalized and categorical features one-hot encoded. To limit computational cost, large datasets ($N \geq 10,000$) are subsampled after preprocessing (Appendix E.4). As a sanity check, we plot empirical KM curves and event/censoring time distributions (see Figure E.1).

5.2 RESULTS AND TAKEAWAYS

Estimation error. Table 2 reports the mean estimation error on the semi-synthetic datasets, measured as the absolute deviation from the oracle IBS computed using the true event times. On average, the CG-based estimator reduces error

by 4-8%, while the uncertainty-weighted variant achieves the largest reductions, at 12-16%. The relative improve-

ment becomes more pronounced when important features are omitted (e.g., Random 25%). This does not imply that

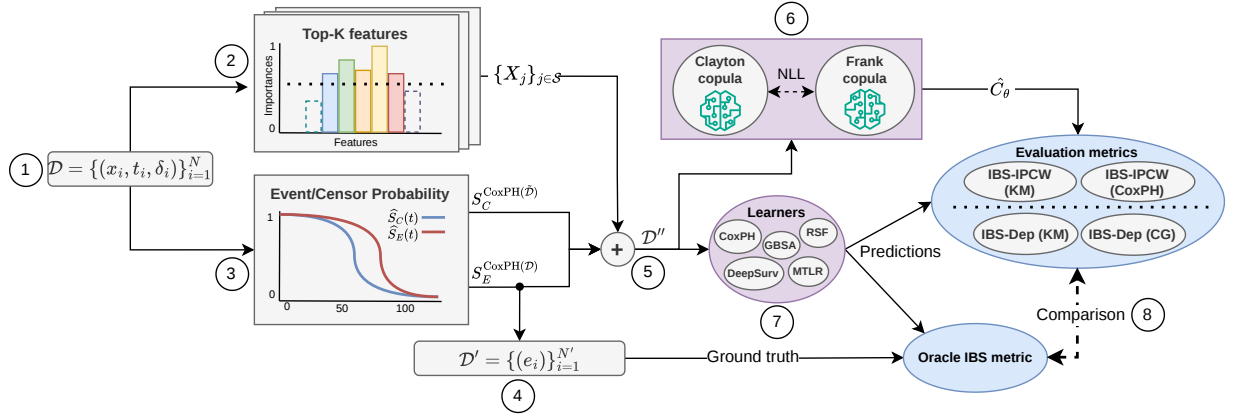


Figure 5: Evaluation pipeline. (1) We begin with a real-world dataset $\mathcal{D}$. (2) A feature selection procedure (e.g., top- k selection) yields a reduced feature set. In parallel, (3) two CoxPH models are fitted to estimate the event- and censoring-time distributions, and (4) uncensored event times are sampled from the event distribution and stored as $\mathcal{D}'$ for oracle evaluation. (5) We construct a semi-synthetic dataset $\mathcal{D}''$ by combining the selected features with event and censoring times sampled from the fitted distributions. Next, (6) two Archimedean copula models are trained to estimate $\hat{C}_\theta$, and (7) five survival learners are trained on $\mathcal{D}''$. Finally, (8) test-set predictions are evaluated using IBS-IPCW with KM or CoxPH censoring estimates and IBS-Dep with KM- or CG-based dependence estimates using $\hat{C}_\theta$, and compared with the oracle IBS computed from $\mathcal{D}'$.

Table 2: Mean absolute estimation error of independent and proposed dependent IBS-based metrics across 12 datasets and 4 feature selection strategies. Results are averaged over 5 learners and 10 experiments. Relative (%) change is shown against the KM-based IBS-IPCW baseline. *IBS-IPCW (CoxPH)* denotes IPCW with a CoxPH censoring model, corresponding to conditional independent censoring. *Dep (KM)* denotes the proposed metric using KM instead of CG as an ablation, and *UW* denotes uncertainty weighting (see Section 3.1). Full results, including variances, are provided in Appendix F.2.

	Metric	WHAS	METABRIC	Chum	GBSG	NACD	FLCHAIN	SUPPORT	Employee	MIMIC (IV)	SEER (brain)	SEER (liver)	SEER (stom.)	Average
Original	IPCW (KM)	0.073	0.061	0.117	0.125	0.092	0.156	0.108	0.152	0.026	0.077	0.104	0.119	0.101
	IPCW (CoxPH)	0.106	0.051	0.115	0.129	0.148	0.162	0.109	0.100	0.039	0.055	0.097	0.095	0.100 (-0.4%)
	Dep (KM)	0.055	0.047	0.120	0.130	0.091	0.135	0.111	0.148	0.023	0.089	0.118	0.125	0.099 (-1.4%)
	Dep (CG)	0.055	0.047	0.120	0.126	0.091	0.135	0.109	0.154	0.024	0.084	0.106	0.112	0.097 (-4.0%)
	Dep (CG) ^{UW}	0.063	0.049	0.124	0.134	0.082	0.090	0.093	0.152	0.024	0.067	0.092	0.097	0.089 (-11.8%)
Top-5	IPCW (KM)	0.071	0.062	0.121	0.124	0.103	0.133	0.109	0.155	0.027	0.076	0.109	0.124	0.101
	IPCW (CoxPH)	0.065	0.061	0.114	0.126	0.098	0.138	0.109	0.103	0.027	0.083	0.109	0.128	0.097 (-4.4%)
	Dep (KM)	0.055	0.048	0.120	0.130	0.098	0.108	0.112	0.149	0.026	0.090	0.120	0.126	0.098 (-2.8%)
	Dep (CG)	0.055	0.048	0.119	0.126	0.097	0.108	0.109	0.156	0.027	0.085	0.111	0.116	0.096 (-4.9%)
	Dep (CG) ^{UW}	0.065	0.049	0.125	0.134	0.089	0.066	0.094	0.154	0.025	0.068	0.096	0.101	0.089 (-12.2%)
Top-10	IPCW (KM)	0.071	0.061	0.119	0.125	0.098	0.143	0.109	0.152	0.026	0.077	0.103	0.119	0.100
	IPCW (CoxPH)	0.082	0.051	0.111	0.130	0.093	0.150	0.109	0.101	0.031	0.055	0.097	0.099	0.092 (-7.9%)
	Dep (KM)	0.054	0.047	0.121	0.131	0.095	0.123	0.111	0.147	0.023	0.090	0.118	0.124	0.099 (-1.4%)
	Dep (CG)	0.054	0.047	0.121	0.126	0.095	0.123	0.109	0.153	0.024	0.085	0.107	0.111	0.096 (-4.0%)
	Dep (CG) ^{UW}	0.063	0.049	0.125	0.134	0.086	0.078	0.093	0.152	0.025	0.068	0.093	0.097	0.088 (-11.7%)
Rand. 25%	IPCW (KM)	0.090	0.061	0.129	0.153	0.109	0.140	0.142	0.161	0.016	0.111	0.128	0.132	0.114
	IPCW (CoxPH)	0.091	0.056	0.122	0.157	0.111	0.147	0.142	0.110	0.017	0.111	0.197	0.133	0.116 (+1.6%)
	Dep (KM)	0.081	0.049	0.126	0.153	0.101	0.140	0.145	0.137	0.031	0.113	0.127	0.135	0.111 (-2.6%)
	Dep (CG)	0.081	0.049	0.117	0.147	0.099	0.140	0.141	0.120	0.032	0.102	0.109	0.120	0.105 (-8.4%)
	Dep (CG) ^{UW}	0.095	0.050	0.131	0.161	0.091	0.032	0.125	0.147	0.021	0.089	0.098	0.106	0.096 (-16.4%)

Algorithm 1 Copula estimation: PyTorch-like Pseudocode

```

# model_e, model_c: Weibull models for E/C
# copula: Copula model with parameter theta
def find_copula(model_e, model_c, copula,
               d_train, d_val):

    opt = Adam(params=[model_e, model_c,
                       copula])

    for epoch in range(n_epochs):
        for batch in DataLoader(d_train):
            train_nll = loss(model_e, model_c,
                             batch, copula)
            opt.zero_grad()
            train_nll.backward()
            opt.step()

        # validate and track best loss
        val_losses = []
        for batch in DataLoader(d_val):
            val_losses.append(loss(model_e,
                                   model_c, batch, copula))
        val_nll = mean(val_losses)

        # check early stopping
        if theta_stabilized(copula) and
           no_val_improve(val_nll): break

    return copula

# Computes dependent NLL given a copula
def loss(model_e, model_c, batch, copula):
    T, E, X = batch.T, batch.E, batch.X
    s1 = model_e.survival(T, X) # S_E(T|X)
    s2 = model_c.survival(T, X) # S_C(T|X)
    f1 = model_e.pdf(T, X) # f_E(T|X)
    f2 = model_c.pdf(T, X) # f_C(T|X)

    # dependent likelihood via copula terms
    S = cat([s1, s2])
    p1 = log(f1) + log(copula.ccdf("u", S))
    p2 = log(f2) + log(copula.ccdf("v", S))

    reg = 0.01 * sum(copula.theta ** 2)
    ll = p1 * E + (1 - E) * p2
    return -mean(ll) + reg

```

Notes: cat concatenates matrices. log calculates the natural logarithm. ccdf computes a conditional CDF via differentiation.

the KM estimate of the censoring distribution becomes more biased; however, the resulting Brier-score bias can still depend on the learner’s prediction errors. When informative features are removed, prediction errors increase; under dependent censoring, IPCW applies misspecified inverse weights to these larger squared errors, amplifying the resulting bias. In addition, with weaker (random) covariates, the marginal model explains less variation in E , leaving

Table 3: Ranking performance of the independent and proposed dependent metrics across 12 datasets and 4 feature selection strategies. The results show how often each metric recovered at least two of the top three survival learners according to the oracle IBS, across 10 experiments per feature selection strategy. Each entry is therefore reported out of 40. Higher is better.

Dataset	IPCW (KM)	IPCW (CoxPH)	Dep (KM)	Dep (CG)	Dep (CG) ^{UW}
WHAS	38/40	39/40	40/40	40/40	39/40
METABRIC	40/40	40/40	39/40	39/40	40/40
Churn	37/40	37/40	34/40	29/40	32/40
GBSG	40/40	40/40	38/40	37/40	35/40
NACD	39/40	40/40	38/40	36/40	38/40
FLCHAIN	39/40	39/40	18/40	18/40	35/40
SUPPORT	37/40	36/40	35/40	34/40	31/40
Employee	33/40	35/40	12/40	11/40	13/40
MIMIC (IV)	33/40	29/40	1/40	1/40	13/40
SEER (brain)	37/40	38/40	38/40	36/40	36/40
SEER (liver)	39/40	39/40	38/40	30/40	35/40
SEER (stom.)	36/40	38/40	40/40	35/40	35/40

more unexplained (residual) dependence between E and C . The CoxPH-based IPCW estimator also depends strongly on the selected feature set: it *reduces* error by up to 8% for the Original, Top-5, and Top-10 strategies, but *increases* error by 2% under the Random 25% strategy. This highlights a key limitation of conditional IPCW, since an uninformative feature set can yield a misspecified censoring model and invalidate conditional independent censoring. The ablation study highlights the importance of the CG estimator. Replacing censored event times with margin times derived from KM yields only modest gains of 1-3% over IBS-IPCW. In contrast, the CG estimator estimates the marginal event-time distribution while accounting for dependence between event and censoring times through the specified copula model, and uses this distribution to compute the conditional margin time for censored instances. This increases the reduction to 4-8% across feature selection strategies, indicating that accounting for dependent censoring when estimating the event-time margin provides substantial improvements over KM-based margin imputation.

Ranking. Table 3 reports the ranking experiment, which evaluates whether a censored metric identifies the same top-performing models as the oracle IBS. We first compute the oracle IBS for each model (*i.e.*, without censoring) and identify the top-3 models, then assess whether each censored metric recovers at least two of them. IBS-IPCW preserves the oracle top-3 most consistently overall, while the dependent metrics perform comparably on several datasets but worse on FLCHAIN, Employee, and MIMIC-IV. Thus, although the copula-based IBS substantially reduces absolute estimation error under dependent censoring, this does not necessarily translate into better separation between closely

performing models. IBS-IPCW may therefore remain reasonable for within-dataset ranking when its bias affects competing models similarly. However, it does not estimate the true IBS under dependent censoring and may give unfair rankings when models differ in how they handle censoring.

Computational analysis. The primary computational cost lies in copula fitting, which jointly optimizes the marginal survival models and copula parameter with model selection over candidate families (Clayton and Frank). Empirically, this required 5-75 minutes, depending on dataset size (Appendix F.1). Once estimated, IBS-Dep adds negligible cost relative to IPCW. The incremental peak PyTorch GPU memory allocation is modest (≤ 11 MiB), indicating that the method requires little additional GPU memory.

6 DISCUSSION

6.1 MAIN FINDINGS AND IMPLICATIONS

Our results show that accounting for marginal dependent censoring at evaluation time can meaningfully change how we judge model performance. Across synthetic and semi-synthetic experiments, IBS-Dep consistently reduced bias relative to IBS-IPCW when event and censoring times were dependent. The copula-based estimator achieved average reductions of 4-8%, while the uncertainty-weighted variant yielded improvements of 12-16%. These effects were observed across 12 generated survival datasets with sample sizes from 500 to 10,000, feature dimensions from 14 to 91, and censoring rates between 32% and 76%. Overall, this suggests that dependent censoring is not a rare edge case, but a relevant trait of many real-world datasets. Although IBS-Dep reduced estimation error, it did not consistently improve the ranking of models, because bias reduction does not necessarily improve separation between models.

Most prior work concerning dependent censoring has concentrated on *model estimation*. Copula-based modeling approaches [Foomani et al., 2023, Zhang et al., 2024, Liu et al., 2025] explicitly model the joint distribution of event and censoring times to capture their dependence. However, these models are typically evaluated using metrics such as Harrell’s C-index or Brier score with IPCW, both of which assume independent censoring. This creates a mismatch: models are designed for dependence, but evaluation still assumes independence. As a result, reported performance in dependent censoring settings may itself be biased. Our work shifts the focus to *model evaluation*, introducing a flexible copula-based extension of the Brier score.

6.2 LIMITATIONS AND FUTURE WORK

Our work also points to several directions for future research. More broadly, the framework introduced here opens the door

to dependence-aware versions of other common evaluation metrics. Widely used measures – such as Harrell’s C-index, Uno’s C-index [Uno et al., 2011], and D-calibration [Haider et al., 2020] – rely on independence assumptions. Extending copula-based corrections to concordance and calibration measures is a natural next step toward having metrics that all support dependent censoring.

At the same time, several limitations remain. First, we assume a stable dependence structure; under temporal distribution shift, the copula parameter must be re-estimated, suggesting the need for adaptive or time-varying dependence models. Second, the true copula is not identifiable in real-world survival datasets because for each individual only one of the competing times is observed [Tsiatis, 1975]. Some misspecification is therefore unavoidable. Third, likelihood-based estimation can be unstable in small samples, particularly under dependent censoring, because weak identification of the copula parameter can make its estimate highly sensitive to sampling variation. These limitations motivate reporting sensitivity analyses across plausible copula families and dependence strengths. Agreement across specifications would strengthen confidence in the evaluation, whereas disagreement would reveal that model comparisons are sensitive to assumptions about the censoring mechanism.

6.3 USER GUIDE

We conclude with a brief guide on how to use IBS-Dep. The central choice is the dependence model used to approximate the relationship between event and censoring times. In standard right-censored data, we observe $t_i = \min(e_i, c_i)$ and $\delta_i = \mathbb{1}(e_i \leq c_i)$, but only the smaller of e_i and c_i . Thus, the joint distribution of the population-level variables E and C , and hence their true dependence structure, is generally not identifiable without additional assumptions. We therefore propose not attempting to recover the true latent copula, but instead using a quantitative model comparison to determine whether an Archimedean copula fits the data better than independence.

In practice, one specifies candidate Archimedean copula families and parametric marginal models for E and C , whose parameters are estimated by minimizing the dependent NLL in Equation 5.1. As a baseline, one should fit the independence copula using the same marginal specification. The model with the best validation likelihood is selected; if none of the candidate dependent copulas improves upon the independence copula, one should default to independence. Algorithm 1 summarizes the copula estimation procedure. The selected copula is treated as a working model for the censoring mechanism, and the CG estimate of the event-time distribution is used to impute margin times for censored instances. The selected family is not claimed to be the true latent copula, but a likelihood-based approximation of the dependence structure between event and censoring times.

Acknowledgements

This research received support from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Institute for Advanced Research (CIFAR), and the Alberta Machine Intelligence Institute (Amii).

References

- Leo Breiman. Random Forests. *Machine Learning*, 45(1): 5–32, 2001.
- Norman E. Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review*, pages 45–57, 1975.
- Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.
- Federico Campigotto and Edie Weller. Impact of informative censoring on the Kaplan-Meier estimate of progression-free survival in phase II clinical trials. *Journal of Clinical Oncology*, 32(27):3068–3074, 2014.
- David G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978.
- David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.
- Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, Carlos Caldas, Samuel Aparicio, and METABRIC Group. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403): 346–352, 2012.
- Angela Dispenzieri, Jerry A. Katzmann, Robert A. Kyle, Dirk R. Larson, Terry M. Therneau, Colin L. Colby, Raynell J. Clark, Graham P. Mead, Shaji Kumar, L. Joseph Melton, and S. Vincent Rajkumar. Use of Nonclonal Serum Immunoglobulin Free Light Chains to Predict Overall Survival in the General Population. *Mayo Clinic Proceedings*, 87(6):517–523, 2012.
- Takeshi Emura and Yi-Hau Chen. *Analysis of Survival Data with Dependent Censoring: Copula-Based Approaches*. Springer, 2018.
- Ali H. G. Foomani, Michael Cooper, Russell Greiner, and Rahul G. Krishnan. Copula-based deep survival models for dependent censoring. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, volume 216, pages 669–680, 2023.
- Stephane Fotso et al. *PySurvival: Open Source Package for Survival Analysis Modeling*, 2019. Available from <https://www.pysurvival.io/>.
- Maurice J. Frank. On the simultaneous associativity of (x, y) and $x + y - f(x, y)$. *Aequationes Mathematicae*, 19(1):194–226, 1979.
- Thomas A. Gerds and Martin Schumacher. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 2006.
- Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.
- Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective Ways to Build and Evaluate Individual Survival Distributions. *Journal of Machine Learning Research*, 21(1):1–63, 2020.
- Frank E. Harrell Jr., Kerry L. Lee, and Daniel B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.
- David W. Hosmer, Stanley Lemeshow, and Susanne May. *Applied Survival Analysis*. John Wiley & Sons, Inc., 2008.
- Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random Survival Forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.
- Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1–9, 2023.
- Edwin L Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.
- Jared Katzman, Uri Shaham, Jonathan Bates, Alexander Cloninger, Tingting Jiang, et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):1–12, 2018.
- Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2017.

- David G. Kleinbaum and Mitchel Klein. *Survival Analysis: A Self-Learning Text*. Statistics for Biology and Health. Springer New York, 1 edition, 2012.
- William A. Knaus. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine*, 122(3):191–203, 1995.
- Christian Marius Lillielund, Ali Hossein Gharari Foomani, Weijie Sun, Shi-ang Qi, Russell Greiner, and The Pooled Resource Open-Access ALS Clinical Trials Consortium (PRO-ACT). MENSA: A Multi-Event Network for Survival Analysis with Trajectory-Based Likelihood Estimation. In *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297, pages 543–571, 2026a.
- Christian Marius Lillielund, Shi-ang Qi, Russell Greiner, and Christian Fischer Pedersen. Position: Stop Chasing the C-index when Evaluating Survival Analysis Models. *arXiv preprint arXiv:2506.02075*, 2026b.
- Xin Liu, Weijia Zhang, and Min-Ling Zhang. HACSURV: A Hierarchical Copula-Based Approach for Survival Analysis with Dependent Competing Risks. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pages 3079–3087, 2025.
- Roger B. Nelsen. *An Introduction to Copulas*. Springer Science & Business Media, 2 edition, 2006.
- Sebastian Pölsterl. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020.
- Shi-ang Qi, Neeraj Kumar, Mahtab Farrokh, Weijie Sun, Li Hao Kuan, et al. An Effective Meaningful Way to Evaluate Survival Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 28244–28276, 2023a.
- Shi-ang Qi, Weijie Sun, and Russell Greiner. SurvivalEVAL: A Comprehensive Open-Source Python Package for Evaluating Individual Survival Distributions. In *Proceedings of the 2023 AAAI Fall Symposia*, volume 2, pages 453–457, 2023b.
- Shi-ang Qi, Yakun Yu, and Russell Greiner. Conformalized Survival Distributions: A Generic Post-Process to Increase Calibration. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 41303–41339, 2024.
- Greg Ridgeway. The State of Boosting. *Computing Science and Statistics*, 31:172–181, 1999.
- Lynn A. Gloeckler Ries, Marsha E. Reichman, Denise Riedel Lewis, Benjamin F. Hankey, and Brenda K. Edwards. Cancer survival and incidence from the Surveillance, Epidemiology, and End Results (SEER) program. *Oncologist*, 8(6):541–552, 2003.
- Louis-Paul Rivest and Martin T. Wells. A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *Journal of Multivariate Analysis*, 79:138–155, 2001.
- Patrick Royston and Douglas G. Altman. External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology*, 13:1–15, 2013.
- Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris*, 8:229–231, 1959.
- Terry M. Therneau. *A Package for Survival Analysis in R*, 2024. URL <https://CRAN.R-project.org/package=survival>. R package version 3.6-4.
- Anastasios Tsiatis. A Nonidentifiability Aspect of the Problem of Competing Risks. *Proceedings of the National Academy of Sciences of the United States of America*, 72(1):20–22, 1975.
- Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D’Agostino, and L. J. Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117, 2011.
- Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors. In *Advances in Neural Information Processing Systems*, volume 24, pages 1845–1853, 2011.
- Weijia Zhang, Chun Kai Ling, and Xuanhui Zhang. Deep Copula-Based Survival Analysis for Dependent Censoring with Identifiability Guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20613–20621, 2024.
- Ming Zheng and John P. Klein. Estimates of Marginal Survival for Dependent Competing Risks Based on an Assumed Copula. *Biometrika*, 82(1):127–138, 1995.

Overcoming Dependent Censoring in the Evaluation of Survival Models (Supplementary Material)

Christian Marius Lillelund^{1,2}

Shi-ang Qi^{1,3,4}

Russell Greiner^{1,3}

¹Department of Computing Science, University of Alberta, Edmonton, Canada

²Department of Electrical and Computer Engineering, Aarhus University, Aarhus, Denmark

³Alberta Machine Intelligence Institute, Edmonton, Canada

⁴Vector Institute, Toronto, Canada

A CENSORING ASSUMPTIONS

Random censoring. Instances censored at time t are representative of all instances still at risk at t , in terms of their survival experience [Kleinbaum and Klein, 2012, Ch. 1]. In a medical study, we would assume that the death rate of censored individuals is the same as that of uncensored individuals who remain in the risk set. Formally, $E \perp\!\!\!\perp C$, where E is the event time and C is the censoring time, meaning $\Pr(E > t \mid C = t) = \Pr(E > t)$ for all t .

Independent Censoring. In the absence of competing risks (*i.e.*, when no alternative event precludes the main event), within any subgroup of interest (*e.g.*, a smoker and a non-smoker group), instances censored at time t are representative of those who remain at risk at time t with respect to their survival experience [Kleinbaum and Klein, 2012]. This is also known as conditional independent censoring, since we assume that the observed covariates account for the relationship between event and censoring. Thus, once we condition on the covariates, the censoring time is assumed to be independent of the event time. Formally, $E \perp\!\!\!\perp C \mid X$, where X are the covariates, meaning $\Pr(E > t \mid C = t, X) = \Pr(E > t \mid X)$ for all t . Independent censoring is a less restrictive assumption than random censoring on the data-generating process.

Dependent Censoring. In the absence of competing risks, within any subgroup of interest, instances censored at time t are *not* representative of those who remain at risk at time t with respect to their survival experience [Foomani et al., 2023, Lillelund et al., 2026b]. Equivalently, censoring is said to be dependent when, even after conditioning on the observed covariates, the censoring mechanism remains related to the event process. Thus, the observed covariates do not fully account for the relationship between E and C , and the censoring time is not independent of the event time given the covariates.

B COPULAS FOR SURVIVAL ANALYSIS

Bivariate survival analysis can be viewed as a *competing risks* problem, where the occurrence of an event (*e.g.*, death) and censoring (*e.g.*, dropout) are mutually exclusive. Tsiatis [1975] referred to this as the *identifiability issue*; competing risks data do not allow one to observe two event times simultaneously (one event censors the other), and hence, the data may not identify the dependence structure between event times. Zheng and Klein [1995] gave a partial solution to the identifiability problem of dependent censoring by an assumed *copula* between two event times. A copula links two or more random variables by specifying their dependence structure [Emura and Chen, 2018]. A mathematician, Abe Sklar, first used the word “copula” in his study of probabilistic metric spaces [Sklar, 1959]. In his work, he gave a mathematical definition of copulas and established the most fundamental theorem about copulas, known as Sklar’s theorem [Sklar, 1959]. It demonstrated that any joint cumulative distribution function can be written in terms of a copula over the quantiles of its marginal cumulative distribution functions. Among the various types of copulas, the Archimedean copula is the most widely used, and enables the modeling of dependencies in arbitrarily high dimensions using a single parameter. This family of copulas is defined by a convex generator function, $\varphi_\theta : [0, 1] \rightarrow [0, \infty]$, which is continuous, strictly decreasing, and satisfies the boundary conditions $\varphi_\theta(0) = \infty$ and $\varphi_\theta(1) = 0$ [Emura and Chen, 2018]. Given event time E and censoring time C , their joint survival function under an Archimedean copula can be written as:

$$\Pr(E > t, C > c) = \varphi_\theta^{-1} [\varphi_\theta(S_E(t)) + \varphi_\theta(S_C(c))], \quad (\text{B.1})$$

where φ_θ is the generator function, and $S_E(t) = \Pr(E > t)$ and $S_C(c) = \Pr(C > c)$ are the marginal survival functions. In the bivariate case, φ_θ describes the dependence structure (*i.e.*, correlation) between two random variables, *e.g.*, E and C , and the (rank) correlation can be described with Kendall's tau (τ):

$$\tau = \Pr(t_2 - t_1)(c_2 - c_1) \geq 0 - \Pr(t_2 - t_1)(c_2 - c_1) < 0, \quad (\text{B.2})$$

where (t_1, c_1) and (t_2, c_2) are sampled from a (survival) copula. The Clayton [Clayton, 1978] and Frank [Frank, 1979] copulas are two common Archimedean copulas in survival literature. Within these families, the copula C_θ is parameterized by θ , interpreted as the degree of dependence between E and C in the bivariate case. When $\theta \approx 0$, it yields the independence copula for Clayton and Frank copulas; thus, a larger value of θ implies greater dependence between the marginal distributions. Formally, we write $C(u_1, \dots, u_d) : [0, 1]^d \rightarrow [0, 1]$, which is a d -dimensional copula (*i.e.*, where d is the number of variables) with $u_1, \dots, u_d$ uniform marginal probabilities and the following properties [Nelsen, 2006]:

1. **Groundedness:** If there exists an $i \in \{1, \dots, d\}$ such that $u_i = 0$, then

$$C(u_1, \dots, u_d) = 0.$$

2. **Uniform Margins:** For all $i \in \{1, \dots, d\}$, if $\forall j \neq i \Rightarrow u_j = 1$, then

$$C(u_1, \dots, u_d) = u_i.$$

3. **d -Increasingness:** For all $u = (u_1, \dots, u_d)$ and $v = (v_1, \dots, v_d)$ where $u_i < v_i$ for all $i = 1, \dots, d$, the following holds:

$$\sum_{\ell \in \{0,1\}^d} (-1)^{\ell_1 + \dots + \ell_d} C(u_1^{\ell_1} v_1^{1-\ell_1}, \dots, u_d^{\ell_d} v_d^{1-\ell_d}) \geq 0.$$

Let E be the event distribution, C the censoring distribution, $\mathbf{x}$ a vector of covariates, and $S_E(t | \mathbf{x}) = \Pr(E > t | \mathbf{x})$ and $S_C(c | \mathbf{x}) = \Pr(C > c | \mathbf{x})$ the marginal survival functions given $\mathbf{x}$. We can then define a bivariate (survival) copula C_θ [Nelsen, 2006] that describes the degree of dependence between E and C as follows:

$$\Pr(E > t, C > c | \mathbf{x}) = C_\theta(S_E(t | \mathbf{x}), S_C(c | \mathbf{x})). \quad (\text{B.3})$$

Kendall's tau (τ) can be defined as:

$$\tau = \Pr(t_2 - t_1)(c_2 - c_1) \geq 0 | \mathbf{x} - \Pr(t_2 - t_1)(c_2 - c_1) < 0 | \mathbf{x}, \quad (\text{B.4})$$

where (t_1, c_1) and (t_2, c_2) are sampled from the model (B.3). Kendall's τ can be solely expressed as a function of the copula C_θ by [Emura and Chen, 2018]:

$$\tau_\theta = 4 \int_0^1 \int_0^1 C_\theta(u, v), C_\theta(du, dv), \quad (\text{B.5})$$

which implies that Kendall's τ does not depend on the marginal survival functions. An Archimedean copula is a type of copula that is defined using a generator function, which allows for a flexible way to model dependence structures between random variables, *e.g.*, E and C . Such a copula is defined as [Emura and Chen, 2018]:

$$C_\theta(u, v) = \varphi_\theta^{-1}(\varphi_\theta(u) + \varphi_\theta(v)). \quad (\text{B.6})$$

where the function $\varphi_\theta : [0, 1] \rightarrow [0, \infty]$ is called a generator of the copula, which is continuous and strictly decreasing from $\varphi_\theta(0) = \infty$ to $\varphi_\theta(1) = 0$. The following Archimedean copulas are considered in this work:

• **The independence copula:**

$$C(u, v) = uv. \quad (\text{B.7})$$

• **The Clayton copula [Clayton, 1978]:**

$$C_\theta(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}, \quad \theta \geq 0. \quad (\text{B.8})$$

• **The Frank copula [Frank, 1979]:**

$$C_\theta(u, v) = -\frac{1}{\theta} \log \left[1 + \frac{(\exp(-\theta u) - 1)(\exp(-\theta v) - 1)}{e^{-\theta} - 1} \right], \quad \theta \neq 0. \quad (\text{B.9})$$

C SURVIVAL LIKELIHOOD

In the standard survival analysis framework, we define the survival time E and censoring time C such that:

$$\begin{aligned} E &= t_i \text{ and } C \geq t_i & \text{if } \delta_i = 1, \\ E &> t_i \text{ and } C = t_i & \text{if } \delta_i = 0. \end{aligned}$$

Combining these cases, the likelihood for the i -th observation is:

$$\mathcal{L}_i = \Pr(E = t_i, C \geq t_i | x_i)^{\delta_i} \Pr(E \geq t_i, C = t_i | x_i)^{1-\delta_i} \quad (\text{C.1})$$

Under the assumption of independent censoring [Emura and Chen, 2018], we have:

$$\begin{aligned} \mathcal{L}_i &= [\Pr(E = t_i | x_i) \Pr(C \geq t_i | x_i)]^{\delta_i} [\Pr(E \geq t_i | x_i) \Pr(C = t_i | x_i)]^{1-\delta_i} \\ &= [f_E(t_i | x_i) S_C(t_i | x_i)]^{\delta_i} [S_E(t_i | x_i) f_C(t_i | x_i)]^{1-\delta_i} \\ &= [f_E(t_i | x_i)^{\delta_i} S_E(t_i | x_i)^{1-\delta_i}] [f_C(t_i | x_i)^{1-\delta_i} S_C(t_i | x_i)^{\delta_i}] \end{aligned} \quad (\text{C.2})$$

where $S_E(t | x_i) = \Pr(E \geq t | x_i)$, $f_E(t | x_i) = -dS_E(t | x_i)/dt$, $S_C(t | x_i) = \Pr(C \geq t | x_i)$, and $f_C(t | x_i) = -dS_C(t | x_i)/dt$. Under the non-informative censoring assumption, the term $f_C(t_i | x_i)^{1-\delta_i} S_C(t_i | x_i)^{\delta_i}$ is unrelated to the likelihood for the event times and can simply be ignored [Emura and Chen, 2018]. Therefore, the likelihood function can be written as:

$$\mathcal{L} = \prod_{i=1}^N f_E(t_i | \mathbf{x}_i)^{\delta_i} S_E(t_i | \mathbf{x}_i)^{1-\delta_i}, \quad (\text{C.3})$$

where $f_E(t_i | \mathbf{x}_i)$ is the density function of the event time and $S_E(t_i | \mathbf{x}_i)$ is the survival function of the event time. Note how the censoring terms $f_C(t_i)$ and $S_C(t_i)$ are absent in (C.3). If the event is observed, the likelihood contribution is the probability density of the event occurring at that time. If the observation is censored, the likelihood contribution is the probability of surviving beyond that time. The overall likelihood is the product of these contributions across all instances.

However, if censoring is dependent, we can no longer rely on this clean decomposition of the likelihood as in Equation C.2. Instead, given some Archimedean copula (e.g., Clayton or Frank), we can rewrite Equation C.2 as the log-likelihood under dependent censoring [Foomani et al., 2023]:

$$\begin{aligned} \ell(\theta; \mathcal{D}) &= \sum_{i=1}^N \delta_i \log[f_E(t_i | \mathbf{x}_i)] + \delta_i \log \left[\frac{\partial}{\partial u_1} C(u_1, u_2) \right]_{\substack{u_1 = S_E(t_i | \mathbf{x}_i) \\ u_2 = S_C(t_i | \mathbf{x}_i)}} \\ &+ (1 - \delta_i) \log[f_C(t_i | \mathbf{x}_i)] + (1 - \delta_i) \log \left[\frac{\partial}{\partial u_2} C(u_1, u_2) \right]_{\substack{u_1 = S_E(t_i | \mathbf{x}_i) \\ u_2 = S_C(t_i | \mathbf{x}_i)}} \end{aligned} \quad (\text{C.4})$$

In this expression, the first term represents the log-likelihood of observing the event at time t_i . The second term captures the conditional probability of the censoring time occurring after the event time, given that the event occurred at t_i . By symmetry, the third and fourth terms mirror these quantities but with respect to the censoring time. Although Equation C.4 may appear complex, the partial derivatives of any Archimedean copula have closed-form expressions, which allows us to maximize the log-likelihood using gradient descent.

D THEORETICAL PROPERTIES AND DERIVATIONS

This appendix develops the theoretical foundations of the proposed CG margin-time estimator under dependent censoring. We first derive the KM and CG estimators to establish the survival estimation framework under independent and dependent censoring, respectively, which forms the basis for the subsequent margin-time analysis. We then recall the asymptotic target of the CG survival estimator under possible copula misspecification and introduce the margin-time functional and its plug-in CG estimator.

D.1 THE KAPLAN-MEIER ESTIMATOR

The following derivation follows the presentation of Emura and Chen [2018], generalized to accommodate tied event times.

Recall that (t_i, δ_i) , $i = 1, \dots, N$, denotes survival data without covariates, where $t_i = \min(e_i, c_i)$ and $\delta_i = \mathbb{1}[e_i \leq c_i]$. Let $u_1 < \dots < u_J$ denote the distinct observed event times. For each u_j , define

$$d_j = \sum_{i=1}^N \mathbb{1}[t_i = u_j, \delta_i = 1]$$

as the number of events at u_j , and

$$n_j = \sum_{i=1}^N \mathbb{1}[t_i \geq u_j]$$

as the number of instances at risk immediately before u_j .

We represent $S_E(t)$ as a decreasing step function whose jumps occur at the observed event times. Its value at time t can therefore be written as

$$S_E(t) = \prod_{u_j \leq t} \frac{S_E(u_j^-)}{S_E(u_j^-)} \quad (\text{D.1})$$

$$= \prod_{u_j \leq t} \left(1 - \frac{\Pr(E = u_j)}{\Pr(E \geq u_j)} \right), \quad (\text{D.2})$$

where $S_E(u_j^-)$ denotes the survival probability immediately before the jump at u_j .

Suppose that E and C are independent. Then,

$$S_E(t) = \prod_{u_j \leq t} \left(1 - \frac{\Pr(E = u_j, C \geq u_j)}{\Pr(E \geq u_j, C \geq u_j)} \right) \quad (\text{D.3})$$

$$= \prod_{u_j \leq t} \left(1 - \frac{\Pr(\min(E, C) = u_j, E \leq C)}{\Pr(\min(E, C) \geq u_j)} \right). \quad (\text{D.4})$$

Replacing each probability ratio by its empirical estimate gives

$$\hat{S}_E(t) = \prod_{u_j \leq t} \left(1 - \frac{\sum_{i=1}^N \mathbb{1}[t_i = u_j, \delta_i = 1]}{\sum_{i=1}^N \mathbb{1}[t_i \geq u_j]} \right) \quad (\text{D.5})$$

$$= \prod_{u_j \leq t} \left(1 - \frac{d_j}{n_j} \right), \quad 0 \leq t \leq t_{\max}. \quad (\text{D.6})$$

Thus, all events tied at u_j are incorporated through the common factor $1 - d_j/n_j$. The empty product is defined as one, and the estimator is undefined for $t > t_{\max}$.

D.2 THE COPULA-GRAPHIC ESTIMATOR

The following derivation is based on the Copula-Graphic estimator introduced by Zheng and Klein [1995] and its explicit Archimedean representation derived by Rivest and Wells [2001], generalized here to accommodate tied event times.

Consider the Archimedean copula model

$$\Pr(E > t, C > c) = \varphi_\theta^{-1} [\varphi_\theta(S_E(t)) + \varphi_\theta(S_C(c))], \quad (\text{D.7})$$

where $\varphi_\theta : [0, 1] \rightarrow [0, \infty]$ is a continuous, strictly decreasing generator satisfying $\varphi_\theta(0) = \infty$ and $\varphi_\theta(1) = 0$ [Emura and Chen, 2018]. The marginal survival functions are $S_E(t) = \Pr(E > t)$ and $S_C(t) = \Pr(C > t)$.

Let $u_1 < \dots < u_J$ denote the distinct observed event times, and let d_j and n_j be defined as in Section 2. We treat all events tied at u_j as a single jump of the event survival function. When events and censorings occur at the same observed time, we use the event-before-censoring convention: the event jump is applied before any censoring removals at that time.

Immediately before the event jump at u_j , the empirical joint survival probability is

$$\widehat{\Pr}(E \geq u_j, C \geq u_j) = \frac{n_j}{N}.$$

Immediately after removing the d_j events at u_j , but before removing any censorings tied at that time, it is

$$\frac{n_j - d_j}{N}.$$

Because only the event survival function changes across this event jump, the copula representation gives

$$\varphi_\theta\left(\frac{n_j - d_j}{N}\right) = \varphi_\theta\left(\widehat{S}_E(u_j)\right) + \varphi_\theta\left(\widehat{S}_C(u_j^-)\right), \quad (\text{D.8})$$

whereas immediately before the jump,

$$\varphi_\theta\left(\frac{n_j}{N}\right) = \varphi_\theta\left(\widehat{S}_E(u_j^-)\right) + \varphi_\theta\left(\widehat{S}_C(u_j^-)\right). \quad (\text{D.9})$$

Subtracting Equation D.9 from Equation D.8 yields the jump recursion

$$\begin{aligned} \varphi_\theta\left(\widehat{S}_E(u_j)\right) - \varphi_\theta\left(\widehat{S}_E(u_j^-)\right) \\ = \varphi_\theta\left(\frac{n_j - d_j}{N}\right) - \varphi_\theta\left(\frac{n_j}{N}\right). \end{aligned} \quad (\text{D.10})$$

Since $\widehat{S}_E(t) = 1$ before the first observed event and $\varphi_\theta(1) = 0$, summing the jumps up to time t gives

$$\varphi_\theta\left(\widehat{S}_E(t)\right) = \sum_{u_j \leq t} \left[\varphi_\theta\left(\widehat{S}_E(u_j)\right) - \varphi_\theta\left(\widehat{S}_E(u_j^-)\right) \right] \quad (\text{D.11})$$

$$= \sum_{u_j \leq t} \left[\varphi_\theta\left(\frac{n_j - d_j}{N}\right) - \varphi_\theta\left(\frac{n_j}{N}\right) \right]. \quad (\text{D.12})$$

Applying φ_θ^{-1} therefore gives the CG estimator

$$\widehat{S}_E(t) = \varphi_\theta^{-1} \left(\sum_{u_j \leq t} \left[\varphi_\theta\left(\frac{n_j - d_j}{N}\right) - \varphi_\theta\left(\frac{n_j}{N}\right) \right] \right), \quad 0 \leq t \leq t_{\max}. \quad (\text{D.13})$$

The empty sum is zero, and the estimator is undefined for $t > t_{\max}$.

For the independence generator $\varphi_{\text{ind}}(u) = -\log u$, Equation D.13 becomes

$$\widehat{S}_E(t) = \exp \left(- \sum_{u_j \leq t} \left[-\log\left(\frac{n_j - d_j}{N}\right) + \log\left(\frac{n_j}{N}\right) \right] \right) \quad (\text{D.14})$$

$$= \prod_{u_j \leq t} \frac{n_j - d_j}{n_j} = \prod_{u_j \leq t} \left(1 - \frac{d_j}{n_j} \right), \quad (\text{D.15})$$

which is the Kaplan-Meier estimator.

For the Clayton generator $\varphi_\theta(u) = u^{-\theta} - 1$ [Clayton, 1978], the CG estimator becomes

$$\widehat{S}_E(t) = \left[1 + \sum_{u_j \leq t} \left\{ \left(\frac{n_j - d_j}{N} \right)^{-\theta} - \left(\frac{n_j}{N} \right)^{-\theta} \right\} \right]^{-1/\theta}. \quad (\text{D.16})$$

D.3 THEORETICAL ANALYSIS

Rivest and Wells [2001] study the CG estimator for $\hat{S}_E(t)$ under dependent censoring. Without assuming the fitted Archimedean copula is the true copula, their estimator targets a marginal survival function $S_E^*(t)$ defined via their equation (8) (at page 6). Moreover, when the dependence is truly Archimedean with the same generator used to construct the estimator, $\hat{S}_E(t)$ is uniformly consistent for $S_E(t)$ [Rivest and Wells, 2001, Theorem 1].

In this section, we provide theoretical properties of this margin time. We start the analysis by formally (re-)stating the calculation.

For any $c \geq 0$ with $S_E(c) > 0$ and $\int_c^\infty S_E(u) du < \infty$, define

$$e^{\text{margin}}(c; S_E) := c + \frac{\int_c^\infty S_E(u) du}{S_E(c)}. \quad (\text{D.17})$$

We estimate it by the plug-in CG estimator

$$\hat{e}^{\text{margin}}(c) := e^{\text{margin}}(c; \hat{S}_E) = c + \frac{\int_c^\infty \hat{S}_E(u) du}{\hat{S}_E(c)}. \quad (\text{D.18})$$

Similarly, define the large-sample target

$$e^{*,\text{margin}}(c) := e^{\text{margin}}(c; S_E^*) = c + \frac{\int_c^\infty S_E^*(u) du}{S_E^*(c)}. \quad (\text{D.19})$$

When the copula model is correctly specified, $S_E^* = S_E$ and hence $e^{*,\text{margin}}(c) = e^{\text{margin}}(c; S_E)$.

To connect (D.17)–(D.18) to Rivest and Wells [2001] results (proved on $[0, t_0)$), without loss of generality, we assume the right tail beyond some finite horizon (*i.e.*, t_0) does not contribute. Under such a setting, we can rewrite the integrals in (D.18)–(D.19) using

$$\begin{aligned} \int_c^\infty \hat{S}_E(u) du &= \int_c^{t_0} \hat{S}_E(u) du \\ \int_c^\infty S_E^*(u) du &= \int_c^{t_0} S_E^*(u) du, \end{aligned}$$

so the theory reduces to controlling $\hat{S}_E$ on the finite interval $[0, t_0)$ without any tail extrapolation.

D.3.1 Margin time as conditional expectation

Now we prove our method in (D.17) corresponds to the conditional expectation. This following theorem and proof are adapted from Haider et al. [2020, Theorem B.1]. Here, we rephrase the proof for the completeness.

Lemma D.1. *Let $E \geq 0$ have survival function $S_E(t) = \Pr(E > t)$. Fix $c \geq 0$ with $S_E(c) > 0$ and assume $\int_c^\infty S_E(u) du < \infty$. Then*

$$\mathbb{E}[E \mid E > c] = c + \frac{\int_c^\infty S_E(u) du}{S_E(c)}. \quad (\text{D.20})$$

Proof. Define $(E - c)_+ := \max\{E - c, 0\}$. For any nonnegative Y , $\mathbb{E}[Y] = \int_0^\infty \Pr(Y > v) dv$. Apply this conditionally

on $\{E > c\}$:

$$\begin{aligned}
\mathbb{E}[(E - c)_+ \mid E > c] &= \int_0^\infty \Pr((E - c)_+ > v \mid E > c) dv \\
&= \int_0^\infty \Pr(E > c + v \mid E > c) dv \\
&= \int_0^\infty \frac{\Pr(E > c + v)}{\Pr(E > c)} dv \\
&= \frac{1}{S_E(c)} \int_0^\infty S_E(c + v) dv \\
&= \frac{1}{S_E(c)} \int_c^\infty S_E(u) du,
\end{aligned}$$

where we substituted $u = c + v$. Finally, on $\{E > c\}$, $E = c + (E - c)$, hence

$$\begin{aligned}
\mathbb{E}[E \mid E > c] &= c + \mathbb{E}[E - c \mid E > c] \\
&= c + \mathbb{E}[(E - c)_+ \mid E > c] \\
&= c + \frac{\int_c^\infty S_E(u) du}{S_E(c)}.
\end{aligned}$$

□

D.3.2 Consistency of the margin time

To simplify the notation, we denote the numerator term in (D.17) as

$$A(c; S) := \int_c^\infty S(u) du.$$

Lemma D.2. *Let $s_{\min}$ be a small positive constant that satisfies $S(c) \geq s_{\min} > 0$, $A(c; S) < \infty$, and $A(c; \hat{S}) < \infty$. Let*

$$\Delta A := A(c; \hat{S}) - A(c; S) = \int_c^\infty (\hat{S}(u) - S(u)) du, \quad \Delta S(c) := \hat{S}(c) - S(c).$$

Assume $|\Delta S(c)| \leq s_{\min}/2$ (so that $\hat{S}(c) \geq s_{\min}/2$). Then

$$|e^{\text{margin}}(c; \hat{S}) - e^{\text{margin}}(c; S)| \leq \frac{1}{s_{\min}} |\Delta A| + \frac{2A(c; S)}{s_{\min}^2} |\Delta S(c)| + \frac{2}{s_{\min}^2} |\Delta A| |\Delta S(c)|. \quad (\text{D.21})$$

In particular, for small perturbations, the last term is second order in the errors.

Proof. By definition,

$$e^{\text{margin}}(c; \hat{S}) - e^{\text{margin}}(c; S) = \frac{A(c; \hat{S})}{\hat{S}(c)} - \frac{A(c; S)}{S(c)}.$$

Write $A(c; \hat{S}) = A(c; S) + \Delta A$ and $\hat{S}(c) = S(c) + \Delta S(c)$:

$$\frac{A(c; \hat{S})}{\hat{S}(c)} - \frac{A(c; S)}{S(c)} = \frac{A(c; S) + \Delta A}{S(c) + \Delta S(c)} - \frac{A(c; S)}{S(c)}.$$

Subtract and re-add $\frac{A(c; S) + \Delta A}{S(c)}$:

$$\frac{A(c; S) + \Delta A}{S(c) + \Delta S(c)} - \frac{A(c; S)}{S(c)} = \underbrace{\left(\frac{A(c; S) + \Delta A}{S(c) + \Delta S(c)} - \frac{A(c; S) + \Delta A}{S(c)} \right)}_I + \underbrace{\left(\frac{A(c; S) + \Delta A}{S(c)} - \frac{A(c; S)}{S(c)} \right)}_{II}. \quad (\text{D.22})$$

Term II equals $\Delta A/S(c)$, hence $|(II)| \leq |\Delta A|/s_{\min}$.

For term I,

$$(I) = (A(c; S) + \Delta A) \left(\frac{1}{S(c) + \Delta S(c)} - \frac{1}{S(c)} \right) = (A(c; S) + \Delta A) \frac{-\Delta S(c)}{S(c)(S(c) + \Delta S(c))}.$$

Taking absolute values and using $S(c) \geq s_{\min}$ and $S(c) + \Delta S(c) \geq s_{\min}/2$ gives

$$|(I)| \leq |A(c; S) + \Delta A| \cdot \frac{2|\Delta S(c)|}{s_{\min}^2} \leq \frac{2A(c; S)}{s_{\min}^2} |\Delta S(c)| + \frac{2}{s_{\min}^2} |\Delta A| |\Delta S(c)|.$$

Combining the bounds for both terms yields (D.21). $\square$

Corollary D.3 (Consistency). *Assume:*

- (a) *Theorem 1 in Rivest and Wells [2001] holds at some $t_0 > c$, i.e., $\hat{S}_E(t)$ is uniformly consistent for $S_E^*(t)$ on $[0, t_0]$;*
- (b) *no tail contribution beyond t_0 ;*
- (c) $S_E^*(c) \geq s_{\min} > 0$.

Then $\hat{e}^{\text{margin}}(c) \xrightarrow{P} e^{*, \text{margin}}(c)$. If the copula model is correctly specified so that $S_E^* = S_E$, then $\hat{e}^{\text{margin}}(c) \xrightarrow{P} e^{\text{margin}}(c; S_E)$.

Proof. Because (b) assumes no tail contribution beyond t_0 , we have

$$\Delta A = \int_c^\infty (\hat{S}_E(u) - S_E^*(u)) du = \int_c^{t_0} (\hat{S}_E(u) - S_E^*(u)) du.$$

Hence

$$\begin{aligned} |\Delta A| &\leq (t_0 - c) \\ \sup_{u \in [c, t_0]} |\hat{S}_E(u) - S_E^*(u)| &\leq (t_0 - c) \\ \sup_{u \in [0, t_0]} |\hat{S}_E(u) - S_E^*(u)| &\xrightarrow{P} 0, \end{aligned}$$

by uniform consistency on $[0, t_0]$. Also $|\Delta S(c)| = |\hat{S}_E(c) - S_E^*(c)| \leq \sup_{u \in [0, t_0]} |\hat{S}_E(u) - S_E^*(u)| \xrightarrow{P} 0$.

Since $\sup_{u \in [0, t_0]} |\hat{S}_E(u) - S_E^*(u)| \xrightarrow{P} 0$ and $s_{\min} > 0$, we have $\Pr(|\Delta S(c)| < \frac{s_{\min}}{2}) \xrightarrow{P} 1$, hence $\hat{S}_E(c) \geq \frac{s_{\min}}{2}$ with probability close to 1.

Then we can apply Lemma D.2 (because we have $|\Delta S(c)| \leq s_{\min}/2$) with $S = S_E^*$ and $\hat{S} = \hat{S}_E$ to conclude $\hat{e}^{\text{margin}}(c) - e^{*, \text{margin}}(c) \xrightarrow{P} 0$. The correctly specified case follows from $S_E^* = S_E$. $\square$

D.3.3 First-order expansion and delta method

In the preceding section, we prove the “law of large numbers property” of the margin time – in the limit, we have asymptotic consistency. Now, we want to demonstrate the convergence rate.

Let us first define some notation for convenience. Fix $c \geq 0$. Define

$$\mathcal{T}_c(S) := e^{\text{margin}}(c; S) = c + \frac{A(c; S)}{S(c)}.$$

Lemma D.4. *Let S be a survival function with $S(c) > 0$ and $A(c; S) < \infty$. Let h be any bounded function on $[c, \infty)$ such that $S(c) + h(c) > 0$ and $A(c; S + h) < \infty$. Write $S_h := S + h$ and define $\Delta A := A(c; S_h) - A(c; S) = \int_c^\infty h(u) du$. Then*

$$\mathcal{T}_c(S_h) - \mathcal{T}_c(S) = \underbrace{\frac{1}{S(c)} \int_c^\infty h(u) du - \frac{A(c; S)}{S(c)^2} h(c)}_{\dot{\mathcal{T}}_{c; S}(h)} + R_{c; S}(h), \quad (\text{D.23})$$

where the remainder is

$$R_{c;S}(h) = -\frac{\Delta A h(c)}{S(c)^2} + \frac{A(c; S_h) h(c)^2}{S(c)^2 (S(c) + h(c))}. \quad (\text{D.24})$$

In particular, if $|h(c)| \leq S(c)/2$, then $R_{c;S}(h) = O(|h(c)|^2 + |\Delta A| |h(c)|)$.

Proof. The proof is identical to the finite-horizon case, replacing $\int_c^{t_0}$ by $\int_c^\infty$. Start from

$$\mathcal{T}_c(S_h) - \mathcal{T}_c(S) = \frac{A(c; S_h)}{S_h(c)} - \frac{A(c; S)}{S(c)}.$$

Add and subtract $A(c; S_h)/S(c)$ to obtain

$$\begin{aligned} \frac{A(c; S_h)}{S_h(c)} - \frac{A(c; S)}{S(c)} &= A(c; S_h) \left(\frac{1}{S_h(c)} - \frac{1}{S(c)} \right) + \frac{A(c; S_h) - A(c; S)}{S(c)} \\ &= A(c; S_h) \left(\frac{1}{S_h(c)} - \frac{1}{S(c)} \right) + \frac{\Delta A}{S(c)}. \end{aligned}$$

Use the identity $\frac{1}{a+b} - \frac{1}{a} = -\frac{b}{a^2} + \frac{b^2}{a^2(a+b)}$ with $a = S(c)$ and $b = h(c)$, and then decompose $A(c; S_h) = A(c; S) + \Delta A$. Collecting terms yields (D.23)–(D.24). $\square$

Theorem 2 in Rivest and Wells [2001] states that, for any t_0 satisfying $\pi(t_0) > 0$ and suitable bounded-derivative conditions, $\sqrt{n}(\hat{S}_E(t) - S_E^*(t))$ converges weakly on $D[0, t_0]$ to a mean-zero Gaussian process, where $\pi(t) = \Pr(\min(E, C) \geq t)$. Because the margin-time functional depends only on $\hat{S}_E$ over $[0, t_0]$, this functional CLT is sufficient.

Theorem D.5. *Assume:*

- (a) Rivest–Wells Theorem 2 holds at some $t_0 > c$, i.e., $\sqrt{n}(\hat{S}_E - S_E^*) \Rightarrow \mathbb{G}_E$ on $D[0, t_0]$, where $\mathbb{G}_E$ has almost surely continuous sample paths;
- (b) no tail contribution beyond t_0 ;
- (c) $S_E^*(c) > 0$.

Then

$$\sqrt{n}(\hat{e}^{\text{margin}}(c) - e^{*, \text{margin}}(c)) \Rightarrow \dot{\mathcal{T}}_{c; S_E^*}(\mathbb{G}_E) = \frac{1}{S_E^*(c)} \int_c^\infty \mathbb{G}_E(u) du - \frac{\int_c^\infty S_E^*(u) du}{(S_E^*(c))^2} \mathbb{G}_E(c). \quad (\text{D.25})$$

Moreover, under no tail contribution assumption, the integrals in (D.25) equal $\int_c^{t_0}(\cdot) du$.

Proof. Let $h_n := \hat{S}_E - S_E^*$. By Lemma D.4 applied with $S = S_E^*$ and $h = h_n$,

$$\hat{e}^{\text{margin}}(c) - e^{*, \text{margin}}(c) = \dot{\mathcal{T}}_{c; S_E^*}(h_n) + R_{c; S_E^*}(h_n).$$

Multiply by $\sqrt{n}$:

$$\sqrt{n}(\hat{e}^{\text{margin}}(c) - e^{*, \text{margin}}(c)) = \dot{\mathcal{T}}_{c; S_E^*}(\sqrt{n}h_n) + \sqrt{n}R_{c; S_E^*}(h_n).$$

Under no tail contribution, $\hat{S}_E(u) = S_E^*(u) = 0$ for $u \geq t_0$ (w.p. $\rightarrow 1$), so $\int_c^\infty h_n(u) du = \int_c^{t_0} h_n(u) du$ and likewise for $A(c; S_E^*)$. Thus $\dot{\mathcal{T}}_{c; S_E^*}$ is linear and is continuous at every path that is continuous at c .

By Theorem 1 in Rivest and Wells [2001], $\hat{S}_E$ is uniformly consistent for S_E^* on $[0, t_0]$, so $\|h_n\|_{\infty, [0, t_0]} \xrightarrow{p} 0$. The remainder in Lemma D.4 is second order in $h_n(c)$ and $\Delta A = \int_c^{t_0} h_n$, hence $\sqrt{n}R_{c; S_E^*}(h_n) = o_p(1)$ under the same tightness argument implied by the functional CLT in (a).

Since $\sqrt{n}h_n \Rightarrow \mathbb{G}_E$ on $D[0, t_0]$ and $\mathbb{G}_E$ has almost surely continuous sample paths, $\dot{\mathcal{T}}_{c; S_E^*}$ is continuous at $\mathbb{G}_E$ almost surely. The continuous mapping theorem therefore yields

$$\dot{\mathcal{T}}_{c; S_E^*}(\sqrt{n}h_n) \Rightarrow \dot{\mathcal{T}}_{c; S_E^*}(\mathbb{G}_E).$$

Combining this with $\sqrt{n}R_{c; S_E^*}(h_n) = o_p(1)$ and Slutsky's theorem gives (D.25). $\square$

E EXPERIMENTAL DETAILS

E.1 SYNTHETIC EXPERIMENT DETAILS

We evaluate the behavior of survival evaluation metrics under controlled censoring mechanisms using a fully synthetic data-generating process (DGP) with known ground-truth event times. The primary goal of these experiments is to isolate the effect of dependent censoring on metric bias while keeping the predictive model fixed.

All experiments use Weibull marginal distributions for both event and censoring times and induce dependence through a copula function. Synthetic datasets are generated using a linear Weibull DGP, and all experimental factors other than the evaluation metric are held fixed.

Covariates. For each experiment, we generate N independent covariate vectors

$$\mathbf{x}_i \sim \text{Unif}([0, 1]^d),$$

with $d = 10$ and $N = 10,000$. Covariates are fixed across different censoring dependence levels for each random seed to isolate the effect of the censoring mechanism.

Event and censoring models. Both the event time E and censoring time C follow Weibull distributions with covariate-dependent hazards. Conditional on $\mathbf{x}_i$, the hazard and survival functions take the form

$$h(t \mid \mathbf{x}_i) = \left(\frac{v}{\rho}\right) \left(\frac{t}{\rho}\right)^{v-1} \exp(g(\mathbf{x}_i)), \quad (\text{E.1})$$

$$S(t \mid \mathbf{x}_i) = \exp\left(-\left(\frac{t}{\rho}\right)^v \exp(g(\mathbf{x}_i))\right). \quad (\text{E.2})$$

We use the following parameters throughout:

$$(v_E, \rho_E) = (4, 17), \quad (v_C, \rho_C) = (3, 12).$$

The risk functions are linear:

$$g_E(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}_E, \quad g_C(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}_C,$$

where $\boldsymbol{\beta}_E, \boldsymbol{\beta}_C \in \mathbb{R}^d$ are sampled independently for each random seed from $\text{Unif}([-1, 1]^d)$.

Copula-based dependence. Dependence between event and censoring times is introduced using copula-based constructions. For each instance i , a pair of uniform random variables

$$(u_i, v_i) \in [0, 1]^2$$

is sampled from a copula with a specified Kendall's τ .

For Archimedean copulas (Clayton and Frank), (u_i, v_i) is sampled directly using the standard Archimedean construction, with the copula parameter chosen to match the target Kendall's τ via the corresponding analytical relationship. For the Gaussian copula, dependence is induced by first sampling a bivariate normal vector (z_{1i}, z_{2i}) with zero mean, unit variance, and correlation coefficient ρ . The correlation ρ is chosen such that the resulting Kendall's τ matches the target value. The uniform variables are then obtained via the probability integral transform:

$$u_i = \Phi(z_{1i}), \quad v_i = \Phi(z_{2i}).$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

When $\tau = 0$, u_i and v_i are sampled independently, yielding conditional independence between E and C given $\mathbf{X}$.

Event and censoring times are subsequently generated via inverse transform sampling:

$$e_i = S_E^{-1}(v_i \mid \mathbf{x}_i), \quad c_i = S_C^{-1}(u_i \mid \mathbf{x}_i).$$

Censoring rate calibration. To ensure that differences in metric behavior are not confounded by changes in censoring severity, the censoring distribution is calibrated to achieve approximately matched marginal censoring rates across copula families and dependence strengths.

This is accomplished by introducing a multiplicative scaling factor on the censoring hazard. For each copula family, a grid of scaling factors is evaluated in pilot simulations, and the resulting empirical censoring rates are recorded. Scaling factors are then selected to span a fixed set of target censoring-rate intervals, ensuring that each copula family contributes comparable censoring levels despite differences in dependence structure.

The selected scaling factors are reused across random seeds, yielding stable censoring rates with low variability while preserving the intended dependence between event and censoring times.

Observed data. The observed time and event indicator are given by

$$t_i = \min(e_i, c_i), \quad \delta_i = \mathbb{I}\{e_i \leq c_i\}.$$

For oracle evaluation, the true event times e_i are retained for all instances.

Model fitting. For each simulated dataset, we fit a CoxPH model using the observed data $(t_i, \delta_i, \mathbf{x}_i)$. Survival curves $\hat{S}(t | \mathbf{x}_i)$ are obtained using the Breslow estimator. The CoxPH model specification and hyperparameters are held fixed across experiments, while the model is refitted for each simulated dataset.

Metric evaluation. We compare two evaluation approaches:

- **IPCW Brier Score**, computed using inverse probability of censoring weights as implemented in the `SurvivalEVAL` package [Qi et al., 2023b].
- **Dependent Brier Score**, computed using the proposed dependent Brier score based on CG margin-time imputation with uncertainty weighting and with the copula family and dependence strength assumed known.

Metric error is defined as the absolute deviation from the oracle Brier score computed using the true event times. All results are averaged over 10 random seeds. Empirical censoring rates are additionally reported to verify calibration stability across dependence levels.

E.2 DATASETS AND PREPROCESSING

WHAS: Worcester Heart Attack Study dataset (WHAS) [Hosmer et al., 2008] contains 500 patients with acute myocardial infarction, focusing on the time to death post-hospital admission. The data was already post-processed and can be downloaded from the `scikit-survival` package [Pölsterl, 2020].

METABRIC: The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset [Curtis et al., 2012] contains survival information for breast cancer patients. This dataset includes a diverse range of feature sets that encompass clinical traits, expression profiles, copy number variation (CNV) profiles, and single-nucleotide polymorphism (SNP) genotypes. All these features are derived from breast tumor samples collected during the METABRIC trial. The dataset can be downloaded from https://www.cbioportal.org/study/summary?id=brca_metabric.

Churn: Customer churn prediction dataset (Churn) focuses on predicting customer attrition. Each record in the file corresponds to a customer and contains the same set of features, such as age, gender, tenure, usage frequency, support calls, payment delay, subscription type, contract length, total spend, and last interaction. We apply one-hot encoding on the US region feature and exclude subjects who are censored at time 0. The dataset can be downloaded from the `PySurvival` package [Fotso et al., 2019].

GBSG: The German Breast Cancer Study Group (GBSG) dataset [Royston and Altman, 2013] includes 686 patients with node-positive breast cancer and complete prognostic information. It is publicly available in R’s `survival` package [Therneau, 2024]. While the original GBSG dataset features a higher censoring rate and a broader set of variables, we use a modified version that combines the dataset with uncensored cases from the Rotterdam dataset. This results in fewer variables, a lower censoring rate, and a larger sample size than the original. This combined dataset, which was already post-processed in the DeepSurv study [Katzman et al., 2018], is available for download at <https://github.com/jaredleekatzman>.

NACD: The Northern Alberta Cancer Dataset (NACD) [Haider et al., 2020] contains 2402 patients and 53 features. It is a conglomerate of many different cancer patients, including lung, colorectal, head and neck, esophageal, stomach, and other cancers. The event of interest in this dataset is failure time. We drop patients with negative or zero survival time leaving 2,396 patients. There are no missing values in this dataset. The dataset can be downloaded from <http://pssp.srv.ualberta.ca> under “Public Predictors”.

FLCHAIN: The Serum Free Light Chain (FLCHAIN) dataset is a stratified random sample comprising half of the participants from a study investigating the association between serum free light chain (FLC) levels and mortality [Dispenzieri et al., 2012]. It is available in R’s `survival` package [Therneau, 2024]. After downloading, we perform several preprocessing steps: we remove three subjects with events occurring at time zero, impute missing values in the `creatinine` feature using its median, and exclude the `chapter` feature—which encodes the cause of death according to ICD chapter headings—since it is only defined for deceased (uncensored) subjects and would therefore leak the event indicator to the model.

SUPPORT: The Study to Understand Prognoses, Preferences, Outcomes, and Risks of Treatment (SUPPORT) dataset [Knaus, 1995] includes 9,105 participants and was designed to investigate survival outcomes and clinical decision-making among seriously ill hospitalized patients. The dataset contains a substantial proportion of missing values across many features. We obtained the dataset from Qi et al. [2024] and follow their preprocessing steps.

Employee: Employee dataset contains employee activity information that can be used to predict when an employee will quit. The dataset can be downloaded from the PySurvival package [Fotso et al., 2019]. It contains duplicate entries; after dropping these duplicates, the number of subjects in the dataset is reduced from 14,999 to 11,991. We also apply one-hot encoding to the department information.

MIMIC-IV: The Medical Information Mart for Intensive Care IV (MIMIC-IV) [Johnson et al., 2023] contains critical care data from patients admitted to hospitals and intensive care units (ICU). We create a version using the MIMIC-IV database, which contains patients who are alive at least 24 hours after being admitted to ICU. Their date of death is derived from hospital records or state records, which means the cause of mortality is not limited to the reason for ICU admission. We follow the instructions in [Qi et al., 2023a] to process the dataset, and the code is available in their GitHub repository: <https://github.com/shi-ang/CensoredMAE>.

SEER: The SEER Program dataset [Ries et al., 2003] is a comprehensive collection of cancer patient data from approximately 49% of the U.S. population. Our study focuses on three subsets: SEER-brain, SEER-liver, and SEER-stomach, containing data on patients diagnosed with brain, liver, and stomach cancers, respectively. The goal is to model the time from diagnosis to failure events, such as death or disease progression. We preprocess the dataset as in [Qi et al., 2024]. Features analyzed include age, sex, cancer stage, grade, and various clinical factors. We excluded features with over 70% missing data, patients without follow-up times, duplicates, and those with a survival time of zero. The SEER cohort is available for download at www.seer.cancer.gov.

Figure E.1 shows the empirical KM survival curves and the distribution of event and censoring times for the generated semi-synthetic datasets used for training and evaluation. Reported censoring rates are computed from the observed indicators.

E.3 SURVIVAL LEARNERS

CoxPH: The Cox Proportional Hazards model (CoxPH) is a semi-parametric survival analysis model [Cox, 1972], consisting of a non-parametric baseline hazard function $h_0(t)$ and a parametric partial hazard function $f(\theta, \mathbf{x}_i)$, typically estimated using the Breslow method [Breslow, 1975]. The hazard function is $h(t | \mathbf{x}_i) = h_0(t) \exp(f(\theta, \mathbf{x}_i))$, where f is linear. The MLE of θ maximizes the Cox partial log-likelihood [Cox, 1972]. CoxPH is implemented from the `scikit-survival` package [Pölsterl, 2020].

GBSA: GBSA [Ridgeway, 1999] is a likelihood-based boosting model that optimizes an ℓ_2 -norm penalized Cox partial log-likelihood [Cox, 1972]. It updates candidate variables flexibly at each boosting step using an offset-based gradient boosting method, unlike traditional gradient boosting, which updates either a single component or all features for fitting the gradient. GBSA is implemented from the `scikit-survival` package [Pölsterl, 2020].

RSF: Random Survival Forests (RSF) [Ishwaran et al., 2008] extend decision trees for survival analysis by constructing multiple trees on bootstrapped samples. Each tree predicts an individual’s survival function, with splits based on criteria like the log-rank test. RSF handles censored data as right-censored and accounts for it during tree construction, with final predictions as the average of all trees’ survival functions. RSF is implemented from the `scikit-survival` package [Pölsterl, 2020].

DeepSurv: DeepSurv [Katzman et al., 2018] is a neural network implementation of the CoxPH model of the form $h(t | \mathbf{x}_i) = h_0(t) \exp(f(\theta, \mathbf{x}_i))$, where $f(\theta, \mathbf{x}_i)$ denotes a risk score as a nonlinear function of the features, *i.e.*, $f(\theta, \mathbf{x}_i) = \sigma(\mathbf{x}_i \theta)$, where σ is a nonlinear activation function. Popular functions include the hyperbolic tangent (Tanh) function or the Rectified Linear Unit (ReLU). The MLE for θ is derived by numerically maximizing the partial Cox log-likelihood. DeepSurv is implemented from <https://github.com/shi-ang/BNN-ISR/>.

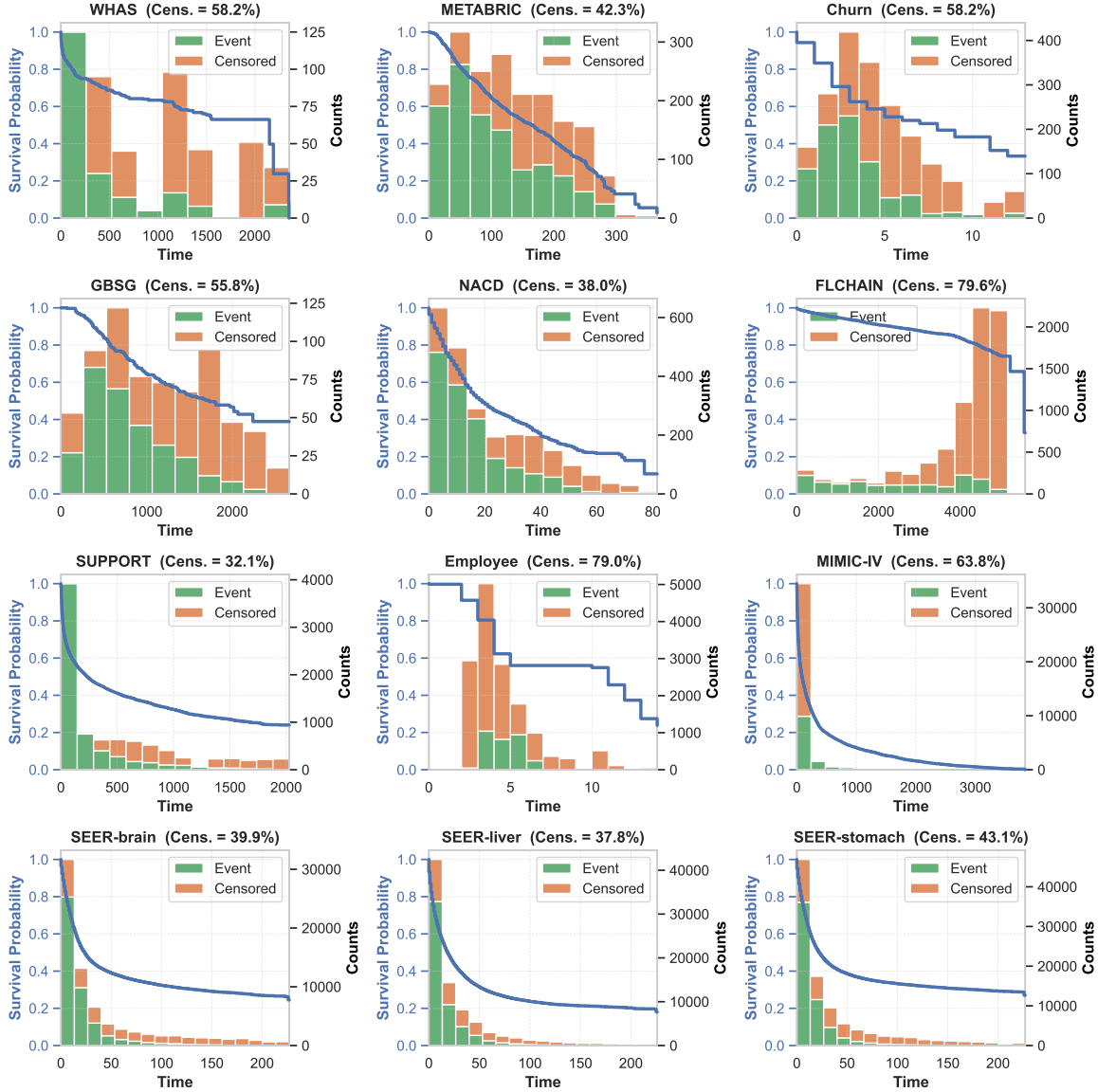


Figure E.1: Empirical KM survival curve and event/censoring time distribution for the generated semi-synthetic datasets used in training and evaluation.

MTLR: Multi-Task Logistic Regression (MTLR) [Yu et al., 2011] is a neural network implementation of a discrete-time model that predicts the survival distribution using a sequence of dependent logistic regressors. The number of discrete times is determined by the square root of the number of uncensored patients, and we use quantiles to divide those uncensored instances evenly into each time interval. MTLR is implemented from <https://github.com/shi-ang/BNN-ISD/>.

Table E.1 reports the selected hyperparameters for the survival learners.

E.4 IMPLEMENTATION DETAILS

Subsampling. For computational efficiency, we subsample the largest semi-synthetic datasets while preserving their event-time distributions. For MIMIC-IV, SEER, and Employee, we draw $N = 10,000$ rows stratified by event status and discretized time intervals.

Reproducibility. Experiments used Python 3.9, PyTorch 1.13.1, NumPy 1.24.3, and Pandas 1.5.3, with tensor computations in double precision (fp64). They were run on a workstation with an AMD Ryzen 5 5600X 3.70 GHz CPU, 32 GB RAM,

Table E.1: Hyperparameters for the survival learners.

Model	Parameter	Value
CoxPH	alpha	0.01
	ties	'Breslow'
	n_iter	100
	tol	1e-9
GBSA	n_estimators	100
	max_depth	1
	min_samples_split	2
	min_samples_leaf	1
	max_features	'sqrt'
	sub_sample	0.8
	random_state	0
RSF	n_estimators	100
	max_depth	1
	min_samples_split	2
	min_samples_leaf	1
	max_features	'sqrt'
	random_state	0
DeepSurv	hidden_size	100
	verbose	False
	lr	0.001
	c1	0.01
	num_epochs	1000
	batch_size	32
	dropout	0.25
	early_stop	True
MTLR	patience	10
	verbose	False
	lr	0.001
	c1	0.01
	num_epochs	1000
	batch_size	32
	early_stop	True
	patience	10

and an NVIDIA GeForce RTX 2060 (6 GB) GPU using CUDA 11.7. Parts of the pipeline and preprocessing utilities build on the *MENSA* codebase [Lillelund et al., 2026a].

Copula estimation details. For semi-synthetic experiments, copula and marginal parameters were optimized using Adam [Kingma and Ba, 2017] for up to 3×10^4 full-batch epochs. Learning rates were 10^{-3} for the Weibull marginals and 10^{-2} for the copula parameter θ . Early stopping was triggered when the validation NLL failed to improve for 100 epochs and θ changed negligibly across successive epochs.

F ADDITIONAL RESULTS

F.1 COMPUTATIONAL ANALYSIS

Time complexity. Table F.1 reports the runtime of copula fitting, the dominant cost of dependent metric evaluation. For each candidate copula family, two Weibull marginals and a copula parameter are fitted jointly, followed by validation-based model selection. Runtime ranges from approximately 5 to 75 minutes. After fitting, the dependent metrics add less than 1ms relative to standard estimators across all datasets.

Space complexity. Table F.1 also reports the increase in peak PyTorch-allocated GPU tensor memory during fitting. This excludes CPU RAM, non-PyTorch allocations, and reserved but unallocated GPU memory. The increase ranges from approximately 0.3 to 11MiB. Fitted models are discarded afterward and therefore do not affect metric-evaluation memory.

F.2 ESTIMATION ERROR

Table F.2 reports the full bias and variability results corresponding to the main text estimation error analysis.

Table F.1: Copula fitting time and incremental peak PyTorch GPU tensor memory across datasets under the *Original* feature selection strategy. Fitting includes training Clayton and Frank copula models jointly with their Weibull marginals.

Dataset	Copula Time (min)	Copula Memory (MiB)
METABRIC	15.67	1.12
GBSG	5.22	0.43
NACD	14.76	1.91
SUPPORT	67.36	5.90
FLChain	55.98	5.40
WHAS	10.61	0.35
Employee	54.41	6.08
Churn	5.59	1.32
MIMIC-IV (all)	74.24	10.68
SEER (brain)	28.97	5.72
SEER (liver)	68.35	5.92
SEER (stomach)	38.01	5.92

Table F.2: Mean ($\pm$ SD) absolute estimation error of IBS-IPCW and proposed dependent metrics across semi-synthetic datasets and feature selection strategies, averaged over 5 survival learners and 10 experiments.

		WHAS	METABRIC	Churn	GBSG	NACD	FLCHAIN
Original	IPCW (KM)	0.073 $\pm$ 0.038	0.061 $\pm$ 0.012	0.117 $\pm$ 0.016	0.125 $\pm$ 0.039	0.092 $\pm$ 0.013	0.156 $\pm$ 0.008
	IPCW (CoxPH)	0.106 $\pm$ 0.070	0.051 $\pm$ 0.012	0.115 $\pm$ 0.019	0.129 $\pm$ 0.040	0.148 $\pm$ 0.106	0.162 $\pm$ 0.008
	Dep (KM)	0.055 $\pm$ 0.025	0.047 $\pm$ 0.011	0.120 $\pm$ 0.012	0.130 $\pm$ 0.038	0.091 $\pm$ 0.010	0.135 $\pm$ 0.008
	Dep (CG)	0.055 $\pm$ 0.025	0.047 $\pm$ 0.011	0.120 $\pm$ 0.013	0.126 $\pm$ 0.037	0.091 $\pm$ 0.010	0.135 $\pm$ 0.008
	Dep (CG) ^{UW}	0.063 $\pm$ 0.025	0.049 $\pm$ 0.010	0.124 $\pm$ 0.014	0.134 $\pm$ 0.037	0.082 $\pm$ 0.011	0.090 $\pm$ 0.008
Top-5	IPCW (KM)	0.071 $\pm$ 0.037	0.062 $\pm$ 0.012	0.121 $\pm$ 0.018	0.124 $\pm$ 0.039	0.103 $\pm$ 0.014	0.133 $\pm$ 0.008
	IPCW (CoxPH)	0.065 $\pm$ 0.037	0.061 $\pm$ 0.013	0.114 $\pm$ 0.016	0.126 $\pm$ 0.038	0.098 $\pm$ 0.014	0.138 $\pm$ 0.008
	Dep (KM)	0.055 $\pm$ 0.024	0.048 $\pm$ 0.011	0.120 $\pm$ 0.013	0.130 $\pm$ 0.038	0.098 $\pm$ 0.010	0.108 $\pm$ 0.008
	Dep (CG)	0.055 $\pm$ 0.024	0.048 $\pm$ 0.011	0.119 $\pm$ 0.014	0.126 $\pm$ 0.037	0.097 $\pm$ 0.012	0.108 $\pm$ 0.008
	Dep (CG) ^{UW}	0.065 $\pm$ 0.024	0.049 $\pm$ 0.010	0.125 $\pm$ 0.016	0.134 $\pm$ 0.037	0.089 $\pm$ 0.012	0.066 $\pm$ 0.008
Top-10	IPCW (KM)	0.071 $\pm$ 0.038	0.061 $\pm$ 0.011	0.119 $\pm$ 0.016	0.125 $\pm$ 0.039	0.098 $\pm$ 0.013	0.143 $\pm$ 0.013
	IPCW (CoxPH)	0.082 $\pm$ 0.051	0.051 $\pm$ 0.012	0.111 $\pm$ 0.015	0.130 $\pm$ 0.041	0.093 $\pm$ 0.017	0.150 $\pm$ 0.014
	Dep (KM)	0.054 $\pm$ 0.025	0.047 $\pm$ 0.011	0.121 $\pm$ 0.012	0.131 $\pm$ 0.038	0.095 $\pm$ 0.010	0.123 $\pm$ 0.017
	Dep (CG)	0.054 $\pm$ 0.025	0.047 $\pm$ 0.011	0.121 $\pm$ 0.012	0.126 $\pm$ 0.037	0.095 $\pm$ 0.010	0.123 $\pm$ 0.017
	Dep (CG) ^{UW}	0.063 $\pm$ 0.025	0.049 $\pm$ 0.011	0.125 $\pm$ 0.014	0.134 $\pm$ 0.037	0.086 $\pm$ 0.011	0.078 $\pm$ 0.009
Rand. 25%	IPCW (KM)	0.090 $\pm$ 0.030	0.061 $\pm$ 0.018	0.129 $\pm$ 0.014	0.153 $\pm$ 0.032	0.109 $\pm$ 0.013	0.140 $\pm$ 0.020
	IPCW (CoxPH)	0.091 $\pm$ 0.031	0.056 $\pm$ 0.018	0.122 $\pm$ 0.013	0.157 $\pm$ 0.036	0.111 $\pm$ 0.017	0.147 $\pm$ 0.013
	Dep (KM)	0.081 $\pm$ 0.022	0.049 $\pm$ 0.017	0.126 $\pm$ 0.009	0.153 $\pm$ 0.026	0.101 $\pm$ 0.010	0.140 $\pm$ 0.012
	Dep (CG)	0.081 $\pm$ 0.022	0.049 $\pm$ 0.017	0.117 $\pm$ 0.009	0.147 $\pm$ 0.026	0.099 $\pm$ 0.010	0.140 $\pm$ 0.012
	Dep (CG) ^{UW}	0.095 $\pm$ 0.026	0.050 $\pm$ 0.017	0.131 $\pm$ 0.009	0.161 $\pm$ 0.026	0.091 $\pm$ 0.010	0.032 $\pm$ 0.013
		SUPPORT	Employee	MIMIC (IV)	SEER (brain)	SEER (liver)	SEER (stomach)
Original	IPCW (KM)	0.108 $\pm$ 0.052	0.152 $\pm$ 0.009	0.026 $\pm$ 0.016	0.077 $\pm$ 0.005	0.104 $\pm$ 0.013	0.119 $\pm$ 0.018
	IPCW (CoxPH)	0.109 $\pm$ 0.052	0.100 $\pm$ 0.006	0.039 $\pm$ 0.027	0.055 $\pm$ 0.005	0.097 $\pm$ 0.015	0.095 $\pm$ 0.012
	Dep (KM)	0.111 $\pm$ 0.051	0.148 $\pm$ 0.007	0.023 $\pm$ 0.012	0.089 $\pm$ 0.005	0.118 $\pm$ 0.004	0.125 $\pm$ 0.008
	Dep (CG)	0.109 $\pm$ 0.051	0.154 $\pm$ 0.007	0.024 $\pm$ 0.012	0.084 $\pm$ 0.006	0.106 $\pm$ 0.004	0.112 $\pm$ 0.008
	Dep (CG) ^{UW}	0.093 $\pm$ 0.051	0.152 $\pm$ 0.008	0.024 $\pm$ 0.013	0.067 $\pm$ 0.005	0.092 $\pm$ 0.004	0.097 $\pm$ 0.008
Top-5	IPCW (KM)	0.109 $\pm$ 0.052	0.155 $\pm$ 0.010	0.027 $\pm$ 0.017	0.076 $\pm$ 0.005	0.109 $\pm$ 0.014	0.124 $\pm$ 0.018
	IPCW (CoxPH)	0.109 $\pm$ 0.052	0.103 $\pm$ 0.005	0.027 $\pm$ 0.018	0.083 $\pm$ 0.022	0.109 $\pm$ 0.017	0.128 $\pm$ 0.020
	Dep (KM)	0.112 $\pm$ 0.051	0.149 $\pm$ 0.007	0.026 $\pm$ 0.011	0.090 $\pm$ 0.005	0.120 $\pm$ 0.005	0.126 $\pm$ 0.008
	Dep (CG)	0.109 $\pm$ 0.051	0.156 $\pm$ 0.007	0.027 $\pm$ 0.013	0.085 $\pm$ 0.005	0.111 $\pm$ 0.005	0.116 $\pm$ 0.008
	Dep (CG) ^{UW}	0.094 $\pm$ 0.051	0.154 $\pm$ 0.008	0.025 $\pm$ 0.013	0.068 $\pm$ 0.005	0.096 $\pm$ 0.005	0.101 $\pm$ 0.008
Top-10	IPCW (KM)	0.109 $\pm$ 0.052	0.152 $\pm$ 0.009	0.026 $\pm$ 0.017	0.077 $\pm$ 0.005	0.103 $\pm$ 0.013	0.119 $\pm$ 0.017
	IPCW (CoxPH)	0.109 $\pm$ 0.051	0.101 $\pm$ 0.005	0.031 $\pm$ 0.026	0.055 $\pm$ 0.005	0.097 $\pm$ 0.015	0.099 $\pm$ 0.015
	Dep (KM)	0.111 $\pm$ 0.051	0.147 $\pm$ 0.007	0.023 $\pm$ 0.011	0.090 $\pm$ 0.005	0.118 $\pm$ 0.004	0.124 $\pm$ 0.008
	Dep (CG)	0.109 $\pm$ 0.051	0.153 $\pm$ 0.007	0.024 $\pm$ 0.011	0.085 $\pm$ 0.006	0.107 $\pm$ 0.004	0.111 $\pm$ 0.008
	Dep (CG) ^{UW}	0.093 $\pm$ 0.051	0.152 $\pm$ 0.008	0.025 $\pm$ 0.013	0.068 $\pm$ 0.005	0.093 $\pm$ 0.004	0.097 $\pm$ 0.009
Rand. 25%	IPCW (KM)	0.142 $\pm$ 0.058	0.161 $\pm$ 0.016	0.016 $\pm$ 0.009	0.111 $\pm$ 0.027	0.128 $\pm$ 0.015	0.132 $\pm$ 0.017
	IPCW (CoxPH)	0.142 $\pm$ 0.058	0.110 $\pm$ 0.010	0.017 $\pm$ 0.011	0.111 $\pm$ 0.028	0.197 $\pm$ 0.105	0.133 $\pm$ 0.011
	Dep (KM)	0.145 $\pm$ 0.057	0.137 $\pm$ 0.008	0.031 $\pm$ 0.013	0.113 $\pm$ 0.017	0.127 $\pm$ 0.004	0.135 $\pm$ 0.011
	Dep (CG)	0.141 $\pm$ 0.057	0.120 $\pm$ 0.008	0.032 $\pm$ 0.014	0.102 $\pm$ 0.016	0.109 $\pm$ 0.005	0.120 $\pm$ 0.013
	Dep (CG) ^{UW}	0.125 $\pm$ 0.057	0.147 $\pm$ 0.009	0.021 $\pm$ 0.007	0.089 $\pm$ 0.017	0.098 $\pm$ 0.004	0.106 $\pm$ 0.013