# ISOTONIC SURVIVAL REGRESSION: CALIBRATED SURVIVAL DISTRIBUTIONS FROM DEEP COX MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Time-to-event data is widespread across the life sciences and engineering, but it is typically encountered together with censoring, which complicates the application of standard machine learning methods. Deep Cox models have emerged as a popular method for analyzing time-to-event data because they gracefully handle censoring and can be used with unstructured data such as clinical text reports, genomic sequences, and pathology images. However, their predicted survival probabilities are often poorly calibrated, thus limiting their practical utility. In this paper, we propose a novel post hoc calibration method for Deep Cox models that uses isotonic regression to refine predicted survival probabilities without affecting discriminative power. We establish favorable theoretical guarantees, including a double-robustness property and asymptotic calibration. Experiments on synthetic and real-world clinical data demonstrate the empirical effectiveness of our method.

## 1 INTRODUCTION

Survival analysis uses statistical methods to predict the probability that an event time exceeds a specified time (Klein and Moeschberger, 1997; Kalbfleisch and Prentice, 2002). Accurate prediction of these survival probabilities has vital applications in medicine, epidemiology, and engineering (Masciocchi et al., 2022; Noone et al., 2017; Papathanasiou et al., 2023; Luo et al., 2021). The unique challenge in this setting is *right-censoring*, where for some instances, we only observe that the event time exceeds a known censoring time (Leung et al., 1997; Collett, 2014; Lagakos, 1979; Flynn, 2012).

Historically, the Cox proportional hazards model has been an extremely popular survival modeling approach due to its flexibility and interpretability (Cox, 1972; Breslow, 1975; Schober and Vetter, 2018; Van Ness et al., 2023). Recently, modern applications involving high-dimensional, unstructured data, such as medical images (Wang et al., 2024a; Lu et al., 2024a), genomic measurements (Hao et al., 2024), and electronic health records (Rasmy et al., 2021) have motivated the *Deep Cox model* (Katzman et al., 2018), which replaces the linear risk predictor with a neural network. Deep Cox models have since become a dominant method in contemporary biomedical applications (Kang et al., 2025; Lu et al., 2021; Zhang et al., 2025; Vaidya et al., 2025; Ding et al., 2025; Chen et al., 2024; Lu et al., 2024b; 2023; Wang et al., 2024b; Zhao et al., 2021).

For informed decision-making in safety-critical domains, good discrimination alone is not sufficient. Models must also be *calibrated*, meaning that predicted probabilities reflect the true event frequencies (Guo et al., 2017). Recent work establishes that deep learning methods are often poorly calibrated (Minderer et al., 2021; Thulasidasan et al., 2019; Nixon et al., 2019), a concern that extends to Deep Cox models (Kvamme et al., 2019).

A popular approach to post hoc calibration is isotonic regression because it learns a flexible mapping from predicted to calibrated probabilities while preserving performance (Zadrozny and Elkan, 2002). Yet, extending it to survival analysis remains challenging, as it requires targeting an entire distributional forecast while accounting for censoring.

**Our contributions.** In this work, we study the problem of calibrating survival predictions from classical Cox and Deep Cox models. We propose a lightweight post hoc calibration procedure based on isotonic regression that uses estimated risk scores together with an initial survival and censoring model. Our main contributions are as follows: (i) we introduce a novel post hoc method for calibrating



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

survival probability predictions under right-censoring while preserving the discriminative ranking up to potential ties, (ii) we establish theoretical guarantees for the calibration and consistency of our algorithm, including a double-robustness result, and (iii) we demonstrate empirically on synthetic and real-world datasets that our method improves calibration for Cox and Deep Cox models even when theoretical assumptions are likely violated.

## 2 PRELIMINARIES

We introduce the notation and preliminary background for survival analysis and calibration.

**Survival Analysis.** Let $X \in \mathcal{X}$ denote the covariates for an individual, $T$ denote the time until the event of interest occurs, and $C$ denote the censoring time. Due to right-censoring, we only observe the time $Y = \min(T, C)$, together with the event indicator $\delta = \mathbf{1}[T \leq C]$. The goal of survival prediction is to estimate the conditional survival function $S(t \mid X) = \mathbb{P}(T > t \mid X)$, which gives the probability that the event time exceeds $t$ given $X$. The survival function can also be written using the hazard function $\lambda_{T|X}(t \mid X) = -\frac{d}{dt} \log S(t \mid X)$, which describes the instantaneous event rate at time $t$ among individuals who have survived up to that time. Similarly, we define the censoring function $G(t \mid X) = \mathbb{P}(C > t \mid X)$ as the probability that the censoring time exceeds $t$ given $X$.

**Cox and Deep Cox models.** The Cox model is a popular semi-parametric approach for learning the survival function. The model assumes that the hazard function can be represented in the form

$$\lambda(t \mid X) = \lambda_0(t) \exp(r_\theta(X))$$

where $\lambda_0(t)$ is the baseline hazard and $r_\theta(X) = \theta^\top X$ is a covariate-dependent risk score. Deep Cox models (Katzman et al., 2018) replace the linear risk with a neural network $r_\theta : \mathcal{X} \to \mathbb{R}$. Intuitively, the model scales a shared baseline hazard, so higher-risk individuals experience the event at a faster rate. Typically, the parameters $\theta$ are estimated by maximizing the partial likelihood while the baseline hazard is estimated using Breslow's estimator (Breslow, 1972).

A common way to evaluate Cox models is through discrimination, or the ability to correctly rank individuals by risk. This is typically quantified using Harrell's concordance index (C-index) (Harrell Jr. et al., 1996), which estimates the frequency with which among comparable pairs of samples, the one that experiences the event earlier is assigned a higher predicted risk.

**Calibration.** We desire survival models that are well-calibrated, meaning that predicted probabilities are statistically reliable. To evaluate this, we focus on two related notions of calibration. The first is *threshold calibration* (Henzi et al., 2021b), which requires that for all $t \geq 0$ and predicted probabilities $p$, $\mathbb{P}(T > t \mid \widehat{S}(t \mid X) = p) = p$ almost surely. Another notion is *PIT calibration*, based on the probability integral transform (PIT). We say a survival function $\widehat{S}(t \mid X)$ is PIT calibrated if the corresponding PIT values $\widehat{F}(T \mid X) = 1 - \widehat{S}(T \mid X)$ are uniformly distributed on $[0, 1]$.

**Isotonic regression.** Given data $\{(x_i, y_i)\}_{i=1}^n$, one-dimensional isotonic regression estimates a monotonic function by solving $f_{\text{IR}} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^n |f(x_i) - y_i|^2$, where $\mathcal{F}$ is the set of non-increasing functions. Isotonic regression is widely used for post hoc calibration in classification tasks (Zadrozny and Elkan, 2002), since it typically preserves model accuracy. Algorithmically, this problem can be solved efficiently using the pool adjacent violators algorithm (PAVA) (Barlow, 1972).

## 3 DISTRIBUTIONAL REGRESSION FOR SURVIVAL ANALYSIS

We now introduce two methods for estimating the conditional survival function under right censoring using isotonic regression. We begin with a simpler reweighting estimator and then proceed to our primary proposed calibration algorithm, which uses a more sophisticated doubly robust estimator.

Suppose we have fit a Deep Cox model with learned risk score $r_\theta$. Since the predicted risks determine the discriminative power of the model, our goal is to calibrate the survival function without altering the ordering induced by the learned risks. We therefore restrict our attention to estimators $S(t \mid X)$ that are non-increasing in time $t$ and in the predicted risk $r_\theta(X)$, so that individuals assigned higher risk have uniformly lower survival probabilities (up to ties). We call this class of survival curves *valid* and refer to our general method as Isotonic Survival Regression (ISR).

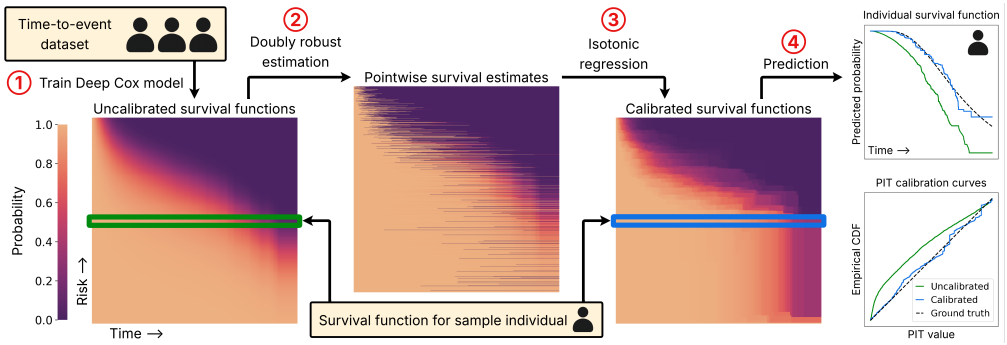



Figure 1: Our DR-ISR calibration method first uses uncalibrated estimates from an initial Deep Cox model to construct doubly robust pointwise survival estimates. We then use isotonic regression to project onto the set of valid survival functions that are monotonic in both time and risk. We demonstrate our approach improves the distributional accuracy and calibration of survival forecasts.

Let $(X_i, Y_i, \delta_i)_{i=1}^n$ denote the observed data, and let $S_0$ and $G_0$ denote the true survival and censoring functions, respectively. Following the standard post hoc calibration paradigm, we split the data into a training set $\mathcal{D}_{\text{train}}$ and a calibration set $\mathcal{D}_{\text{cal}}$. The training set is used to fit an initial survival and censoring model, while the calibration set is reserved for constructing the final calibrated estimator.

### 3.1 Reweighted Isotonic Survival Regression (RW-ISR)

We now turn to a relatively simple reweighting estimator. To this end, consider first predicting the binary outcome $\mathbf{1}[T > t]$. In the absence of censoring, predicting the survival function can be done with isotonic regression. However, censoring causes some of the indicators to be unobserved. One popular approach to handling missing data is Inverse Probability of Censoring Weighting (IPCW), which upweights observed samples depending on their probability of being censored. To apply this approach, we fit a censoring model $\widehat{G}$ on the training data, for example using a Cox model with Breslow's estimate for the baseline hazard function. With this in hand, we define the RW-ISR estimator as the solution to the following weighted isotonic regression problem:

$$\widehat{S}^{\text{RW}}_{\text{ISR}}(t \mid x) = f_t(r_\theta(x)), \quad \text{where } f_t = \underset{f:\text{non-increasing}}{\arg\min} \sum_{X_i \in \mathcal{D}_{\text{cal}}} \frac{\delta_i (f(r_\theta(X_i)) - \mathbf{1}[Y_i > t])^2}{\widehat{G}(Y_i \mid X_i)}. \quad (1)$$

To extend this idea to estimate the entire survival curve, we can solve the isotonic regression for each time $t$. Since the target labels $\mathbf{1}[Y_i > t]$ are non-increasing across $t$, the resulting estimator is non-increasing in time as well. Thus, the estimator yields a valid survival curve.

**Shortcomings.** RW-ISR has two main weaknesses: (i) the estimator depends only on uncensored individuals, thereby discarding information from censored cases, and (ii) its calibration depends on the quality of the estimated censoring probabilities. In missing-data problems, parameter estimation under potentially misspecified weights is commonly addressed using *doubly robust* estimators, which are more robust to imperfection in the reweighting model and remain consistent if either the outcome model or the weight model is correctly specified. We turn to such an approach next.

### 3.2 Doubly Robust Isotonic Survival Regression (DR-ISR)

While RW-ISR provides a simple censoring-adjusted calibration procedure, we can improve upon it using a doubly robust alternative. To do so, we first construct pointwise estimates of the survival probabilities, which we refer to as *pseudo-outcomes*, over the grid $(X, t) \in \mathcal{D}_{\text{cal}} \times \mathcal{T}$, where $\mathcal{T}$ is a time grid, either user-specified or induced by the natural discretization of observed events (e.g., days). We then project these estimates onto the set of valid survival functions using isotonic regression, as shown in Figure 1. DR-ISR both uses more information from censored cases and is less sensitive to misspecification of the censoring model compared to RW-ISR (see Section 4 for formal results).

**Pseudo-outcome construction.** We construct pseudo-outcomes by estimating pointwise survival probabilities using the augmented IPCW estimator motivated by standard semiparametric theory (Robins and Rotnitzky, 1992; Tsiatis, 2006; Farina et al., 2026). The estimator starts with an initial estimate of the survival function $\widehat{S}$ and censoring probability $\widehat{G}$, then adds an augmentation term based on the observed calibration data. The correction is carefully constructed by measuring the error in the models $\widehat{S}$ and $\widehat{G}$ across time. This can be done in a way that is doubly robust, and thus is less sensitive to model misspecification.

To formulate the estimator, define the event counting process $dN_T(u) = \mathbf{1}[Y \in du, \delta = 1]$, which indicates whether an uncensored event is observed at time $u$, and the estimated cumulative hazard function $\widehat{\Lambda}_{T|X}(u \mid X) = -\log \widehat{S}(u \mid X)$. We then define the following event difference process $d\widehat{M}_{T|X}(u \mid X) = dN_T(u) - \mathbf{1}[Y \geq u]\, d\widehat{\Lambda}_{T|X}(u \mid X)$, whose expectation tracks the difference between the frequency of observed events and the model-based probabilities. Note that $d\widehat{M}_{T|X}$ is a random variable, while $\widehat{\Lambda}_{T|X}$ is fixed. We propose the following doubly robust pointwise survival estimate as the pseudo-outcome:

$$\tilde{S}^{\text{DR}}\big(t \mid X; \widehat{S}, \widehat{G}\big) = \widehat{S}(t \mid X) - \int_0^t \frac{\widehat{S}(t \mid X)}{\widehat{S}(u \mid X)\widehat{G}(u \mid X)}\, d\widehat{M}_{T|X}(u \mid X). \tag{2}$$

The integral is a residual correction accumulated over times $u \leq t$ that adjusts the initial model when the observed event process differs from the model predictions. Similar to RW-ISR, we estimate $\widehat{S}$ and $\widehat{G}$ on the training data, for example using a Cox model with Breslow's estimator of the baseline hazard function. Note that the difference process is a discrete measure for piecewise constant survival estimators, including Cox models. Thus, the integral reduces to a discrete sum over times in $\mathcal{T}$.

**From pseudo-outcomes to distributional forecasts.** First, the pseudo-outcomes $\tilde{S}^{\text{DR}}\big(t \mid X; \widehat{S}, \widehat{G}\big)$ are separately constructed for each $X \in \mathcal{D}_{\text{cal}}$ and $t \in \mathcal{T}$. Although the conditional expectation of the pseudo-outcome does target $S_0(t \mid X)$ under correct specification of either $S_0$ or $G_0$ (Theorem 1), the pseudo-outcomes do not directly yield satisfactory conditional survival probabilities because each is constructed with only a single observation. As such, the pseudo-outcomes can fluctuate rapidly across nearby covariate values (see Figure 1).

To correct for this behavior, we can regress the pseudo-outcomes on the risk, thereby sharing information across similar covariates. This smoothing step improves statistical efficiency and stabilizes the otherwise noisy pointwise estimates. Moreover, the pseudo-outcomes do not necessarily define a valid family of survival functions and may have worse discriminative performance than the initial survival model. We address both of these issues by projecting the pseudo-outcomes onto the class of survival functions that are monotonic in both time and risk.

Let $\mathcal{F}$ denote the class of functions $f : \mathbb{R}_{\geq 0} \times \mathbb{R} \to [0, 1]$, where the two arguments correspond to time and risk. Let $\mathcal{F}_{\leq} \subset \mathcal{F}$ denote the subset of functions that are non-increasing in both coordinates. We define the calibrated survival estimate as $\widehat{S}^{\text{DR}}_{\text{ISR}}(t \mid x) = f(t, r_\theta(x))$, where $f$ solves

$$f = \underset{f \in \mathcal{F}_{\leq}}{\arg\min} \sum_{t \in \mathcal{T}} \sum_{X_i \in \mathcal{D}_{\text{cal}}} \left[ \Big( f(t, r_\theta(X_i)) - \tilde{S}^{\text{DR}}\big(t \mid X_i; \widehat{S}, \widehat{G}\big) \Big)^2 \right]. \tag{3}$$

While the two-dimensional isotonic regression optimization in Equation 3 is convex, solving it with a generic convex optimization solver requires $\mathcal{O}(n|\mathcal{T}|)$ constraints, which can be computationally burdensome. Instead, we apply Dykstra's projection algorithm (Dykstra and Robertson, 1982), which exploits that $\mathcal{F}_{\leq}$ can be written as the intersection of two convex sets: non-increasing functions in time and non-increasing functions in risk. We optimize by alternating between projections onto the two sets. Each projection reduces to a collection of one-dimensional isotonic regression problems along time or risk, which can then be computed efficiently using PAVA (Barlow, 1972).

Putting both steps together, our calibration procedure is outlined in Algorithm 1. The time complexity is $\mathcal{O}(T_{\text{Dyk}} \cdot n|\mathcal{T}|)$, where $T_{\text{Dyk}}$ is the number of Dykstra iterations. In practice, PAVA is extremely fast which makes DR-ISR a lightweight post hoc calibration procedure that is practical to deploy on top of modern Deep Cox models. We also highlight that the calibration procedure does not require tuning any hyperparameters, making it especially straightforward to deploy.

**Algorithm 1** DR-ISR calibration algorithm

**Require:** Survival $\widehat{S}$, censoring $\widehat{G}$, and risk $r_\theta$ models from an initial Deep Cox model, $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i, \delta_i)\}_{i=1}^{n_{\text{cal}}}$ sorted by increasing risk, time grid $\mathcal{T} = \{t_1, \ldots, t_K\}$

**for** $i \in [K]$ and $j \in [n_{\text{cal}}]$ **do** ▷ Doubly robust point estimation

  $d\widehat{\Lambda}_{T|X}(t_i \mid X_j) \leftarrow 1 - \widehat{S}(t_i \mid X_j)/\widehat{S}(t_{i-1} \mid X_j)$ ▷ Define $\widehat{S}(t_0 \mid X_j) = 1$

  $d\widehat{M}_{T|X}(t_i \mid X_j) \leftarrow \mathbf{1}[Y_j = t_i, \delta_j = 1] - \mathbf{1}[Y_j \geq t_i]\, d\widehat{\Lambda}_{T|X}(t_i \mid X_j)$

  $\tilde{S}^{\text{DR}}[i, j] \leftarrow \widehat{S}(t_i \mid X_j) - \sum_{k=1}^{i} \frac{\widehat{S}(t_i|X_j)}{\widehat{S}(t_k|X_j)\widehat{G}(t_k|X_j)} d\widehat{M}_{T|X}(t_k \mid X_j)$ ▷ Equation 2

**end for**

Initialize $S \leftarrow \tilde{S}_{\text{DR}}$ and residuals $\epsilon_{\text{time}} \leftarrow \mathbf{0}^{|\mathcal{T}| \times n_{\text{cal}}}$, $\epsilon_{\text{risk}} \leftarrow \mathbf{0}^{|\mathcal{T}| \times n_{\text{cal}}}$ ▷ Dykstra's algorithm

**while** not converged **do**

  $U \leftarrow S + \epsilon_{\text{time}}, \quad S[i, :] \leftarrow \texttt{PAVA}(U[i, :]) \ \ \forall i, \quad \epsilon_{\text{time}} \leftarrow U - S$ ▷ Monotonicity in risk

  $V \leftarrow S + \epsilon_{\text{risk}}, \quad S[:, j] \leftarrow \texttt{PAVA}(V[:, j]) \ \ \forall j, \quad \epsilon_{\text{risk}} \leftarrow V - S$ ▷ Monotonicity in time

**end while**

Interpolate $S$ to form 2D function $\hat{f}$, then make predictions $\widehat{S}^{\text{DR}}_{\text{ISR}}(t \mid X_{\text{test}}) \leftarrow \hat{f}(t, r_\theta(X_{\text{test}}))$

## 4 Theoretical Guarantees

We next describe the theoretical properties of DR-ISR. First, we formalize the double-robustness property of the pseudo-outcomes (Theorem 1). We then show that this property is preserved by the isotonic regression projection step, so that our procedure remains consistent if either the event model or the censoring model converges locally uniformly (Theorem 2). Lastly, we show that even when the risk score is incorrect, the procedure still yields calibrated survival curves (Theorem 3). Our results are stated with respect to a *fixed* risk mapping $r : \mathcal{X} \to \mathbb{R}$ and will refer to the following two assumptions:

A1. The censoring and event times are conditionally independent, i.e. $T \perp\!\!\!\perp C \mid X$.

A2. The risk scores induce a monotone ordering over survival functions, i.e.

$$r(x) \geq r(x') \implies \mathbb{P}(T > t \mid X = x) \leq \mathbb{P}(T > t \mid X = x'), \quad \text{for all } t \geq 0.$$

Assumption A1 is a standard identifiability condition in survival analysis, while assumption A2 imposes the additional structure that the risk score correctly orders and is a sufficient statistic for the survival probability. Note that Assumption A2 does not demand proportional hazard rates and is hence strictly weaker than correct specification of the Cox model with the risk function $r$.

We first make the double-robustness property of the point estimates from Equation 2 precise. Let $\Lambda_{C|X}(u \mid X) = -\log G(u \mid X)$ be the censoring analog of the cumulative hazard function.

**Theorem 1** (Pseudo-outcome Double-robustness)**.** *Suppose that assumption A1 holds and we also have the mild regularity condition S1 from Appendix A. Define the bias of the pseudo-outcomes as* $B\big(t, x; \widehat{S}, \widehat{G}\big) = |\mathbb{E}[\tilde{S}^{\text{DR}}\big(t \mid X; \widehat{S}, \widehat{G}\big) \mid X = x] - S_0(t \mid x)|$*. Then, we have that* $B\big(t, x; \widehat{S}, \widehat{G}\big) \leq \min\{B_1\big(t, x; \widehat{S}, \widehat{G}\big), B_2\big(t, x; \widehat{S}, \widehat{G}\big)\}$*, where the bounds* $B_1$ *and* $B_2$ *admit the following mixed-product form,*

$$B_1\big(t, x; \widehat{S}, \widehat{G}\big) = c_1(t) \int_0^t \left| \frac{1}{\widehat{G}(u \mid x)} - \frac{1}{G_0(u \mid x)} \right| \left| d\Lambda_{0,T|X}(u \mid x) - d\widehat{\Lambda}_{T|X}(u \mid x) \right|, \quad (4)$$

$$B_2\big(t, x; \widehat{S}, \widehat{G}\big) = c_2(t) \int_0^t \left| \frac{\widehat{S}(t \mid x)}{\widehat{S}(u \mid x)} - \frac{S_0(t \mid x)}{S_0(u \mid x)} \right| \left| d\Lambda_{0,C|X}(u \mid x) - d\widehat{\Lambda}_{C|X}(u \mid x) \right|. \quad (5)$$

*Here,* $c_1(t)$ *and* $c_2(t)$ *are finite positive constants depending on* $t$ *as defined in Assumption S1. In particular, if either* $\widehat{S} = S_0$ *or* $\widehat{G} = G_0$*, then the DR-ISR pseudo-outcome is conditionally unbiased for* $S_0(t \mid x)$*.*

The double-robustness property similarly extends to uniform consistency of the survival function estimator when using a sufficiently informative and correctly ordered fixed score mapping, and under locally uniformly consistent estimation of either $S_0$ or $G_0$.

**Theorem 2** (Survival Function Double-robustness). *Suppose that assumptions A1–A2 hold and we also have the mild regularity conditions S1–S3 from Appendix A. Assume either $\widehat{S} \to S_0$ or $\widehat{G} \to G_0$ locally uniformly, i.e. for all times $t \in \mathcal{T}$ and covariates $x$, we have*

$$\sup_{u\in[0,t]} |\widehat{S}(u \mid x) - S_0(u \mid x)| \xrightarrow{p} 0, \quad \text{or} \quad \sup_{u\in[0,t]} |\widehat{G}(u \mid x) - G_0(u \mid x)| \xrightarrow{p} 0.$$

*For any compact $\mathcal{K} \subset \mathrm{int}(r(\mathcal{X}))$, we have,*

$$\sup_{x\in\mathcal{X}: r(x)\in\mathcal{K}} |\widehat{S}^{\mathrm{DR}}_{\mathrm{ISR}}(t \mid x) - S_0(t \mid x)| \xrightarrow{p} 0, \quad \text{for all } t \in \mathcal{T}.$$

Theorem 2 provides a consistency (asymptotic conditional calibration) guarantee since the estimator converges to the true survival function. In practice the risk scores are estimates from the model we wish to calibrate. These may be incorrect or insufficiently characterize the true survival probabilities, thereby violating Assumption A2. Nonetheless, DR-ISR is still well-calibrated in the following sense:

**Theorem 3** (Calibration under Risk Misspecification). *Suppose all assumptions from Theorem 2 hold except Assumption A2. Define $\widehat{S}^{\infty}_{\text{DR-ISR}}(t \mid x) = \lim_{n\to\infty} \widehat{S}^{\mathrm{DR}}_{\mathrm{ISR}}(t \mid x)$ as the pointwise, in-probability limit of the DR-ISR estimator. If either $\widehat{S} \to S_0$ or $\widehat{G} \to G_0$ locally uniformly, then for almost all $p$ in the range of $\widehat{S}^{\infty}_{\text{DR-ISR}}$,*

$$\mathbb{P}\big(T > t \;\big|\; \widehat{S}^{\infty}_{\text{DR-ISR}}(t, X) = p\big) = p, \qquad \text{for all } t \in \mathcal{T}. \tag{6}$$

Intuitively, Theorem 3 guarantees that the average of true survival probabilities over all individuals that are assigned a probability $p$ by the DR-ISR estimator will also be $p$, even if the risk scores do not satisfy Assumption A2. This is a form of threshold calibration (Sec. 2).

Lastly, we comment on the discrimination (C-index) of the calibrated model. Although our algorithm preserves the risk ordering induced by the Cox model, the isotonic projection may pool patients with distinct risk levels to receive identical predicted survival curves. However, because ISR enforces monotonicity in risk, this pooling can only introduce ties, never inversions.

## 5 Related Work

**Conformal methods for survival analysis.** Much of the recent work on statistically valid survival prediction has focused on *conformal prediction* (Candès et al., 2023; Gui et al., 2024; Farina et al., 2026). Rather than calibration, the typical goal in this setting is to construct a valid *lower prediction bound* for the event time, i.e., a time beyond which an individual is expected to survive with probability at least $1 - \alpha$. We defer discussion of this complementary line of work to Appendix B.

**Post hoc calibration of survival distributions.** Turning to the calibration of probability estimates, the methods most closely related to ours are CSD (Qi et al., 2024b) and CiPOT (Qi et al., 2024a). CSD calibrates survival curves by discretizing them at multiple percentile levels and applying conformal quantile regression at each level. CiPOT extends this line of work by directly mapping the empirical distribution of PIT values to the uniform scale. CSD handles censoring by sampling missing event times from an estimated survival model, while CiPOT directly samples on the PIT scale. We elaborate on the guarantees and properties of these methods in Appendix B. Concurrent to our work, Bladt et al. (2026) propose a method for obtaining calibrated survival curves by aggregating Kaplan-Meier estimates at each covariate value. While their method can be used for post hoc calibration of Deep Cox models, their assumptions necessitate binning the risks. Finally, several modifications to the training procedure (Avati et al., 2020; Goldstein et al., 2020; Chapfuwa et al., 2023) of survival models have been proposed for improving the calibration, but these often come at the cost of worse discriminative performance (Qi et al., 2024b).

Table 1: Comparison of relevant calibration method properties. Calibration guarantees are asymptotic.

| Method | C-index preserving | Consistency Guarantees | Calibration under Misspecification |
|---|---|---|---|
| CSD | ✓ | ✗ | ✗ |
| CiPOT | ✗ | ✓ | ✗ |
| **Ours** | ✓[†] | ✓ | ✓ |

[†]C-index is preserved up to ties. Exact preservation follows by breaking ties according to the risk score.

**Isotonic Distributional Regression.** Our methodology also builds on Isotonic Distributional Regression (Henzi et al., 2021b) and related work (Jiménez and El Barmi, 2003; Davidov and Iliopoulos, 2012; Mösching and Dümbgen, 2020). Given a partial ordering over the input space, Isotonic Distributional Regression is a nonparametric approach for estimating the conditional CDF of the response variable subject to this ordering. We elaborate on related extensions in Appendix B.

# 6 EXPERIMENTS

We evaluate ISR for calibrating survival functions on synthetic data and on three real-world settings using unstructured data from The Cancer Genome Atlas. For all Cox models, we estimate the baseline hazard using Breslow's estimator. Hyperparameter and training details are provided in Appendix E. Ablations are included in Appendix F.2.

We evaluate the estimated survival distributions along three metrics: PIT calibration, quantile estimation, and the Integrated Brier Score (IBS). On the real data, we use IPCW to correct for censoring since we do not have access to the true event times. For all three metrics, lower values indicate better performance. We compare ISR with two existing baselines for post hoc calibration of survival curves: CSD (Qi et al., 2024b) and CiPOT (Qi et al., 2024a).

**PIT Calibration (AUPIT).** Recall PIT calibration requires PIT values be distributed uniformly on the interval $[0, 1]$. We quantify miscalibration as the unsigned area between the empirical cumulative distribution function (CDF) of the predicted PIT values and the uniform CDF:

$$\mathrm{AUPIT} = \int_0^1 \left| \widehat{F}_{\mathrm{PIT}}(u) - u \right| du,$$

where $\widehat{F}_{\mathrm{PIT}}$ is the IPCW-adjusted empirical CDF of the PIT values.

**Quantile Estimation (IPCW Pinball Loss).** We also evaluate the quality of the survival curves by assessing the accuracy of predicted quantiles using quantile regression loss (Koenker, 2005). The IPCW estimate reweights the observable indicators in the pinball loss as follows:

$$\mathrm{QS}(\tau) = \frac{1}{|\mathcal{I}_{\mathrm{test}}|} \sum_{i \in \mathcal{I}_{\mathrm{test}}} \left[ (1-\tau) \int_0^{\hat{q}_\tau(X_i)} \frac{\delta_i \mathbf{1}[Y_i \le t]}{G(Y_i \mid X_i)}\, dt + \tau \int_{\hat{q}_\tau(X_i)}^{t_{\max}} \frac{\mathbf{1}[Y_i > t]}{G(t \mid X_i)}\, dt \right],$$

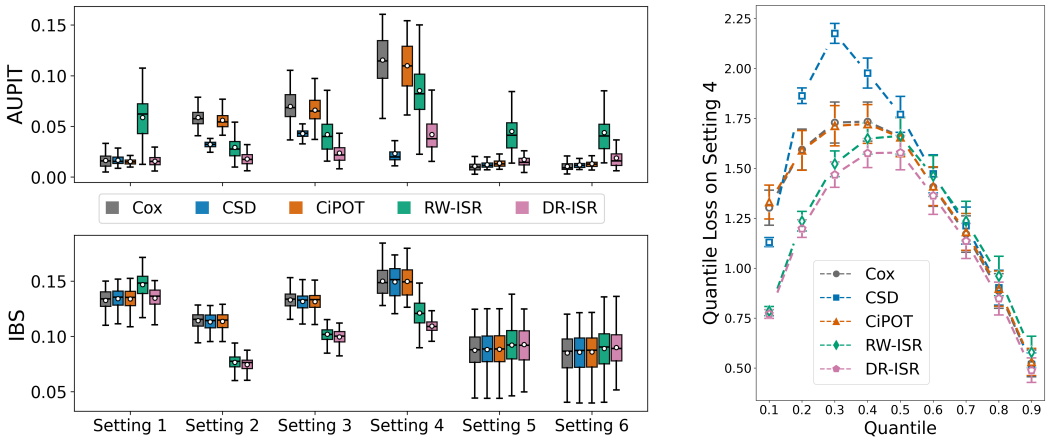
where $t_{\max}$ is the maximum time up to which predictions are made (e.g., the largest observed time across the training and calibration samples) and $\hat{q}_\tau(X_i)$ is the time at which $\widehat{S}(t|X_i)$ reaches $1-\tau$. The left and right integrals compute the weighted penalty for overestimation and underestimation respectively, analogous to the standard quantile loss (see Appendix E.1). If the survival curve does not reach the $1-\tau$ level before $t_{\max}$, then $\mathrm{QS}(\tau)$ is ill-defined; we discuss ways to handle this in Appendix E.1.

**Overall Distributional Accuracy (IBS).** While the quantile score evaluates predictive accuracy at a specific level of the event-time distribution, it does not capture the calibration of the *entire* predicted survival curve. To assess the full forecast, we use IBS (Graf et al., 1999), which measures the calibration and sharpness of the entire prediction. The Brier score measures the mean squared error of a probabilistic prediction at a single time point $t$. IBS integrates this penalty across all times up to $t_{\max}$. We show in Appendix E.2 that IBS is also equivalent to integrating the quantile score across all levels $\tau \in (0, 1)$, up to scaling factors.

$$\mathrm{IBS} = \frac{1}{|\mathcal{I}_{\mathrm{test}}|} \sum_{i \in \mathcal{I}_{\mathrm{test}}} \frac{1}{t_{\max}} \int_0^{t_{\max}} \left[ \frac{\delta_i \mathbf{1}[Y_i \le t] \hat{S}(t \mid X_i)^2}{G(Y_i \mid X_i)} + \frac{\mathbf{1}[Y_i > t](1 - \hat{S}(t \mid X_i))^2}{G(t \mid X_i)} \right] dt.$$

## 6.1 SIMULATIONS

We first evaluate our method in synthetic settings where the true event times are known, including for individuals who are censored in the data. This allows us to compute oracle values of all metrics. We use the six settings from Gui et al. (2024), where $\log T \mid X \sim \mathcal{N}(\mu(X), \sigma^2(X))$ and $C \mid X \sim P_{C|X}$. The first four simulations use a univariate covariate $X$, while Settings 5–6 use multivariate covariates




(a) (b)

Figure 2: AUPIT, IBS, and quantile loss across the six experimental settings. Boxplots in (2a) summarize the variation over seeds, where the center line indicates the median and the white dot indicates the mean. In (2b), we show the quantile loss across nine levels for Setting 4. All simulations were repeated with 100 different random seeds, and error bars in (2b) denote $\pm 2$ standard errors.

with dimension 10. In each setting, we sample a dataset of size 5,000 and split it evenly into training and calibration sets, as recommended in Candès et al. (2023). We evaluate all metrics on an independent test sample of size 5,000. Further details on the data-generating mechanisms are provided in Appendix E.4 (reproduced from Gui et al. (2024)).

Because these simulation settings are relatively low-dimensional, we use a canonical linear Cox model as the initial survival estimator. We then apply each calibration method to the fitted survival curves and compare performance across the six settings.

Figure 2a summarizes the PIT calibration and distributional accuracy results. DR-ISR achieves the lowest AUPIT in Settings 2–3 and remains competitive with the best-performing method elsewhere. On IBS, both ISR methods yield substantial improvements over the uncalibrated Cox model in Settings 2–4. In the remaining settings, where the Cox model is already moderately well calibrated, RW-ISR performs worse because it relies solely on the censoring model and therefore cannot take advantage of the good fit of the initial Cox model. The other estimators perform similarly, with the baseline methods providing little to no improvement over the uncalibrated Cox model on IBS.

We study threshold calibration more directly in Figure 2b by evaluating the quantile score across nine levels in Setting 4, where the variation in IBS is the greatest. Both RW-ISR and DR-ISR achieve lower quantile loss compared to other methods across all levels, with DR-ISR performing better at medium quantiles. We include results for the quantile score on remaining settings in Appendix F.1.

### 6.2 Real-data results

We now evaluate our calibration procedures on models trained using real-world datasets to demonstrate that they typically continue to provide an empirical advantage over existing post hoc calibration methods *even when Assumptions A1 and A2 are (likely) violated.*

In recent years, an increasingly common approach is to combine the features extracted from powerful foundation models with Deep Cox networks for downstream survival analysis tasks (Bioptimus, 2025; Hemker et al., 2024; Kang et al., 2025; Lu et al., 2021; Zhang et al., 2025; Vaidya et al., 2025; Ding et al., 2025; Chen et al., 2024; Lu et al., 2024b; 2023; Wang et al., 2024b; Zhao et al., 2021; Zhu et al., 2017). We explore exactly this common paradigm in our real-data experiments.

We use The Cancer Genome Atlas dataset for our experiments which consists of about 10,000 cancer patients, of which we use 60% for model fitting, 25% for calibration and 15% for testing. We aim to predict the progression-free survival (PFS), the length of time during and after treatment that a patient lives with the disease without it worsening.

Table 2: IPCW estimates of AUPIT, IBS, and quantile loss at the $20\%, 50\%$ and $80\%$ levels over the test set for the text reports, RNA-seq, and H&E datasets. Values are mean over 100 seeds with the p-value in brackets (calculated by taking the maximum across paired one sided t-test p-values comparing pairwise with all other methods). Entries are bolded if $p < 0.05$.

| | Method | AUPIT | IBS | 20% Quantile | 50% Quantile | 80% Quantile |
|---|---|---|---|---|---|---|
| Text Reports | Deep Cox | 0.0970 (1.00) | 0.125 (1.00) | 371 (1.00) | 700 (1.00) | 596 (1.00) |
| | CSD | 0.1256 (1.00) | 0.125 (1.00) | 370 (1.00) | 704 (1.00) | 598 (1.00) |
| | CiPOT | 0.0948 (1.00) | 0.125 (1.00) | 363 (1.00) | 696 (0.86) | **590 (0.001)** |
| | RW-ISR (ours) | 0.1394 (1.00) | 0.123 (0.68) | 357 (0.72) | 700 (1.00) | 621 (1.00) |
| | DR-ISR (ours) | **0.0819 (<0.001)** | 0.122 (0.32) | 357 (0.28) | 692 (0.14) | 613 (1.00) |
| RNA-seq | Deep Cox | 0.0546 (1.00) | 0.102 (1.00) | 375 (0.99) | 740 (1.00) | 603 (1.00) |
| | CSD | 0.0826 (1.00) | 0.102 (1.00) | 376 (1.00) | 735 (1.00) | 608 (1.00) |
| | CiPOT | 0.0544 (1.00) | 0.102 (1.00) | 375 (0.99) | 732 (1.00) | **601 (0.004)** |
| | RW-ISR (ours) | 0.0923 (1.00) | **0.092 (<0.001)** | 375 (0.99) | 699 (1.00) | 619 (1.00) |
| | DR-ISR (ours) | **0.0517 (<0.001)** | 0.098 (1.00) | **373 (0.01)** | **682 (<0.001)** | 625 (1.00) |
| H&E Slides | Deep Cox | 0.1007 (1.00) | 0.113 (1.00) | 441 (1.00) | 781 (1.00) | 661 (1.00) |
| | CSD | 0.0787 (1.00) | 0.110 (1.00) | 429 (1.00) | 775 (1.00) | **628 (0.01)** |
| | CiPOT | 0.0800 (1.00) | 0.111 (1.00) | 396 (1.00) | 771 (1.00) | 653 (1.00) |
| | RW-ISR (ours) | 0.0889 (1.00) | **0.096 (<0.001)** | **371 (0.004)** | **729 (<0.001)** | 651 (0.99) |
| | DR-ISR (ours) | **0.0568 (<0.001)** | 0.110 (1.00) | 380 (0.99) | 761 (1.00) | 752 (1.00) |

For our first experiment, we use the text-based pathology reports from Kefeli and Tatonetti (2024). To train our model, we first extract the final-layer embeddings from the BioClinical BERT foundation model (Alsentzer et al., 2019) as covariate features. We then train a three-layer MLP head as our Deep Cox model to predict risk scores for the PFS (as compiled by Liu et al. (2018)) by maximizing the Cox partial likelihood (Katzman et al., 2018). A separate three-layer MLP head is trained to obtain the censoring model. In the second experiment, we use bulk RNA-seq data from Vivian et al. (2017) and the BulkFormer foundation model (Kang et al., 2025) to extract features, but otherwise keep the exact same training procedure. For our third experiment, we use H&E slide data (Weinstein et al., 2013) and the image-only SPARC model (Ayed et al., 2026) to extract features, again keeping the same procedure. All C-indices for the PFS and censoring models are reported in Appendix E.5.

We report performance using the IPCW-estimated metrics in Table 2. To avoid biasing evaluation towards our methods, we obtain IPCW weights from independent Random Survival Forests (Ishwaran and Lu, 2019) trained on top of the features extracted by the various foundation models. These models achieve C-indices comparable to the censoring Deep Cox models. DR-ISR achieves the best performance on AUPIT, even improving over the baselines, which directly target PIT calibration. We believe this is because DR-ISR explicitly models the dependence of the censoring distribution on the covariates, whereas CSD and CiPOT do not directly account for this structure. On IBS, we find that both ISR variants perform well and are statistically significantly better than any of the baselines for text reports and RNA-seq. Decomposing IBS by quantile level shows gains up to the 50% quantile.

## 7 Discussion

In this work, we propose a new algorithm for calibrating survival functions under right censoring using Isotonic Survival Regression (ISR). Our method extends principled pointwise survival probability estimates to valid, monotone survival functions while preserving discriminative performance. Theoretically, we establish calibration and consistency guarantees for the proposed procedure. Across experiments, we find that DR-ISR gives the best overall calibration performance.

**Limitations.** As with many survival analysis methods, our algorithm relies on assumptions that may be difficult to verify in practice. Nonetheless, our real-data experiments suggest that the proposed method is moderately robust to violations of these assumptions. Our approach relies on a partial ordering of individuals by way of a univariate risk score. While simplistic, we believe this modeling choice is useful in many applications, and it is supported by the practical success of Cox models.

Our theoretical results are stated for a fixed risk score. In practice, we use risk scores learned by the Cox model, and our results do not immediately extend to this setting. Future work could establish consistency or calibration guarantees when the learned risk mapping converges to a limiting function.

### AI USE STATEMENT

In this work, we used generative AI tools for polishing the writing and improving the clarity of the proofs. All authors have reviewed the AI-assisted edits in the text and for any theoretical claims. We have not used generative AI tools for writing code, generating synthetic data, or drafting sections of the paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES


Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in medicine*, 24(24):3927–3944, 2005.

Mohammad Ashhad, Robert Hoehndorf, and Ricardo Henao. Graft: Decoupling ranking and calibration for survival analysis. *arXiv preprint arXiv:2602.07884*, 2026.

Anand Avati, Tony Duan, Sharon Zhou, Kenneth Jung, Nigam H. Shah, and Andrew Y. Ng. Countdown regression: Sharp and calibrated survival predictions. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 145–155. PMLR, 22–25 Jul 2020.

Aziz Ayed, Gabrielle Cohn, Noé Bertramo, Genevieve Boland, Justin Gainor, Ömer H. Yilmaz, and Regina Barzilay. Sparc: A mechanism-aware spatial representation from routine histology predicts cancer survival and therapy response. *medRxiv*, 2026. doi: 10.64898/2026.05.04.26352410.

R.E. Barlow. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. J. Wiley, 1972. ISBN 9780471049708.

Bioptimus. H-optimus-1, 2025. URL `https://huggingface.co/bioptimus/H-optimus-1`.

Martin Bladt, Alexander Henzi, Bram van den Heuvel, and Johanna Ziegel. Survival isotonic distributional regression. *arXiv preprint arXiv:2608.02914*, 2026.

Norman E. Breslow. Discussion of "Regression models and life-tables" by D.R. Cox. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):216–217, 1972.

Norman E. Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review*, 43, 1975.

Emmanuel Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 85, 2023.

Paidamoyo Chapfuwa, Chenyang Tao, Chunyuan Li, Irfan Khan, Karen J. Chandross, Michael J. Pencina, Lawrence Carin, and Ricardo Henao. Calibration and uncertainty in neural time-to-event modeling. *IEEE Transactions on Neural Networks and Learning Systems*, 34(4):1666–1680, 2023.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

David Collett. *Modelling Survival Data in Medical Research*. Chapman and Hall, 2014.

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

Hen Davidov, Shai Feldman, Gil Shamai, Ron Kimmel, and Yaniv Romano. Conformalized survival analysis for general right-censored data. In *International Conference on Learning Representations*, 2025.

Ori Davidov and George Iliopoulos. Estimating a distribution function subject to a stochastic order restriction: a comparative study. *Journal of Nonparametric Statistics*, 24(4):923–933, 2012.

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2):12–22, 1983.

Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, pages 1–13, 2025.

Richard L. Dykstra and Tim Robertson. An Algorithm for Isotonic Regression for Two or More Independent Variables. *The Annals of Statistics*, 10(3):708 – 716, 1982. doi: 10.1214/aos/1176345866.

Rebecca Farina, Eric J. Tchetgen Tchetgen, and Arun Kumar Kuchibhotla. Doubly robust and efficient calibration of prediction sets for right-censored time-to-event outcomes. *CoRR*, abs/2501.04615, 2026.

Robert Flynn. Survival analysis. *Journal of Clinical Nursing*, 21, 2012.

Tilmann Gneiting and Roopesh Ranjan. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29:411–422, 2011.

Mark Goldstein, Xintian Han, Aahlad Puli, Adler Perotte, and Rajesh Ranganath. X-cal: Explicit calibration for survival analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18296–18307. Curran Associates, Inc., 2020.

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.

P. Groeneboom and G. Jongbloed. *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2014. ISBN 9781316194126.

Yu Gui, Rohan Hore, Zhimei Ren, and Rina Foygel Barber. Conformalized survival analysis with adaptive cut-offs. *Biometrika*, 111, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 2017.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21, 2024.

Frank E. Harrell Jr., Kerry L. Le, 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.

Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. HEALNet: Multimodal fusion for heterogeneous biomedical data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Alexander Henzi, Gian-Reto Kleger, and Johanna F. Ziegel. Distributional (single) index models. *Journal of the American Statistical Association*, 118, 2021a.

Alexander Henzi, Johanna F. Ziegel, and Tilmann Gneiting. Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83, 2021b.

Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000.

Chris Holmes and Ariane Marandon. Two-sided conformalized survival analysis. *CoRR*, abs/2410.24136, 2024.

Hemant Ishwaran and Min Lu. *Random Survival Forests*, pages 1–13. John Wiley and Sons, Ltd, 2019. ISBN 9781118445112. doi: https://doi.org/10.1002/9781118445112.stat08188.

Javier Rojo Jiménez and Hammou El Barmi. Estimation of distribution functions under second order stochastic dominance. *Statistica Sinica*, pages 903–926, 2003.

Alistair Johnson, Tom Pollard, and Roger Mark. MIMIC-III Clinical Database. *PhysioNet*, September 2016. Version 1.4.

John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, 2002.

Boming Kang, Rui Fan, Meizheng Yi, Chunmei Cui, and Qinghua Cui. A large-scale foundation model for bulk transcriptomes. *bioRxiv*, 2025. doi: 10.1101/2025.06.11.659222.

Edward L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 1958.

Jared L. Katzman, Uri Shaham, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18, 2018.

Jenna Kefeli and Nicholas Tatonetti. Tcga-reports: A machine-readable pathology report resource for benchmarking text-based ai models. *Patterns*, 5(3):100933, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR (Poster)*, 2015.

John P. Klein and Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 1997.

Roger Koenker. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20, 2019.

Stephen W. Lagakos. General right censoring and its impact on the analysis of survival data. *Biometrics*, 35(1), 1979.

Nicholas R. Latimer. Survival analysis for economic evaluations alongside clinical trials–extrapolation with patient-level data: Inconsistencies, limitations, and a practical guide. *Medical Decision Making*, 33(6):743–754, 2013.

K. M. Leung, Robert M. Elashoff, and Abdelmonem A. Afifi. Censoring issues in survival analysis. *Annual Review of Public Health*, 1997.

Jianfang Liu, Tara Lichtenberg, Katherine A. Hoadley, Laila M. Poisson, Alexander J. Lazar, and et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416.e11, 2018.

Yu Liu, Weiyao Tao, Tong Xia, Simon Knight, and Tingting Zhu. Survunc: A meta-model based uncertainty quantification framework for survival analysis. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 1903–1914, 2025.

Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.

Ming Y. Lu, Bowen Chen, Andrew Zhang, Drew F. K. Williamson, Richard J. Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19764–19775, June 2023.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30, 2024a.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024b.

Huiyan Luo, Jin Lu, Yuxian Bai, and et al. Effect of camrelizumab vs placebo added to chemotherapy on survival and progression-free survival in patients with advanced or metastatic esophageal squamous cell carcinoma: The escort-1st randomized clinical trial. *JAMA*, 326(10), 2021.

Carlotta Masciocchi, Benedetta Gottardelli, Mariachiara Savino, Luca Boldrini, Antonella Martino, Ciro Mazzarella, Mariangela Massaccesi, Vincenzo Valentini, and Andrea Damiani. Federated cox proportional hazards model with multicentric privacy-preserving lasso feature selection for survival analysis from the perspective of personalized medicine. In *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, 2022.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Alexandre Mösching and Lutz Dümbgen. Monotone least squares and isotonic quantiles. *Electronic Journal of Statistics*, 14(1):24 – 49, 2020.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

Anne-Michelle Noone, Kathleen A. Cronin, Sean F. Altekruse, Nadia Howlader, Denise R. Lewis, Valentina I. Petkov, and Lynne Penberthy. Cancer incidence and survival trends by subtype using data from the surveillance epidemiology and end results program, 1992–2013. *Cancer Epidemiology, Biomarkers & Prevention*, 26, 2017.

Dimitris Papathanasiou, Konstantinos Demertzis, and Nikos Tziritas. Machine failure prediction using survival analysis. *Future Internet*, 15(5), 2023.

Shi-Ang Qi, Yakun Yu, and Russell Greiner. Toward conditional distribution calibration in survival prediction. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

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, 2024b.

Jing Qin, Jin Piao, Jing Ning, and Yu Shen. Conformal predictive intervals in survival analysis: A resampling approach. *Biometrics*, 81, 2025.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4, 2021.

James M. Robins and Andrea Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In Nicholas P. Jewell, Klaus Dietz, and Vern T. Farewell, editors, *AIDS Epidemiology*. Birkhäuser, 1992.

Patrick Schober and Thomas R. Vetter. Survival analysis and interpretation of time-to-event data: The tortoise and the hare. *Anesthesia & Analgesia*, 127, 2018.

Matteo Sesia and Vladimir Svetnik. Doubly robust conformalized survival analysis with right-censored data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, 2025a.

Matteo Sesia and Vladimir Svetnik. Conformal survival bands for risk screening under right-censoring. In *Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 266, 2025b.

Wenwen Si and Hongxiang Qiu. Training-set conditionally valid prediction sets with right-censored data. *CoRR*, abs/2505.02213, 2025.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A. Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Anastasios A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, 2006.

Anurag Vaidya, Andrew Zhang, Guillaume Jaume, Andrew H Song, Tong Ding, Sophia J Wagner, Ming Y Lu, Paul Doucet, Harry Robertson, Cristina Almagro-Perez, et al. Molecular-driven foundation model for oncologic pathology. *arXiv preprint arXiv:2501.16652*, 2025.

Aad W Van Der Vaart and Jon A Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer, 1996.

Mike Van Ness, Tomas Bosschieter, Natasha Din, Andrew Ambrosy, Alexander Sandhu, and Madeleine Udell. Interpretable survival analysis for heart failure risk prediction. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225, 2023.

John Vivian, Arjun Arkal Rao, Frank Austin Nothaft, Christopher Ketchum, Joel Armstrong, and et al. Toil enables reproducible, open source, big biomedical data analyses. *Nature Biotechnology*, 35(4):314–316, 2017. doi: 10.1038/nbt.3772.

Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, and et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634, 2024a.

Zhaonian Wang, Chundan Zheng, Xu Han, Wufan Chen, and Lijun Lu. An innovative and efficient diagnostic prediction flow for head and neck cancer: A deep learning approach for multi-modal survival analysis prediction based on text and multi-center pet/ct images. *Diagnostics*, 14(4), 2024b. ISSN 2075-4418.

John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, and et al. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, October 2013. ISSN 1061-4036. doi: 10.1038/ng.2764.

Menghan Yi, Ze Xiao, Huixia Judy Wang, and Yanlin Tang. Survival conformal prediction under random censoring. *Stat*, 14, 2025.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 694–699, New York, NY, USA, 2002. Association for Computing Machinery.

Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating data processing and benchmarking of ai models for pathology. *arXiv preprint arXiv:2502.06750*, 2025.

Yun Zhao, Qinghang Hong, Xinlu Zhang, Yu Deng, Yuqing Wang, and Linda Petzold. Bertsurv: Bert-based survival models for predicting outcomes of trauma patients. *arXiv preprint arXiv:2103.10928*, 2021.

Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6855–6863, 2017. doi: 10.1109/CVPR.2017.725.

# Supplementary Material for Isotonic Survival Regression: Calibrated Survival Predictions from Deep Cox Models

## A PROOFS

We first establish a key result used in our arguments: under the risk-ordering assumption, the survival function can be expressed solely in terms of the risk score produced by the uncalibrated model.

**Lemma 1** (Risk Score Sufficiency)**.** *Under Assumption A2, the event time depends on the covariates only through the risk score $r(X)$. Equivalently,*

$$\mathbb{P}(T > t \mid X) = \mathbb{P}(T > t \mid r(X)), \quad \textit{for all } t \geq 0.$$

*Proof.* Let $x, x' \in \mathcal{X}$ such that $r(x) = r(x')$. Then, we have $r(x) \geq r(x')$ and $r(x') \geq r(x)$. Applying Assumption A2 twice gives $S(t \mid x) = S(t \mid x')$. By the tower property,

$$\begin{aligned}
\mathbb{P}(T > t \mid r(X)) &= \mathbb{E}[\mathbf{1}[T > t] \mid r(X)] \\
&= \mathbb{E}[\mathbb{E}[\mathbf{1}[T > t] \mid X] \mid r(X)] \\
&= \mathbb{E}[S(t \mid X) \mid r(X)] \\
&= S(t \mid X).
\end{aligned}$$

□

Lemma 1 states that the survival function can be expressed in terms of the risk. Thus, for ease of exposition, we write $S_{0,r}(t \mid r) = \mathbb{P}(T > t \mid r(X) = r)$. We also introduce notation for the nuisance parameters $\eta = (S, G)$, with true value $\eta_0 = (S_0, G_0)$ and estimated value $\widehat{\eta} = (\widehat{S}, \widehat{G})$. Our proofs will make use of the following additional regularity assumptions as specified in the corresponding theorem statements.

- S1. We assume that there exists constants $0 < c_1(t) < \infty$ and $0 < c_2(t) < \infty$ depending on $t$ such that $S_0(t \mid x) \wedge \widehat{S}(t \mid x) > 1/c_1(t)$ and $G_0(t \mid x) \wedge \widehat{G}(t \mid x) > 1/c_2(t)$ for all $x$.
- S2. The covariates $X \overset{\text{i.i.d.}}{\sim} P_X$ induce a distribution for the risks $r(X) \overset{\text{i.i.d.}}{\sim} P$. We assume $P$ has bounded support $[a, b]$ and a Lebesgue density $p$ which is strictly positive over the support.
- S3. We assume the survival function in terms of risk $S_{0,r}(t \mid r)$ is continuous in $r$ at all times $t$.

Assumption S1 is a pointwise-in-time positivity condition. At each fixed time $t$, the survival and censoring probabilities are uniformly bounded away from zero over $x$, but the lower bound may vanish as $t \to \infty$.

### A.1 PROOF OF THEOREM 1

We first discuss the augmented IPCW (AIPCW) estimating equation for the conditional survival probability using standard semiparametric analysis (Robins and Rotnitzky, 1992; Tsiatis, 2006). In particular, we will demonstrate that the AIPCW estimator has two equivalent representations using either the censoring or event martingale difference. Next, we will show that the two bias bounds hold separately.

$$B(t, x; \widehat{\eta}) = |\mathbb{E}[\tilde{S}^{\mathrm{DR}}(t \mid x; \widehat{\eta}) \mid X = x] - S_0(t \mid x)| \leq B_i(t, x; \widehat{\eta}).$$

We use the event martingale representation to get the bound with $B_1$ and the censoring martingale representation to get the bound with $B_2$. The full result then follows by taking the minimum bound between $B_1$ and $B_2$.

For simplicity and notational convenience, we will assume all of the survival and censoring functions ($S_0$, $G_0$, $\widehat{S}$, and $\widehat{G}$) are continuous; the general case follows by using the appropriate left-limits. We will first prove two useful identities regarding the survival and censoring probabilities.

**Lemma 2.** *For any time $t > 0$ and continuous survival function $S$,*

$$d\left(\frac{1}{S(t \mid X)}\right) = \frac{1}{S(t \mid X)}\, d\Lambda_{T|X}(t \mid X),$$

*where $\Lambda_{T|X}(t \mid X) = -\log S(t \mid X)$ is the cumulative hazard function. The same result holds when replacing $S$ with the censoring model equivalent $G$.*

*Proof.* We differentiate $\frac{1}{S(u|X)} = \exp\{\Lambda_{T|X}(u \mid X)\}$ with respect to $\Lambda_{T|X}$. Then,

$$d\left(\frac{1}{S(u \mid X)}\right) = \exp\{\Lambda_{T|X}(u \mid X)\} d\Lambda_{T|X}(u \mid X) = \frac{1}{S(u \mid X)} d\Lambda_{T|X}(u \mid X).$$

The proof is identical for any censoring model $G$ using the cumulative hazard function $\Lambda_{C|X}$. □

**Lemma 3.** *For any time $t > 0$ and pair of continuous survival functions $S$ and $S'$,*

$$S(t \mid X) - S'(t \mid X) = S'(t \mid X) \int_0^t \frac{S(u \mid X)}{S'(u \mid X)} \left(d\Lambda'_{T|X}(u \mid X) - d\Lambda_{T|X}(u \mid X)\right),$$

*where $\Lambda_{T|X}(u \mid X) = -\log S(u \mid X)$ and $\Lambda'_{T|X}(u \mid X) = -\log S'(u \mid X)$ are the respective cumulative hazard functions. The same result holds when replacing $S$ and $S'$ with their censoring model equivalents $G$ and $G'$, respectively.*

*Proof.* We express the ratio of survival probabilities in terms of the cumulative hazard.

$$\frac{S(u \mid X)}{S'(u \mid X)} = \exp\left\{\Lambda'_{T|X}(u \mid X) - \Lambda_{T|X}(u \mid X)\right\}.$$

Taking the derivative, we have

$$d\left(\frac{S(u \mid X)}{S'(u \mid X)}\right) = \frac{S(u \mid X)}{S'(u \mid X)} \left(d\Lambda'_{T|X}(u \mid X) - d\Lambda_{T|X}(u \mid X)\right).$$

We integrate and use $S(0 \mid X) = S'(0 \mid X) = 1$ to get

$$\frac{S(t \mid X)}{S'(t \mid X)} - 1 = \int_0^t \frac{S(u \mid X)}{S'(u \mid X)} \left(d\Lambda'_{T|X}(u \mid X) - d\Lambda_{T|X}(u \mid X)\right).$$

Multiplying both sides by $S'(t \mid X)$ gives the desired result. The proof is identical for any pair of censoring models $G$ and $G'$ using the cumulative hazard functions $\Lambda_{C|X}$ and $\Lambda'_{C|X}$. □

### A.1.1 AIPCW ESTIMATORS OF THE SURVIVAL PROBABILITY

We now analyze AIPCW estimator for the conditional survival probability. For a fixed $t$, the target can be written as the conditional mean of the complete-data outcome $\mathbf{1}[T > t]$. Let

$$\beta(x) = \mathbb{P}(T > t \mid X = x) = \mathbb{E}[\mathbf{1}[T > t] \mid X = x].$$

This perspective allows us to use standard semiparametric results for AIPCW estimators, including their efficiency and robustness properties. At a high level, right censoring is a monotone coarsening mechanism, and the AIPCW estimator can be expressed using the martingale difference process associated with the censoring mechanism. We refer the reader to Robins and Rotnitzky (1992) and Tsiatis (2006) for additional details on semiparametric theory and AIPCW estimation.

Define the counting process $dN_C(u) = \mathbf{1}[Y \in du, \delta = 0]$ and censoring martingale increment

$$dM_{C|X}(u \mid X) = dN_C(u) - \mathbf{1}[Y \geq u]\, d\Lambda_{C|X}(u \mid X),$$

where $\Lambda_{C|X}(u \mid X) = -\log G(u \mid X)$ is the censoring analog of the cumulative hazard function. Following the construction in Section 9.3 of Tsiatis (2006), the AIPCW estimation equation for $\beta$ is

$$\mathbb{E}\Bigg[\underbrace{\frac{\delta(\mathbf{1}[T > t] - \beta(X))}{G(Y \mid X)}}_{\text{IPCW estimate}} + \underbrace{\int_0^\infty \frac{dM_{C|X}(u \mid X)}{G(u \mid X)} h(u, X)}_{\text{augmentation}} \;\Bigg|\; X = x\Bigg] = 0. \tag{7}$$

The first term is the IPCW estimate of $\beta(X)$, while the second augmentation term lies in the censoring augmentation space $\mathcal{A}$ for any square integrable function $h \in L^2$. The choice of $h$ determines the efficiency of the resulting estimator. By Theorem 10.4 of Tsiatis (2006), the efficient choice is such that the augmentation is the projection of the IPCW estimating function onto $\mathcal{A}$, which for survival probability estimation is

$$h(u, X) = \mathbb{E}[\mathbf{1}[T > t] - \beta(X) \mid X,\ T \geq u].$$

We simplify $h(u, X)$ in terms of the conditional survival function. Since $\beta(X) = S(t \mid X)$,

$$\begin{aligned}
h(u, X) &= \mathbb{E}\left[\mathbf{1}(T > t) - \beta(X) \mid X,\ T \geq u\right] \\
&= \mathbb{P}(T > t \mid X,\ T \geq u) - \beta(X) \\
&= \frac{\mathbb{P}(T > t,\ T \geq u \mid X)}{\mathbb{P}(T \geq u \mid X)} - \beta(X) \\
&= \frac{\mathbb{P}(T > \max\{u, t\} \mid X)}{\mathbb{P}(T \geq u \mid X)} - \beta(X) \\
&= \frac{S(\max\{u, t\} \mid X)}{S(u \mid X)} - \beta(X).
\end{aligned}$$

Substituting into Equation 7, we obtain the observed-data estimating equation

$$\beta(x) = \mathbb{E}\left[\frac{\delta \mathbf{1}[Y > t]}{G(Y \mid X)} + \int_0^\infty \frac{S(\max\{u, t\} \mid X)}{S(u \mid X) G(u \mid X)} dM_{C|X}(u \mid X) \;\middle|\; X = x\right] \Big/ W(x),$$

where

$$W(x) = \mathbb{E}\left[\frac{\delta}{G(Y \mid X)} + \int_0^\infty \frac{dM_{C|X}(u \mid X)}{G(u \mid X)} \;\middle|\; X = x\right].$$

We now show that $W(x) = 1$. Since $dN_C(u)$ increments at $Y$ if and only if $\delta = 0$,

$$\begin{aligned}
W(x) &= \mathbb{E}\left[\frac{\delta}{G(Y \mid X)} + \int_0^\infty \frac{1}{G(u \mid X)}\, dN_C(u) - \int_0^\infty \frac{\mathbf{1}[Y \geq u]}{G(u \mid X)}\, d\Lambda_{C|X}(u \mid X) \;\middle|\; X = x\right] \\
&= \mathbb{E}\left[\frac{\delta}{G(Y \mid X)} + \frac{1 - \delta}{G(Y \mid X)} - \int_0^Y \frac{d\Lambda_{C|X}(u \mid X)}{G(u \mid X)} \;\middle|\; X = x\right] \\
&= \mathbb{E}\left[\frac{\delta}{G(Y \mid X)} + \frac{1 - \delta}{G(Y \mid X)} - \int_0^Y d\left(\frac{1}{G(u \mid X)}\right) \;\middle|\; X = x\right] \qquad \text{(Lemma 2)} \\
&= \mathbb{E}\left[\frac{1}{G(Y \mid X)} - \left\{\frac{1}{G(Y \mid X)} - 1\right\} \;\middle|\; X = x\right] = 1.
\end{aligned}$$

Thus, we have the AIPCW estimating equation

$$\beta(x) = \mathbb{E}\left[\frac{\delta \mathbf{1}[Y > t]}{G(Y \mid X)} + \int_0^\infty \frac{S(\max\{u, t\} \mid X)}{S(u \mid X) G(u \mid X)} dM_{C|X}(u \mid X) \;\middle|\; X = x\right].$$

### A.1.2 Censoring and event martingale representations for AIPCW estimators

The construction in Tsiatis for the efficient AIPCW estimator uses the censoring martingale difference $dM_{C|X}$. However, for our doubly robust pseudo-outcomes in Equation 2, we use the event martingale difference $dM_{T|X}$. We will now show that these representations are equivalent.

**Lemma 4.** *Suppose Assumption A1 holds. For any time $t > 0$ and continuous functions $S$ and $G$, the censoring and event martingale representations of the AIPCW estimator are equivalent, i.e.*

$$\frac{\delta \mathbf{1}[Y > t]}{G(Y \mid X)} + \int_0^\infty \frac{S(\max\{u,t\} \mid X)}{S(u \mid X)G(u \mid X)}\, dM_{C|X}(u \mid X) = S(t \mid X) - \int_0^t \frac{S(t \mid X)}{S(u \mid X)G(u \mid X)}\, dM_{T|X}(u \mid X).$$

*Proof.* Define $H(u) = \frac{S(t|X)}{S(u|X)G(u|X)}$ for $0 \le u \le t$. Then,

$$dH(u) = H(u)\left(d\Lambda_{T|X}(u \mid X) + d\Lambda_{C|X}(u \mid X)\right).$$

Define the at-risk indicator $R(u) = \mathbf{1}[Y \ge u]$, with $dR(u) = -dN_T(u) - dN_C(u)$. Then,

$$\begin{aligned} d[H(u)R(u)] &= H(u)R(u)\left(d\Lambda_{T|X}(u \mid X) + d\Lambda_{C|X}(u \mid X)\right) - H(u)\left(dN_T(u) + dN_C(u)\right) \\ &= H(u)\left(-dM_{T|X}(u \mid X) - dM_{C|X}(u \mid X)\right). \end{aligned}$$

Integrating and using $S(0 \mid X) = G(0 \mid X) = R(0) = 1$ gives

$$-\int_0^t H(u) dM_{T|X}(u \mid X) - \int_0^t H(u) dM_{C|X}(u \mid X) = H(t)R(t) - H(0)R(0) = \frac{\mathbf{1}[Y > t]}{G(t \mid X)} - S(t \mid X).$$

Rearranging the result and expressing in terms of $S$ and $G$ gives,

$$\frac{\mathbf{1}[Y > t]}{G(t \mid X)} - S(t \mid X) + \int_0^t \frac{S(t \mid X)}{S(u \mid X)G(u \mid X)}\, dM_{C|X}(u \mid X) = -\int_0^t \frac{S(t \mid X)}{S(u \mid X)G(u \mid X)} dM_{T|X}(u \mid X). \tag{8}$$

We now analyze the component of the censoring martingale integral from $t$ to $\infty$. Using Lemma 2,

$$\begin{aligned} d\left(\frac{R(u)}{G(u \mid X)}\right) &= -\frac{dN_T(u) + dN_C(u)}{G(u \mid X)} + \frac{R(u)}{G(u \mid X)} d\Lambda_{C|X}(u \mid X). \\ &= -\frac{dM_{C|X}(u \mid X)}{G(u \mid X)} - \frac{dN_T(u)}{G(u \mid X)}. \end{aligned}$$

By Assumption S1, we have $G(Y \mid X) > 0$, so

$$-\int_t^\infty \frac{dM_{C|X}(u \mid X)}{G(u \mid X)} - \int_t^\infty \frac{dN_T(u)}{G(u \mid X)} = \frac{R(u)}{G(u \mid X)}\Big|_t^\infty = -\frac{\mathbf{1}[Y > t]}{G(t \mid X)}.$$

Since $dN_T(u)$ increments at $u = Y$ only when $Y > t$ and $\delta = 1$,

$$\int_t^\infty \frac{dN_T(u)}{G(u \mid X)} = \frac{\delta \mathbf{1}[Y > t]}{G(Y \mid X)}.$$

Thus, we have that

$$\frac{\mathbf{1}[Y > t]}{G(t \mid X)} = \frac{\delta \mathbf{1}[Y > t]}{G(Y \mid X)} + \int_t^\infty \frac{dM_{C|X}(u \mid X)}{G(u \mid X)}. \tag{9}$$

Since,

$$\frac{S(\max\{u,t\} \mid X)}{S(u \mid X)G(u \mid X)} = \begin{cases} \frac{S(t|X)}{S(u|X)G(u|X)} & u < t \\ \frac{1}{G(u|X)} & u \ge t, \end{cases}$$

substituting Equation 9 into Equation 8 gives the desired equivalence. □

The equivalence between the censoring and event martingale representation holds for the plug-in AIPCW estimator using $\widehat{S}$ and $\widehat{G}$ as well under our continuity simplification. We will use the two different expressions to derive separate bounds for the bias.

### A.1.3 EVENT MARTINGALE BOUND

Define the target functional

$$\psi(t, x; \eta) = \mathbb{E}\big[\tilde{S}^{\mathrm{DR}}(t \mid x; \widehat{\eta}) \;\big|\; X = x\big].$$

The bias of the pseudo-outcome is

$$B(t, x; \widehat{\eta}) = |\psi(t, x; \widehat{\eta}) - S_0(t \mid x)|.$$

Using the event martingale representation,

$$B(t, x; \widehat{\eta}) = \left| \mathbb{E}\left[ \widehat{S}(t \mid X) - S_0(t \mid X) - \int_0^t \frac{\widehat{S}(t \mid X)}{\widehat{S}(u \mid X)\widehat{G}(u \mid X)} d\widehat{M}_{T|X}(u \mid X) \;\Big|\; X = x \right] \right|.$$

Note that the only randomness in this expression is in the event process difference $d\widehat{M}_{T|X}$. Let $\Lambda_{0,T|X}(t \mid X) = -\log S_0(t \mid X)$ and $\widehat{\Lambda}_{T|X}(t \mid X) = -\log \widehat{S}(t \mid X)$ be the true and estimated cumulative hazard functions, respectively. Under assumption A1,

$$\begin{aligned}
\mathbb{E}\left[dN_T(u) \mid X = x\right] &= \mathbb{P}(Y \in du, \delta = 1 \mid X = x) \\
&= \mathbb{P}(T \in du,\, C \geq u \mid X = x) \\
&= \mathbb{P}(C \geq u \mid X = x)\, \mathbb{P}(T \in du \mid X = x) \\
&= S_0(u \mid x)\, G_0(u \mid x)\, d\Lambda_{0,T|X}(u \mid x). \\
\mathbb{E}\left[\mathbf{1}[Y \geq u] \mid X = x\right] &= \mathbb{P}(Y \geq u \mid X = x) \\
&= \mathbb{P}(T \geq u, C \geq u \mid X = x) \\
&= S_0(u \mid x)\, G_0(u \mid x).
\end{aligned}$$

Thus, we have that

$$\mathbb{E}\big[d\widehat{M}_{T|X}(u \mid X) \mid X = x\big] = S_0(u \mid x) G_0(u \mid x) \left( d\Lambda_{0,T|X}(u \mid x) - d\widehat{\Lambda}_{T|X}(u \mid x) \right).$$

Substituting this result into the bias expression gives,

$$\begin{aligned}
B(t, x; \widehat{\eta}) = \Bigg| & \widehat{S}(t \mid x) - S_0(t \mid x) \\
& - \int_0^t \frac{\widehat{S}(t \mid x) S_0(u \mid x) G_0(u \mid x)}{\widehat{S}(u \mid x)\widehat{G}(u \mid x)} \left( d\Lambda_{0,T|X}(u \mid x) - d\widehat{\Lambda}_{T|X}(u \mid x) \right) \Bigg|.
\end{aligned}$$

Using Lemma 3, we have

$$B(t, x; \widehat{\eta}) = \left| \int_0^t \frac{\widehat{S}(t \mid x) S_0(u \mid x)}{\widehat{S}(u \mid x)} \left( 1 - \frac{G_0(u \mid x)}{\widehat{G}(u \mid x)} \right) \left( d\Lambda_{0,T|X}(u \mid x) - d\widehat{\Lambda}_{T|X}(u \mid x) \right) \right|.$$

Under Assumption S1, there exists some constant $c_1(t)$ dependent on $t$ such that $\widehat{S}(t \mid x) > 1/c_1(t)$. Then, we have the bound

$$0 \leq \frac{\widehat{S}(t \mid x) S_0(u \mid x) G_0(u \mid x)}{\widehat{S}(u \mid x)} \leq c_1(t), \quad \text{for all } u \in [0, t].$$

Therefore, we have that $B(t, x; \widehat{\eta}) \leq B_1(t, x; \widehat{\eta})$, where

$$B_1(t, x; \widehat{\eta}) = c_1(t) \int_0^t \left| \frac{1}{G_0(u \mid x)} - \frac{1}{\widehat{G}(u \mid x)} \right| \left| d\Lambda_{0,T|X}(u \mid x) - d\widehat{\Lambda}_{T|X}(u \mid x) \right|.$$

### A.1.4 CENSORING MARTINGALE BOUND

We now analyze the bias using the censoring martingale representation. Recall the bias of the pseudo-outcome is

$$B(t, x; \widehat{\eta}) = |\psi(t, x; \widehat{\eta}) - S_0(t \mid x)|.$$

Using the censoring martingale representation,

$$B(t,x;\widehat{\eta}) = \left| \mathbb{E}\left[ \frac{\delta \mathbf{1}[Y>t]}{\widehat{G}(Y \mid x)} - S_0(t \mid x) + \int_0^\infty \frac{\widehat{S}(\max\{u,t\} \mid x)}{\widehat{S}(u \mid x)\widehat{G}(u \mid x)} d\widehat{M}_{C|X}(u \mid x) \;\middle|\; X = x \right] \right|,$$

where the randomness is in the IPCW estimation and censoring process difference $d\widehat{M}_{C|X}(u \mid X)$. Let $\Lambda_{0,C|X}(t \mid X) = -\log G_0(t \mid X)$ and $\widehat{\Lambda}_{C|X}(t \mid X) = -\log \widehat{G}(t \mid X)$ be the true and estimated cumulative hazard functions for censoring, respectively. Under assumption A1,

$$\begin{aligned}
\mathbb{E}\left[dN_C(u) \mid X\right] &= \mathbb{P}(Y \in du, \delta = 0 \mid X) \\
&= \mathbb{P}(C \in du, T \geq u \mid X) \\
&= S_0(u \mid X) G_0(u \mid X)\, d\Lambda_{0,C|X}(u \mid X). \\
\mathbb{E}\left[\mathbf{1}[Y \geq u] \mid X\right] &= \mathbb{P}(Y \geq u \mid X) \\
&= \mathbb{P}(T \geq u, C \geq u \mid X) \\
&= S_0(u \mid X) G_0(u \mid X).
\end{aligned}$$

Thus,

$$\mathbb{E}\left[ d\widehat{M}_{C|X}(u \mid X) \mid X \right] = S_0(u \mid X) G_0(u \mid X) \left( d\Lambda_{0,C|X}(u \mid X) - d\widehat{\Lambda}_{C|X}(u \mid X) \right).$$

We can write the bias as

$$\begin{aligned}
B(t,X;\widehat{\eta}) = \Bigg| \mathbb{E}\left[ \frac{\delta \mathbf{1}[Y>t]}{\widehat{G}(Y \mid X)} + \int_t^\infty \frac{1}{\widehat{G}(u \mid X)}\, d\widehat{M}_{C|X}(u \mid X) \;\middle|\; X = x \right] - S_0(t \mid X) \\
+ \int_0^t \frac{\widehat{S}(t \mid x) S_0(u \mid x) G_0(u \mid x)}{\widehat{S}(u \mid x)\widehat{G}(u \mid x)} \left( d\Lambda_{0,C|X}(u \mid x) - d\widehat{\Lambda}_{C|X}(u \mid x) \right) \Bigg|.
\end{aligned}$$

We first simplify the expectation. Using Equation 9, we have

$$\frac{\mathbf{1}[Y>t]}{\widehat{G}(t \mid X)} = \frac{\delta \mathbf{1}[Y>t]}{\widehat{G}(Y \mid X)} + \int_t^\infty \frac{1}{\widehat{G}(u \mid X)}\, d\widehat{M}_{C|X}(u \mid X).$$

Taking conditional expectations yields

$$\mathbb{E}\left[ \frac{\delta \mathbf{1}[Y>t]}{\widehat{G}(Y \mid X)} + \int_t^\infty \frac{1}{\widehat{G}(u \mid X)}\, d\widehat{M}_{C|X}(u \mid X) \;\middle|\; X = x \right] = \frac{S_0(t \mid x) G_0(t \mid x)}{\widehat{G}(t \mid x)}.$$

Therefore,

$$\begin{aligned}
B(t,x;\widehat{\eta}) = \Bigg| S_0(t \mid x) \left( \frac{G_0(t \mid x)}{\widehat{G}(t \mid x)} - 1 \right) \\
+ \int_0^t \frac{\widehat{S}(t \mid x) S_0(u \mid x) G_0(u \mid X)}{\widehat{S}(u \mid x)\widehat{G}(u \mid x)} \left( d\Lambda_{0,C|X}(u \mid x) - d\widehat{\Lambda}_{C|X}(u \mid x) \right) \Bigg|.
\end{aligned}$$

Using Lemma 3,

$$S_0(t \mid x) \left( \frac{G_0(t \mid x)}{\widehat{G}(t \mid x)} - 1 \right) = -S_0(t \mid x) \int_0^t \frac{G_0(u \mid x)}{\widehat{G}(u \mid x)} \left( d\Lambda_{0,C|X}(u \mid x) - d\widehat{\Lambda}_{C|X}(u \mid x) \right).$$

Substituting this identity into the expression for the bias gives

$$B(t,x;\widehat{\eta}) = \left| \int_0^t \frac{G_0(u \mid x)}{\widehat{G}(u \mid x)} \left( \frac{\widehat{S}(t \mid x) S_0(u \mid x)}{\widehat{S}(u \mid x)} - S_0(t \mid x) \right) \left( d\Lambda_{0,C|X}(u \mid x) - d\widehat{\Lambda}_{C|X}(u \mid x) \right) \right|.$$

Under Assumption S1, there exists some constant $c_2(t)$ depending on $t$ such that $\widehat{G}(t \mid x) > 1/c_2(t)$. Then, we have the bound

$$0 \leq \frac{S_0(u \mid x) G_0(u \mid x)}{\widehat{G}(u \mid x)} \leq c_2(t), \quad \text{for all } u \in [0,t].$$

Therefore, we have that $B(t,x;\widehat{\eta}) \leq B_2(t,x;\widehat{\eta})$, where

$$B_2(t,x;\widehat{\eta}) = c_2(t) \int_0^t \left| \frac{\widehat{S}(t \mid x)}{\widehat{S}(u \mid x)} - \frac{S_0(t \mid x)}{S_0(u \mid x)} \right| \left| d\Lambda_{0,C|X}(u \mid x) - d\widehat{\Lambda}_{C|X}(u \mid x) \right|.$$

### A.1.5 POINTWISE CONVERGENCE OF BIAS

The mixed-product bounds imply pointwise convergence of the bias under local uniform consistency of either nuisance parameter. We formalize this in the following lemma.

**Lemma 5.** *Suppose the assumptions of Theorem 1 hold. Fix a finite time $t$ and a covariate value $x$ and let the bias of the pseudo-outcome be*

$$B(t, x; \widehat{\eta}) = |\mathbb{E}[\tilde{S}^{\mathrm{DR}}(t \mid X; \widehat{\eta}) \mid X = x] - S_0(t \mid x)|.$$

*If either of $\widehat{S}$ or $\widehat{G}$ converges to the true value uniformly on $[0, t]$,*

$$\sup_{u \in [0,t]} |\widehat{S}(u \mid x) - S_0(u \mid x)| \xrightarrow{p} 0 \quad \textit{or} \quad \sup_{u \in [0,t]} |\widehat{G}(u \mid x) - G_0(u \mid x)| \xrightarrow{p} 0,$$

*then, $B(t, x; \widehat{\eta}) \xrightarrow{p} 0$.*

*Proof.* By Theorem 1,

$$B(t, x; \widehat{\eta}) \leq \min\{B_1(t, x; \widehat{\eta}), B_2(t, x; \widehat{\eta})\}.$$

Therefore, it suffices to show that either $B_1(t, x; \widehat{\eta}) \xrightarrow{p} 0$ or $B_2(t, x; \widehat{\eta}) \xrightarrow{p} 0$. First suppose

$$\sup_{u \in [0,t]} |\widehat{G}(u \mid x) - G_0(u \mid x)| \xrightarrow{p} 0.$$

By assumption S1,

$$\sup_{u \in [0,t]} \left| \frac{1}{\widehat{G}(u \mid x)} - \frac{1}{G_0(u \mid x)} \right| = \sup_{u \in [0,t]} \frac{|\widehat{G}(u \mid x) - G_0(u \mid x)|}{\widehat{G}(u \mid x) G_0(u \mid x)} \leq c_2^2(t) \sup_{u \in [0,t]} |\widehat{G}(u \mid x) - G_0(u \mid x)|.$$

Moreover, we can bound the total variation distance between hazard difference measures using

$$\int_0^t \left| d\Lambda_{0,T|X}(u \mid x) - d\widehat{\Lambda}_{T|X}(u \mid x) \right| \leq \Lambda_{0,T|X}(t \mid x) + \widehat{\Lambda}_{T|X}(t \mid x) \leq 2 \log c_1(t).$$

Therefore,

$$\begin{aligned} B_1(t, x; \widehat{\eta}) &\leq c_1(t) \sup_{u \in [0,t]} \left| \frac{1}{\widehat{G}(u \mid x)} - \frac{1}{G_0(u \mid x)} \right| \int_0^t \left| d\Lambda_{0,T|X}(u \mid x) - d\widehat{\Lambda}_{T|X}(u \mid x) \right| \\ &\leq 2 c_1(t) \log c_1(t) \sup_{u \in [0,t]} \left| \frac{1}{\widehat{G}(u \mid x)} - \frac{1}{G_0(u \mid x)} \right| \\ &\leq 2 c_1(t) c_2^2(t) \log c_1(t) \sup_{u \in [0,t]} |\widehat{G}(u \mid x) - G_0(u \mid x)|. \end{aligned}$$

Thus, $B_1(t, x; \widehat{\eta}) \xrightarrow{p} 0$. Now suppose instead that

$$\sup_{u \in [0,t]} |\widehat{S}(u \mid x) - S_0(u \mid x)| \xrightarrow{p} 0.$$

By assumption S1,

$$\sup_{u \in [0,t]} \left| \frac{\widehat{S}(t \mid x)}{\widehat{S}(u \mid x)} - \frac{S_0(t \mid x)}{S_0(u \mid x)} \right| \leq c_1^2(t) \sup_{u \in [0,t]} \left| \widehat{S}(t \mid x) S_0(u \mid x) - S_0(t \mid x) \widehat{S}(u \mid x) \right|.$$

Using the triangle inequality,

$$\begin{aligned} \left| \widehat{S}(t \mid x) S_0(u \mid x) - S_0(t \mid x) \widehat{S}(u \mid x) \right| &\leq \left| \widehat{S}(t \mid x) - S_0(t \mid x) \right| S_0(u \mid x) \\ &\quad + \left| S_0(u \mid x) - \widehat{S}(u \mid x) \right| S_0(t \mid x) \\ &\leq 2 \sup_{v \in [0,t]} |\widehat{S}(v \mid x) - S_0(v \mid x)|. \end{aligned}$$

Like before, we can bound the total variation of the hazard difference measures using

$$\int_0^t \left| d\Lambda_{0,C|X}(u \mid x) - d\widehat{\Lambda}_{C|X}(u \mid x) \right| \leq \Lambda_{0,C|X}(t \mid x) + \widehat{\Lambda}_{C|X}(t \mid x) \leq 2 \log c_2(t).$$

Therefore,

$$\begin{aligned} B_2(t, x; \widehat{\eta}) &\leq c_2(t) \sup_{u \in [0,t]} \left| \frac{\widehat{S}(t \mid x)}{\widehat{S}(u \mid x)} - \frac{S_0(t \mid x)}{S_0(u \mid x)} \right| \int_0^t \left| d\Lambda_{0,C|X}(u \mid x) - d\widehat{\Lambda}_{C|X}(u \mid x) \right| \\ &\leq 2c_2(t) \log c_2(t) \sup_{u \in [0,t]} \left| \frac{\widehat{S}(t \mid x)}{\widehat{S}(u \mid x)} - \frac{S_0(t \mid x)}{S_0(u \mid x)} \right| \\ &\leq 4c_2(t) c_1^2(t) \log c_2(t) \sup_{u \in [0,t]} |\widehat{S}(u \mid x) - S_0(u \mid x)|. \end{aligned}$$

Thus, $B_2(t, x; \widehat{\eta}) \xrightarrow{p} 0$. Combining the two cases proves that $B(t, x; \widehat{\eta}) \xrightarrow{p} 0$. □

### A.2 Proof of Theorem 2

Our proof proceeds in 4 main steps.

(i) We first show that the isotonic regression estimate at any specific time converges to the $L_2(P)$ projection of the true survival function. (ii) Next, we prove pointwise consistency at a specific time. (iii) Then, we prove uniform consistency at a specific time. (iv) Finally, we show how these uniformly consistent estimates at each time are also naturally non-increasing in time, leading to Dykstra's projection algorithm terminating after 1 step and forming our DR-ISR estimate.

#### A.2.1 Convergence to $L_2(P)$ projection

**Additional Notation:** For this part of the proof, we first fix a time $t \geq 0$.

Let us define the $L_2(P)$ distance as usual

$$\|f - g\|_P = \left( \int (f(r) - g(r))^2 p(r) dr \right)^{\frac{1}{2}}.$$

Let $\mathcal{M}$ be the class of non-increasing 1D functions. Define

$$S^* = \underset{g \in \mathcal{M}}{\arg\min}\, \mathbb{E}_{X \sim P_X} \left[ (S_0(t \mid X) - g(r(X)))^2 \right].$$

If we add and subtract $S_{0,r}$ inside the square above, we have

$$\begin{aligned} &\mathbb{E}_{X \sim P_X} \left[ (S_0(t \mid X) - g(r(X)))^2 \right] \\ =&\mathbb{E}_X[(S_0(t \mid X) - S_{0,r}(t \mid r(X)))^2] + \mathbb{E}_X[(S_{0,r}(t \mid r(X)) - g(r(X)))^2] \\ &+ 2\mathbb{E}_X[(S_0(t \mid X) - S_{0,r}(t \mid r(X)))(S_{0,r}(t \mid r(X)) - g(r(X)))]. \end{aligned}$$

Conditioning on $r(X)$ first and applying tower rule to the cross term yields

$$\begin{aligned} &2\mathbb{E}_X[(S_0(t \mid X) - S_{0,r}(t \mid r(X)))(S_{0,r}(t \mid r(X)) - g(r(X)))] \\ =&2\mathbb{E}_{r(X)} \left[ (S_{0,r}(t \mid r(X)) - g(r(X))) \cdot \mathbb{E}_{X|r(X)}[S_0(t \mid X) - S_{0,r}(t \mid r(X)) \mid r(X)] \right] \\ =&2\mathbb{E}_{r(X)} \left[ (S_{0,r}(t \mid r(X)) - g(r(X))) \cdot 0 \right] \\ =&0. \end{aligned}$$

We then have,

$$\begin{aligned} &\mathbb{E}_{X \sim P_X} \left[ (S_0(t \mid X) - g(r(X)))^2 \right] \\ =&\mathbb{E}_X[(S_0(t \mid X) - S_{0,r}(t|r(X)))^2] + \mathbb{E}_X[(S_{0,r}(t \mid r(X)) - g(r(X)))^2]. \end{aligned}$$

Since the first term does not depend on $g$, minimizing $\mathbb{E}_{X \sim P_X} \left[ (S_0(t \mid X) - g(r(X)))^2 \right]$ over $g$ is the same as minimizing $\mathbb{E}_X[(S_{0,r}(t \mid r(X)) - g(r(X)))^2]$ over $g$. Hence, $S^*$ is equivalently the $L_2(P)$ projection of $S_{0,r}$ onto $\mathcal{M}$.

Note that if the risk function is well specified such that the true survival probabilities are non increasing in the risk (assumption A2), then $S^*$ is exactly equal to $S_{0,r}(t \mid r)$. However, we will prove convergence of the isotonic regression estimator to $S^*$ in probability in the $L_2(P)$ norm without this assumption since this will be useful in our proof of Theorem 3.

Let $\hat{f}_n$ denote the IR solution when using $n$ calibration datapoints to regress the pseudo-outcomes onto the risks. That is,

$$\hat{f}_n = \underset{g \in \mathcal{M}}{\arg\min} \sum_{i=1}^{n} (\tilde{S}_{\text{DR}}(t \mid X_i; \widehat{\eta}_n) - g(r(X_i)))^2,$$

where $\widehat{\eta}_n = (\widehat{S}_n, \widehat{G}_n)$ are the initial estimates of the survival and censoring models used to construct the pseudo-outcomes when $n$ calibration datapoints were available. Specifically, we assume that corresponding to each possible calibration set size of $n$, there is a training data set over which initial survival and censoring models are estimated.

We define the errors as

$$\epsilon_{n,i} = \tilde{S}_{\text{DR}}(t \mid X_i; \widehat{\eta}_n) - S_0(t \mid X_i).$$

Then, by Lemma 5, we have that if either

$$\sup_{u\in[0,t]} |\widehat{S}(u \mid x) - S_0(u \mid x)| \xrightarrow{p} 0 \quad \text{or} \quad \sup_{u\in[0,t]} |\widehat{G}(u \mid x) - G_0(u \mid x)| \xrightarrow{p} 0,$$

then $\mathbb{E}[\epsilon_{n,i} \mid X_i] \to 0$ as $n \to \infty$.

Substituting the pseudo-label expression from Eq. 2, the errors are given by

$$\epsilon_{n,i} = \widehat{S}_n(t \mid X_i) - S_0(t \mid X_i) - \int_0^t \frac{\widehat{S}_n(t \mid X_i)}{\widehat{S}_n(u \mid X_i)\widehat{G_n}(u \mid X_i)}\, d\widehat{M}_{T|X}(u \mid X_i).$$

Now, $\widehat{S}_n, \widehat{G}_n$ are estimated independently from the calibration data, and the random variable $d\widehat{M}_{T|X}(u|X)$ only depends on $\widehat{S}_n$ and the observed outcome for the $i^{th}$ calibration data point. Hence, the errors across the calibration data are independent at any fixed $n$.

Lastly, since $t$ is fixed, the pseudo-outcomes are bounded by assumption S1, and hence the errors are bounded as well. Let the errors be bounded in absolute value by $M$.

**Step 1:** Let $\mu_{n,i} = \mathbb{E}[\epsilon_{n,i} \mid X_i]$ be the conditional bias and $\nu_{n,i} = \epsilon_{n,i} - \mu_{n,i}$ be the zero-mean stochastic noise.

Note that given $X_i$, $\epsilon_{n,i}$ only depends on the realized outcome for $X_i$. Since the calibration data points are all independent, the conditional biases are also independent.

Since the $\mu_{n,i}$ and $\epsilon_{n,i}$ are independent across the calibration data, the zero-mean stochastic noise $\nu_{n,i}$ is also independent across the calibration data

We know $|\mu_{n,i}| \le M$ and $\mu_{n,i} \xrightarrow{p} 0$. By the bounded convergence theorem, we can interchange the limit and integral, proving that $\mathbb{E}[|\mu_{n,i}|] \to 0$ for all $i$.

By Markov's inequality, the empirical average of the absolute conditional bias goes to zero in probability

$$P\left(\frac{1}{n}\sum_{i=1}^{n} |\mu_{n,i}| > \delta\right) \le \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[|\mu_{n,i}|]}{\delta} \le \frac{\max_{1\le i\le n} \mathbb{E}[|\mu_{n,i}|]}{\delta} \to 0.$$

Thus, $\frac{1}{n}\sum_{i=1}^{n} |\mu_{n,i}| \xrightarrow{p} 0$.

The remaining stochastic noise $\nu_{n,i}$ forms a triangular array satisfying $\mathbb{E}[\nu_{n,i} \mid X_i] = 0$, is conditionally independent across $i$, and is strictly bounded in $[-2M, 2M]$.

**Step 2:** First define the empirical risk over the noisy data, shifted by the errors as

$$\hat{R}_n(g) := \frac{1}{n}\sum_{i=1}^{n} \left[(\tilde{S}_{\text{DR}}(t \mid X_i; \widehat{\eta}_n) - g(r(X_i)))^2 - \epsilon_{n,i}^2\right].$$

Also define the empirical risk over the noiseless data as

$$\tilde{R}_n(g) := \frac{1}{n}\sum_{i=1}^{n} (S_0(t \mid X_i) - g(r(X_i)))^2.$$

Finally, we have the population risk

$$R(g) := \mathbb{E}_{X\sim P_X}[(S_0(t \mid X) - g(r(X)))^2].$$

Because $S_0$ is bounded by 1, $S^*$ is also bounded by 1. Furthermore, since $|\tilde{S}_{\text{DR}}(t \mid X_i; \widehat{\eta}_n)| \le 1 + M$, the isotonic regression estimator $\hat{f}_n$ is bounded by $1 + M$. Let $K = 1 + M$. We can restrict our search space to $\mathcal{M}^K = \{g \in \mathcal{M} : \|g\|_\infty \le K\}$.

Then for any $g \in \mathcal{M}^K$,

$$\begin{aligned}\hat{R}_n(g) - \tilde{R}_n(g) &= \frac{1}{n}\sum_{i=1}^{n}\left((S_0(t \mid X_i) + \epsilon_{n,i} - g(r(X_i)))^2 - \epsilon_{n,i}^2 - (S_0(t \mid X_i) - g(r(X_i)))^2\right) \\ &= \frac{2}{n}\sum_{i=1}^{n}\epsilon_{n,i}(S_0(t \mid X_i) - g(r(X_i))).\end{aligned}$$

Decomposing $\epsilon_{n,i} = \mu_{n,i} + \nu_{n,i}$, we can bound the bias portion uniformly over all $g \in \mathcal{M}^K$ as

$$\sup_{g\in\mathcal{M}^K}\left|\frac{2}{n}\sum_{i=1}^{n}\mu_{n,i}(S_0(t \mid X_i) - g(r(X_i)))\right| \leq 2(1+K)\frac{1}{n}\sum_{i=1}^{n}|\mu_{n,i}|$$

From Step 1, this bias term converges to 0 in probability.

**Step 3:** Note that $\mathcal{M}^K$ is a class of bounded monotonic functions. By Theorem 2.7.5 of Van Der Vaart and Wellner (1996),

$$\log N_{[]}(\epsilon, \mathcal{M}^K, L_1(P)) \leq \frac{K \cdot C}{\epsilon},$$

where $N_{[]}$ is the bracketing number and $C$ is some global constant. Therefore, the bracketing number is finite for all $\epsilon > 0$.

Define the class of difference functions $\mathcal{H} = \{h_g(x) = S_0(t \mid x) - g(r(x)) : g \in \mathcal{M}^K\}$. Since this is simply a bounded shift of $\mathcal{M}^K$, its bracketing number is also finite.

Since the multipliers $\nu_{n,i}$ form a triangular array of conditionally independent, mean-zero random variables strictly bounded by $2M$, and $\mathcal{H}$ is a Glivenko-Cantelli class, the multiplier empirical process converges uniformly to zero in probability

$$\sup_{g\in\mathcal{M}^K}\left|\frac{2}{n}\sum_{i=1}^{n}\nu_{n,i}(S_0(t \mid X_i) - g(r(X_i)))\right| = \sup_{h_g\in\mathcal{H}}\left|\frac{2}{n}\sum_{i=1}^{n}\nu_{n,i}h_g(X_i)\right| \xrightarrow{p} 0.$$

Combined with the uniform bound on the bias from Step 2, we have

$$\sup_{g\in\mathcal{M}^K}|\hat{R}_n(g) - \tilde{R}_n(g)| \xrightarrow{p} 0.$$

Now define $\mathcal{F} = \{h_g(x) = (S_0(t \mid x) - g(r(x)))^2 : g \in \mathcal{M}^K\}$.

For any two $g_1, g_2 \in \mathcal{M}^K$,

$$\begin{aligned}|h_{g_1}(x) - h_{g_2}(x)| &= |(S_0(t \mid x) - g_1(x))^2 - (S_0(t \mid x) - g_2(x))^2| \\ &= |2S_0(t \mid x) - g_1(x) - g_2(x)| \cdot |g_2(x) - g_1(x)| \\ &\leq 2(1+K) \cdot |g_2(x) - g_1(x)|.\end{aligned}$$

Therefore, if $[l, u]$ is an $\epsilon$ bracket for $\mathcal{M}^K$, it is an $2\epsilon(1+K)$ bracket for $\mathcal{F}$. Therefore, the bracketing number for $\mathcal{F}$ is also finite for all $\epsilon > 0$.

By Theorem 2.4.1 of Van Der Vaart and Wellner (1996), $\mathcal{F}$ is Glivenko-Cantelli. That is,

$$\sup_{g\in\mathcal{M}^K}|\tilde{R}_n(g) - R(g)| \xrightarrow{a.s.} 0.$$

**Step 4:** From Steps 2 and 3,

$$\begin{aligned}\sup_{g\in\mathcal{M}^K}|\hat{R}_n(g) - R(g)| &\leq \sup_{g\in\mathcal{M}^K}|\hat{R}_n(g) - \tilde{R}_n(g)| + \sup_{g\in\mathcal{M}^K}|\tilde{R}_n(g) - R(g)| \\ &\xrightarrow{p} 0 + 0.\end{aligned}$$

Now, $\hat{f}_n$ is the minimizer of $\hat{R}_n(g)$ over $\mathcal{M}^K$. Hence, $\hat{R}_n(\hat{f}_n) \leq \hat{R}_n(S^*)$. Also, $S^*$ is minimizer of $R(g)$ over $\mathcal{M}^K$. Hence, $R(S^*) \leq R(\hat{f}_n)$. Combining,

$$\begin{aligned}
0 &\leq R(\hat{f}_n) - R(S^*) \\
&= (R(\hat{f}_n) - \hat{R}_n(\hat{f}_n)) + (\hat{R}_n(\hat{f}_n) - \hat{R}_n(S^*)) + (\hat{R}_n(S^*) - R(S^*)) \\
&\leq (R(\hat{f}_n) - \hat{R}_n(\hat{f}_n)) + 0 + (\hat{R}_n(S^*) - R(S^*)) \\
&\leq 2 \sup_{g \in \mathcal{M}^K} |\hat{R}_n(g) - R(g)| \\
&\xrightarrow{p} 0.
\end{aligned}$$

Hence, $R(\hat{f}_n) \xrightarrow{p} R(S^*)$.

**Step 5:** Note that $\mathcal{M}$ is a closed convex cone. Hence, if $S^*$ is the projection of $S_0$ onto $\mathcal{M}$, then for any $g \in \mathcal{M}$,

$$\begin{aligned}
\mathbb{E}_{X \sim P_X}\left[(S_0(t \mid X) - g(r(X)))^2\right] \geq &\mathbb{E}_{X \sim P_X}\left[(S_0(t \mid X) - S^*(r(X)))^2\right] \\
&+ \mathbb{E}_{X \sim P_X}\left[(g(r(X)) - S^*(r(X)))^2\right].
\end{aligned}$$

Note that $R(g) = \mathbb{E}_{X \sim P_X}[(S_0(t \mid X) - g(r(X)))^2]$ and pick $g = \hat{f}_n$, to get

$$\mathbb{E}_{X \sim P_X}\left[(\hat{f}_n(r(X)) - S^*(r(X)))^2\right] \leq R(\hat{f}_n) - R(S^*).$$

Combining with Step 4,

$$\mathbb{E}_{X \sim P_X}\left[(\hat{f}_n(r(X)) - S^*(r(X)))^2\right] \xrightarrow{p} 0.$$

Since both $\hat{f}_n$ and $S^*$ only depend on $X$ through $r$ and $P$ was defined as the push forward measure of $r$, we can rewrite the above convergence as

$$\|\hat{f}_n - S^*\|_P \xrightarrow{p} 0.$$

**Step 6** Under assumption A2, we have that $S_{0,r}(t \mid r)$ is non-increasing itself ($S_{0,r}(t \mid r) \in \mathcal{M}$) and recall that $S^*$ can alternatively be viewed as the $L_2(P)$ projection of $S_{0,r}$ onto $\mathcal{M}$. Therefore, $S^* = S_{0,r}(t \mid r)$ implying

$$\|\hat{f}_n(r) - S_{0,r}(t \mid r)\|_P \xrightarrow{p} 0.$$

### A.2.2 Pointwise Convergence

We now show that pointwise convergence also holds. That is,

$$\lim_{n \to \infty} |\hat{f}_n(r) - S_{0,r}(t \mid r)| \xrightarrow{p} 0.$$

We proceed by contradiction. Let pointwise convergence fail at $r_0$. Since

$$P(|\hat{f}_n(r_0) - S_{0,r}(t \mid r_0)| \geq \epsilon) = P(\hat{f}_n(r_0) - S_{0,r}(t \mid r_0) \geq \epsilon) + P(\hat{f}_n(r_0) - S_{0,r}(t \mid r_0) \leq -\epsilon),$$

for pointwise convergence to fail, there exists $\epsilon > 0$ such that at least one of $P(\hat{f}_n(r_0) - S_{0,r}(t \mid r_0) \geq \epsilon) \nrightarrow 0$ or $P(\hat{f}_n(r_0) - S_{0,r}(t \mid r_0) \leq -\epsilon) \nrightarrow 0$ must be true.

We consider the case that $P(\hat{f}_n(r_0) - S_{0,r}(t \mid r_0) \geq \epsilon) \nrightarrow 0$; the contradiction for the other case will follow symmetrically.

Since $S_{0,r}(t \mid r)$ is continuous at $r_0$, there exists $\delta > 0$ such that for all $r \in [r_0 - \delta, r_0]$,

$$S_{0,r}(t \mid r) - S_{0,r}(t \mid r_0) < \frac{\epsilon}{2}.$$

Since $\hat{f}_n$ in non-increasing, for all $x \in [r_0 - \delta, r_0]$, we have $\hat{f}_n(r) \geq \hat{f}_n(r_0)$. If, $\hat{f}_n(r_0) \geq S_{0,r}(t \mid r_0) + \epsilon$ then, for all $r \in [r_0 - \delta, r_0]$,

$$\hat{f}_n(r) \geq S_{0,r}(t \mid r_0) + \epsilon.$$

Combining the two,

$$\hat{f}_n(r) - S_{0,r}(t \mid r) > (S_{0,r}(t \mid r_0) + \epsilon) - \left(S_{0,r}(t \mid r_0) + \frac{\epsilon}{2}\right) = \frac{\epsilon}{2} \quad \text{for all } r \in [r_0 - \delta, r_0].$$

But this implies the $L_2(P)$ distance is then at least

$$\|\hat{f}_n(r) - S_{0,r}(t \mid r)\|_P^2 \geq \int_{r_0-\delta}^{r_0} (\hat{f}_n(r) - S_{0,r}(t \mid r))^2 p(r) dr \geq \frac{\epsilon^2}{4} \int_{r_0-\delta}^{r_0} p(r) dr > 0,$$

since $p(t)$ is strictly positive by assumption S2.

Hence,

$$P(\|\hat{f}_n(r) - S_{0,r}(t \mid r)\|_P^2 > 0) \geq P(\hat{f}_n(r_0) - S_{0,r}(t \mid r_0) \geq \epsilon).$$

By our assumption then, for $c = \frac{\epsilon^2}{4} \int_{r_0-\delta}^{r_0} p(r) dr$,

$$P(\hat{f}_n(r_0) - S_{0,r}(t \mid r_0) \geq \epsilon) \nrightarrow 0 \implies P(\|\hat{f}_n(r) - S_{0,r}(t \mid r)\|_P^2 > c) \nrightarrow 0,$$

leading to a contradiction.

### A.2.3 Uniform Convergence

From the previous part, we have that pointwise convergence holds at all $r \in (a, b)$. Consider some compact interval $[c, d] \subset (a, b)$. Note that we have pointwise convergence on $[c, d]$.

Since $S_{0,r}(t \mid r)$ is continuous on $[c, d] \subset (a, b)$ and hence we can pick $M$ points such that $c = u_0 < u_1 < ... < u_M = d$ such that

$$S_{0,r}(t \mid u_{m-1}) - S_{0,r}(t \mid u_m) \leq \epsilon \quad \text{for all } m = 1, \ldots, M$$

for some $\epsilon > 0$.

Fix a $\delta > 0$. From the definition of pointwise convergence, there exist $N_0, ..., N_M$ such that for any $u_m$

$$P(|\hat{f}_n(u_m) - S_{0,r}(t \mid u_m)| > \epsilon) \leq \frac{\delta}{M+1} \quad \forall\, n > N_m.$$

Let $N = \max\{N_0, ..., N_m\}$. Then for any $n > N$, by the union bound and pointwise convergence,

$$P\left(\max_{0 \leq m \leq M} |\hat{f}_n(u_m) - S_{0,r}(t \mid u_m)| > \epsilon\right) \leq \sum_{m=0}^{M+1} P\left(|\hat{f}_n(u_m) - S_{0,r}(t \mid u_m)| > \epsilon\right) \leq \delta.$$

Now, consider some $r \in [c, d]$. Let $u_{m-1} \leq x \leq u_m$ be the interval $r$ is in. $\hat{f}_n$ is non-increasing so we have

$$|\hat{f}_n(r) - S_{0,r}(t \mid r)| \leq \max\{\hat{f}_n(u_{m-1}) - S_{0,r}(t \mid r), -\hat{f}_n(u_m) + S_{0,r}(t \mid r)\}.$$

Note that by construction and non-increasingness of $S_{0,r}(t \mid r)$

$$\begin{aligned}\hat{f}_n(u_{m-1}) - S_{0,r}(t \mid r) &= \hat{f}_n(u_{m-1}) - S_{0,r}(t \mid u_{m-1}) - (S_{0,r}(t \mid r) - S_{0,r}(t \mid u_{m-1})) \\ &\leq \hat{f}_n(u_{m-1}) - S_{0,r}(t \mid u_{m-1}) + \epsilon,\end{aligned}$$

and

$$\begin{aligned}-\hat{f}_n(u_m) + S_{0,r}(t \mid r) &= -\hat{f}_n(u_m) + S_{0,r}(t \mid u_m) + (S_{0,r}(t \mid r) - S_{0,r}(t \mid u_m)) \\ &\leq -\hat{f}_n(u_m) + S_{0,r}(t \mid u_m) + \epsilon.\end{aligned}$$

Combining,

$$\begin{aligned}|\hat{f}_n(r) - S_{0,r}(t \mid r)| &\leq \max\{\hat{f}_n(u_{m-1}) - S_{0,r}(t \mid u_{m-1}), -\hat{f}_n(u_m) + S_{0,r}(t \mid u_m)\} + \epsilon \\ &\leq \max_{0 \leq m \leq M} |\hat{f}_n(u_m) - S_{0,r}(t \mid u_m)| + \epsilon.\end{aligned}$$

Note that the upper bound above does not depend on $r$. Hence, for any $n > N$,

$$\begin{aligned} P\left(\sup_{r\in[c,d]} |\hat{f}_n(r) - S_{0,r}(t \mid r)| > 2\epsilon\right) &\leq P\left(\max_{0\leq m\leq M} |\hat{f}_n(u_m) - S_{0,r}(t \mid u_m)| + \epsilon > 2\epsilon\right) \\ &= P\left(\max_{0\leq m\leq M} |\hat{f}_n(u_m) - S_{0,r}(t \mid u_m)| > \epsilon\right) \\ &\leq \delta. \end{aligned}$$

Since $\epsilon, \delta > 0$ were arbitrary, we have the desired uniform convergence over all possible compact subsets $[c, d] \subset (a, b)$

$$\sup_{r\in[c,d]} |\hat{f}_n(r) - S_{0,r}(t \mid r)| \xrightarrow{p} 0.$$

### A.2.4 Convergence of Dykstra's

We have proved that the isotonic regression estimators $\hat{f}_n$ constructed at each fixed $t$ converge to the population projections. We now show that the population projections $S^*$ of these estimators are non-increasing in time. As such, Dykstra's algorithm which is used to find the solution to the optimization problem from Eq. 3 asymptotically converges in one step after ensuring monotonicity along risks.

Let $t_j \leq t_i$. Let $\hat{f}_n^i$ and $\hat{f}_n^j$ be the isotonic regression estimate from time $t_i$ and $t_j$ respectively. From Sec. A.2.1, we have that $\|\hat{f}_n^i - S^{i*}\|_P \xrightarrow{p} 0$ and $\|\hat{f}_n^j - S^{j*}\|_P \xrightarrow{p} 0$ where

$$S^{i*} = \arg\min_{g\in\mathcal{M}} \int (S_{0,r}(t_i \mid r) - g(r))^2 p(r) dr,$$

and

$$S^{j*} = \arg\min_{g\in\mathcal{M}} \int (S_{0,r}(t_j \mid r) - g(r))^2 p(r) dr.$$

We now proceed by contradiction.

Suppose the set $A = \{r \mid S^{i*}(r) > S^{j*}(r)\}$ has positive measure under $P$. Intuitively, this is the set where a violation of monotonicity along time occurs.

Define the functions $g_{\min}(x) = \min(S^{i*}(r), S^{j*}(r))$ and $g_{\max}(r) = \max(S^{i*}(r), S^{j*}(r))$. Because the class of non-increasing functions $\mathcal{M}$ is closed under minimum and maximum operations, both $g_{\min}$ and $g_{\max}$ are in $\mathcal{M}$.

By the definition of $S^{i*}$ and $S^{j*}$ as the optimal $L_2(P)$ projections onto $\mathcal{M}$, replacing them with $g_{\min}$ and $g_{\max}$ respectively, cannot decrease the objective function. Therefore, we have

$$\begin{aligned} \Delta_{\min} &= \int (S_{0,r}(t_i \mid r) - g_{\min}(r))^2 p(r) dr - \int (S_{0,r}(t_i \mid r) - S^{i*}(r))^2 p(r) dr \geq 0, \\ \Delta_{\max} &= \int (S_{0,r}(t_j \mid r) - g_{\max}(r))^2 p(r) dr - \int (S_{0,r}(t_j \mid r) - S^{j*}(r))^2 p(r) dr \geq 0. \end{aligned}$$

Observe that on the complement $A^c$, $S^{i*}(r) \leq S^{j*}(r)$, which implies $g_{\min} = S^{i*}$ and $g_{\max} = S^{j*}$. Thus, the integrands of $\Delta_{\min}$ and $\Delta_{\max}$ evaluate to exactly zero outside of $A$.

We can therefore restrict the sum of the differences strictly to the set $A$, where $g_{\min} = S^{j*}$ and $g_{\max} = S^{i*}$,

$$\begin{aligned} \Delta_{\min} + \Delta_{\max} = &\int_A \Big[(S_{0,r}(t_i \mid r) - S^{j*}(r))^2 - (S_{0,r}(t_i \mid r) - S^{i*}(r))^2\Big] p(r) dr \\ &+ \int_A \Big[(S_{0,r}(t_j \mid r) - S^{i*}(r))^2 - (S_{0,r}(t_j \mid r) - S^{j*}(r))^2\Big] p(r) dr. \end{aligned}$$

Expanding the squares and simplifying the integrand yields

$$\begin{aligned}
&S_{0,r}(t_i \mid r)^2 - 2S_{0,r}(t_i \mid r)S^{j*}(r) + S^{j*}(r)^2 - S_{0,r}(t_i \mid r)^2 + 2S_{0,r}(t_i \mid r)S^{i*}(r) - S^{i*}(r)^2 \\
&+S_{0,r}(t_j \mid r)^2 - 2S_{0,r}(t_j \mid r)S^{i*}(r) + S^{i*}(r)^2 - S_{0,r}(t_j \mid r)^2 + 2S_{0,r}(t_j \mid r)S^{j*}(r) - S^{j*}(r)^2 \\
&= 2S_{0,r}(t_i \mid r)S^{i*}(r) - 2S_{0,r}(t_i \mid r)S^{j*}(r) - 2S_{0,r}(t_j \mid r)S^{i*}(r) + 2S_{0,r}(t_j \mid r)S^{j*}(r) \\
&= 2(S_{0,r}(t_i \mid r) - S_{0,r}(t_j \mid r))(S^{i*}(r) - S^{j*}(r)).
\end{aligned}$$

Substituting this back into our integral, we get

$$\Delta_{\min} + \Delta_{\max} = \int_A 2(S_{0,r}(t_i \mid r) - S_{0,r}(t_j \mid r))(S^{i*}(r) - S^{j*}(r))p(r)dr.$$

We now analyze the signs of the terms in the integrand on the set $A$.

First, $S_{0,r}(t_i \mid r) \leq S_{0,r}(t_j \mid r)$ since survival curves are non-increasing, so $S_{0,r}(t_i \mid r) - S_{0,r}(t_j \mid r) \leq 0$. Second, $S^{i*}(r) > S^{j*}(r)$ by the definition of $A$, so $S^{i*}(r) - S^{j*}(r) > 0$. Lastly, $p(r) > 0$ strictly everywhere, by assumption.

Multiplying these terms together, the integrand is non-positive everywhere on $A$, which means the entire integral evaluates to $\Delta_{\min} + \Delta_{\max} \leq 0$. However, since $\Delta_{\min} \geq 0$ and $\Delta_{\max} \geq 0$, we must have $\Delta_{\min} + \Delta_{\max} = 0$.

Now, since $\Delta_{\min} = 0$, the function $g_{\min}$ achieves the exact same optimal $L_2(P)$ error as $S^{i*}$. However, the space of monotonic functions $\mathcal{M}$ is a closed convex subset of $L_2(P)$. By the strict convexity of the $L_2(P)$ norm ($P$ has strictly positive density by assumption S2), the projection of $S_{0,r}(t_i \mid \cdot)$ onto $\mathcal{M}$ is strictly unique. Therefore, $g_{\min}$ and $S^{i*}$ must be identical almost everywhere. By definition, $g_{\min}(r) = \min(S^{i*}(r), S^{j*}(r))$, we must have $S^{i*}(r) \leq S^{j*}(r)$ almost everywhere. This contradicts the assumption that the set $A = \{r \mid S^{i*}(r) > S^{j*}(r)\}$ has positive measure.

Thus, $P(A) = 0$, and $S^{i*}(r) \leq S^{j*}(r)$ almost everywhere. Therefore, the set $A$ must have a measure of zero under $P$. Because $p(r)$ is strictly positive on its domain, $S^{i*}(r) \leq S^{j*}(r)$ for almost all $r$.

Hence, the solution to the optimization problem defining $\widehat{S}^{\mathrm{DR}}_{\mathrm{ISR}}$ (Eq. 3), is asymptotically found by simply ensuring monotonicity along risks. Hence, to predict the survival probability at any $t \in \mathcal{T}$, it suffices to find the isotonic regression estimator enforcing monotonicity along risks for the pseudo-outcomes constructed over calibration data at $t$.

More formally, we can once again use the exact same empirical process arguments from A.2.1 to show that $\widehat{S}^{\mathrm{DR}}_{\mathrm{ISR}}$ — defined as the empirical risk mimimizer over the class of product of $\mathcal{T}$ monotonic functions — converges to the population projection since the bracketing entropy of product of $\mathcal{T}$ monotonic functions is finite (Theorem 2.7.5 of Van Der Vaart and Wellner (1996)). We have just showed that the population projection onto the class of functions non-increasing in risk is also non-increasing in time. Hence, at any time $t \in \mathcal{T}$, the DR-ISR estimator converges to the isotonic regression estimator enforcing monotonicity along risks for the pseudo-outcomes constructed over calibration data at $t$.

By assumption A2 we get $S^* = S_{0,r}(t \mid r)$ as the asymptotic solution enforcing non-increasingness along risks. Combining with lemma 1 which ensures $S_0(t \mid x) = S_{0,r}(t \mid r(x))$, we have the desired result

$$\sup_{x \in \mathcal{X}: r(x) \in \mathcal{K}} |\widehat{S}^{\mathrm{DR}}_{\mathrm{ISR}}(t, x) - S_0(t \mid x)| \xrightarrow{p} 0, \qquad \text{for all } t \geq 0, \tag{10}$$

where $\mathcal{K}$ is any compact subset of $\mathrm{int}(r(\mathcal{X}))$.

### A.3 Proof of Theorem 3

Once again, first consider a fixed time $t \geq 0$.

Define $G(x) = \int_a^x (1 - S_{0,r}(t \mid r))p(r)dr$ and let $F(x) = \int_a^x p(r)dr$ be the CDF of $P$.

Consider the parametric curve $C = \{(F(x), G(x)\, x \in [a, b]\}$.

Then, let $G^*(x)$ be such that $C^* = \{(F(x), G^*(x)\, x \in [a, b]\}$ is the greatest convex minorant (GCM) of $C$.

As before, let

$$S^* = \underset{g \in \mathcal{M}}{\arg\min}\, \mathbb{E}_{X \sim P_X} \left[ (S_0(t \mid X) - g(r(X)))^2 \right].$$

From A.2.1 recall that this is equivalently the $L_2(P)$ projection of $S_{0,r}(t \mid r)$ onto the class of non-increasing functions.

Further define $f^* = 1 - S^*$. $f^*$ can then be seen the $L_2(P)$ projection of the survival CDF, $1 - S_{0,r}(t \mid r)$ onto the class of non-decreasing functions.

It is well known that $f^*(r)$ is the subdifferential of $C^*$ at $F(r)$ (Barlow, 1972; Groeneboom and Jongbloed, 2014). Since $C^*$ is convex, the subdifferential is guaranteed to not be empty over the interior $(0, 1)$.

For ease of exposition, we consider the unique $f^*$ which is defined as the right derivative of $C^*$ at $F(r)$. Notably, implementations of isotonic regression algorithms, including ours, use this convention. We note, however, that the proof below easily extends without this right derivative assumption.

**Case I: $f^*(r) = q$ for multiple $r$** Let $[a, b)$ be the maximal interval where $f^*(r) = q$. Note that $a \neq b$ since $f^*(r) = q$ for multiple $r$. Also, note that by monotonicity of $f^*$, $f^*(r) < q$ for $r < a$ and $f^*(r) > q$ for $r > b$.

Since $f^*(r)$ is constant over this interval, the portion of the curve $C^*$ between $F(a)$ and $F(b)$ is a straight line. By the properties of the GCM, this implies that $C^*$ touches $C$ at $F(a)$ and $F(b)$. Hence, $G(a) = G^*(a)$ and $G(b) = G^*(b)$.

Since the portion of the curve $C^*$ between $F(a)$ and $F(b)$ is a straight line, we can write the vertical change as the slope (i.e $q$) times the horizontal change (i.e. $(F(b) - F(a))$) to get

$$\begin{aligned}
&G(b) - G(a) = G^*(b) - G^*(a) = q(F(b) - F(a)) \\
\implies &\int_a^b (1 - S_{0,r}(t \mid r))p(r)dr = q \int_a^b p(r)dr \\
\implies &\frac{\int_a^b (1 - S_{0,r}(t \mid r))p(r)dr}{\int_a^b p(r)dr} = q \qquad \left( \int_a^b p(r)dr > 0 \text{ since } p(r) > 0 \text{ everywhere} \right) \\
\implies &\mathbb{E}[1 - S_{0,r}(t \mid r) | f^*(r) = q] = q,
\end{aligned}$$

where the expectation is taken treating $r$ as a random variable and conditioning on the event $f^*(r) = q$.

**Case II: $f^*(r) = q$ for a unique $r$** Let $r_0$ be the unique point where $f^*(r_0) = q$. Since $f^*(r_0)$ is the (right) derivative of $C^*$ at $F(r_0)$, $C^*$ is strictly increasing at $F(r_0)$.

By the properties of the GCM, the GCM must touch $C$ at $F(r_0)$, i.e. $G(r_0) = G^*(r_0)$. Now consider some $r > r_0$. Since $G^*$ is below $G$, we have

$$G^* r) \leq G(r).$$

Subtract $G(r_0) = G^*(r_0)$ from both sides and divide by $F(r) - F(r_0)$ ($F(r_0) - F(r) > 0$ since the density is positive everywhere) to get

$$\frac{G^*(r) - G^*(r_0)}{F(r) - F(r_0)} \leq \frac{G(r) - G(r_0)}{F(r) - F(r_0)}.$$

Taking limits of $r \to r_0^+$,

$$q = f^*(r_0) = \lim_{r \to r_0^+} \frac{G^*(r) - G^*(r_0)}{F(r) - F(r_0)} \leq \lim_{r \to r_0^+} \frac{G(r) - G(r_0)}{F(r) - F(r_0)}.$$

Next, consider some $r < r_0$. Since $G^*$ is below $G$, we have

$$G^*(r) \leq G(r).$$

Subtract $G(r_0) = G^*(r_0)$ from both sides and divide by $F(r_0) - F(r)$ ($F(r_0) - F(r) > 0$ since the density is positive everywhere and $r < r_0$) to get

$$\begin{aligned} \frac{G^*(r) - G^*(r_0)}{F(r_0) - F(r)} &\leq \frac{G(r) - G(r_0)}{F(r_0) - F(r)} \\ \implies \frac{G(r) - G(r_0)}{F(r) - F(r_0)} &\leq \frac{G^*(r) - G^*(r_0)}{F(r) - F(r_0)}. \end{aligned}$$

Taking limits of $r \to r_0^-$,

$$\lim_{r \to r_0^-} \frac{G(r) - G(r_0)}{F(r) - F(r_0)} \leq \lim_{r \to r_0^-} \frac{G^*(r) - G^*(r_0)}{F(r) - F(r_0)}.$$

Since $C^*$ is a convex curve, its derivative is non decreasing. In particular, the left derivative must be at most the right derivative at any point. That is,

$$\lim_{r \to r_0^-} \frac{G^*(r) - G^*(r_0)}{F(r) - F(r_0)} \leq \lim_{r \to r_0^+} \frac{G^*(r) - G^*(r_0)}{F(r) - F(r_0)}.$$

Putting it all together,

$$\begin{aligned} \lim_{r \to r_0^-} \frac{G(r)) - G(r_0)}{F(r) - F(r_0)} &\leq \lim_{r \to r_0^-} \frac{G^*(r) - G^*(r_0)}{F(r) - F(r_0)} \\ &\leq \lim_{r \to r_0^+} \frac{G^*(r) - G^*(r_0)}{F(r) - F(r_0)} = f^*(r_0) = q \\ &\leq \lim_{r \to r_0^+} \frac{G(r) - G(r_0)}{F(r) - F(r_0)}. \end{aligned}$$

Next, note that

$$\begin{aligned} \lim_{r \to r_0^-} \frac{G(r) - G(r_0)}{F(r) - F(r_0)} &= \lim_{r \to r_0^-} \frac{(1 - S_{0,r}(t \mid r))p(r)}{p(r)} && \text{(L'Hôpital's Rule)} \\ &= 1 - S_{0,r}(t \mid r_0^-). \end{aligned}$$

Similarly,

$$\lim_{r \to r_0^+} \frac{G(r) - G(r_0)}{F(r) - F(r_0)} = 1 - S_{0,r}(t \mid r_0^+).$$

Therefore,

$$1 - S_{0,r}(t \mid r_0^-) \leq f^*(r_0) = q \leq 1 - S_{0,r}(t \mid r_0^+).$$

By assumption S3, $S_{0,r}(t \mid r)$ is continuous at $r_0$, so $1 - S_{0,r}(t \mid r_0^-) = 1 - S_{0,r}(t \mid r_0^+)$ and hence $f^*(r_0) = q = 1 - S_{0,r}(t \mid r_0)$.

Next, note that by assumption S2, $r$ has a strictly positive density and hence the event $f^*(r) = q \iff r = r_0$ happens with probability zero. Since conditional probabilities on zero probability events are not uniquely defined, we write

$$\mathbb{E}[1 - S_{0,r}(t \mid r) | f^*(r) = q] = \mathbb{E}[1 - S_{0,r}(t \mid r) | r = r_0] = 1 - S_{0,r}(t \mid r_0) = q$$

almost surely.

**Combining Cases** Together, from Case I and II, we have shown that

$$\mathbb{E}[1 - S_{0,r}(t \mid r) | f^*(r) = q] = q,$$

for almost all $q$ that are in the range of $f^*$.

Since $f^* = 1 - S^*$, we can define $p = 1 - q$ to get

$$\mathbb{E}[S_{0,r}(t \mid r) | S^*(r) = p] = p.$$

for almost all $p$ in the range of $S^*$.

Now, from the proof of Theorem 2, from Sec. A.2.1 we have that the isotonic regression fit at any specific time $t$ converges to $S^*$ in the $L_2(P)$ norm.

We can proceed exactly as in Sec. A.2.2 to show that this convergence is also pointwise thanks to monotonicity of $S^*$ and assumptions S2, S3.

Finally from Sec. A.2.4, we see that the $\widehat{S}^{\text{DR}}_{\text{ISR}}$ at any time $t$ asymptotically equals the $S^*$ estimate at that time.

Combining, we have that the pointwise limit of the DR-ISR estimator $\widehat{S}^{\infty}_{\text{DR-ISR}}(t, x) = \lim_{n \to \infty} \widehat{S}^{\text{DR}}_{\text{ISR}}(t, x)$ is simply

$$\widehat{S}^{\infty}_{\text{DR-ISR}}(t, x) = S^*_t(r(X)),$$

where,

$$S^*_t = \arg\min_{g \in \mathcal{M}} \int (S_{0,r}(t \mid r) - g(r))^2 p(r) dr.$$

Therefore,

$$\mathbb{E}\left[ S_{0,r}(t \mid r(X)) \;\middle|\; \widehat{S}^{\infty}_{\text{DR-ISR}}(t, X) = p \right] = p.$$

for almost all $p$ in the range of $\widehat{S}^{\infty}_{\text{DR-ISR}}(t, X)$.

Recall that $S_{0,r}(t \mid r) = \mathbb{P}(T > t | r(X) = r) = \mathbb{E}[\mathbf{1}\{T > t\} | r(X) = r]$.

Substituting in the equation above and applying tower rule, we have that at any time $t \in \mathcal{T}$,

$$\mathbb{E}\left[ \mathbf{1}\{T > t\} \;\middle|\; \widehat{S}^{\infty}_{\text{DR-ISR}}(t, X) = p \right] = p.$$

for almost all $p$ in the range of $\widehat{S}^{\infty}_{\text{DR-ISR}}(t, X)$.

Thus, we have the desired result

$$\mathbb{P}\big(T > t \;\big|\; \widehat{S}^{\infty}_{\text{DR-ISR}}(t, X) = p\big) = p, \qquad \text{for all } t \in \mathcal{T},$$

for almost all $p$ in the range of $\widehat{S}^{\infty}_{\text{DR-ISR}}$.

# B Detailed Related Works

## B.1 Conformal Prediction for Survival Analysis

Much of the recent work on statistically valid survival prediction has focused on *conformal prediction* (Candès et al., 2023; Gui et al., 2024; Sesia and Svetnik, 2025b; Qin et al., 2025; Yi et al., 2025; Si and Qiu, 2025; Davidov et al., 2025; Holmes and Marandon, 2024; Sesia and Svetnik, 2025b; Farina et al., 2026). Rather than calibration, the typical goal in this setting is to construct a valid *lower prediction bound* (LPB) for the event time, i.e., a time beyond which an individual is expected to survive with probability at least $1 - \alpha$.

One of the first works to combine conformal inference with survival analysis constructs LPBs under *Type-I censoring*, where censoring times are observed for all individuals (Candès et al., 2023). The method first subsets the data to retain individuals whose censoring times are sufficiently large, and then applies weighted conformal inference to correct for the distribution shift induced by this subsetting step. Subsequent work has refined this idea. For example, Gui et al. (2024) introduces a covariate-dependent, data-adaptive subsetting rule that can produce less conservative bounds. However, both approaches focus on the Type-I censoring setting, which is restrictive in many survival-analysis applications because censoring times are not generally observed for individuals who experience the event.

Other work has adapted conformal survival analysis to the standard right-censoring setting, where censoring times are not observed for individuals who experience the event. Sesia and Svetnik (2025a) address this difficulty by first imputing the missing censoring times using an estimated censoring model. This reduces the problem to one in which Type-I-censoring methods can be applied; they then combine the resulting conformal lower bound with a survival-model-based bound. In a different direction, Davidov et al. (2025) avoid imputing censoring times and instead upper bound the true miscoverage rate using an IPCW estimator, then choose the smallest quantile level satisfying the desired miscoverage constraint. Relatedly, Farina et al. (2026) use doubly robust estimators based on semiparametric efficient influence functions to estimate marginal quantiles of the event-time distribution. Although their estimator has a form similar to our pseudo-label point estimate, we target the survival probability directly instead.

## B.2 Post hoc calibration of survival distributions

The methods most closely related to ours are CSD (Qi et al., 2024b) and CiPOT (Qi et al., 2024a).

CSD calibrates survival curves by discretizing them at multiple percentile levels and applying conformal quantile regression at each level. As noted by the authors, this preserves Harrell's C-index but may cause the predicted survival curves to be non-monotonic and hence invalid. On the other hand, if monotonicity is explicitly enforced, then the C-index may not be preserved. In terms of calibration guarantees, we note that asymptotically, CSD offers neither marginal calibration or consistency (i.e. exact conditional calibration) when the data is right censored.

CiPOT extends this line of work by directly mapping the empirical distribution of PIT values to the uniform scale. To account for censoring, CiPOT directly samples on the PIT scale. CiPOT does not preserve Harrell's C-index, but it does preserve the time-dependent Antolini's C-index (Antolini et al., 2005). In terms of calibration guarantees, CiPOT guarantees consistency (i.e. asymptotic conditional calibration) of the calibrated survival curves if the uncalibrated predictions are also consistent. This is similar to, but weaker than our consistency guarantee in Theorem 2. Roughly, our guarantee holds under consistency of either the uncalibrated survival or censoring mechanism while CiPOT's holds only under consistency of the survival mechanism. Note that in many cases, censoring mechanism may be easier to estimate (for example if the censoring is independent of the covariates) or may be exactly known (for example if censoring happens only due to termination of the clinical study). Finally, CiPOT's marginal calibration guarantee (Theorem 3.1 of Qi et al. (2024a)) also implicitly depends on the uncalibrated survival prediction being consistent. Specifically, when an instance is censored, CiPOT imputes the true PIT value as being uniform between the PIT value calculated at the censored time and 1. Importantly, this is only true if the uncalibrated survival function is already correct. As such, if the uncalibrated survival function is incorrect, CiPOT offers no calibration guarantees, unlike Theorem 3 for our method.

Concurrent to our work, Bladt et al. (2026) also propose a fundamentally different method for obtaining calibrated survival curves with rigorous theoretical guarantees. Roughly, their method assumes repeated covariates (binning when necessary), then constructs Kaplan-Meier estimates (Kaplan and Meier, 1958) over each bin, and finally applies a isotonic regression adjustment to obtain final survival curves. They offer strong theoretical guarantees, including a minimax optimality result for their exact setting. Note however, in our setting, applying their method would necessitate binning the risks from the Deep Cox models.

Recently, there has been an increased interest on addressing the calibration of survival predictions when using deep learning methods. Ashhad et al. (2026) proposes a new model combining linear AFT models with deep networks to allow the decoupling of discriminative performance and calibration for survival analysis. Liu et al. (2025) proposes a meta learning algorithm to augment the predictions of any model for survival analysis with scalar uncertainty scores for each instance.

### B.3 Extension of Isotonic Distributional Regression

An important extension of Isotonic Distributional Regression is the Distributional Single-Index Model (Henzi et al., 2021a) which first uses an index model to order the covariate space, and then applies Isotonic Distributional Regression with respect to the learned index. We likewise learn an ordering, which is induced by the Deep Cox risk score, and then adapt the Distributional Single-Index Model framework to right-censored survival data with two key extensions. First, we account for censoring using doubly robust pseudo-outcomes. Second, because the calibrated survival function must be monotone in both risk and time, we use a two-dimensional isotonic regression procedure to enforce both constraints which we implement efficiently using Dykstra's algorithm with PAVA.

# C ADDITIONAL ESTIMATORS

We now present the motivation behind the RW-ISR and DR-ISR estimators as developed in Sec 3 in full generality. This allows us to identify three new estimators RW$^+$-ISR, HT-ISR and HT$^+$-ISR.

We have considered two broad approaches to calibration under censoring. The first uses constrained regression on the uncensored binary survival indicators in the calibration data. In this case, censoring is handled by weighting the optimization objective to account for unequal probabilities of observing $\mathbf{1}\{T_i > t\} = Z_i(t)$. The second method uses a pseudo-outcome approach, which first constructs pointwise estimates of the conditional survival probability on the calibration set. These estimates are then projected onto a family of survival functions that are monotone in both risk and time.

Because these two approaches incorporate censoring in fundamentally different ways, we develop the resulting estimators separately.

## C.1 WEIGHTED ISOTONIC REGRESSION USING SURVIVAL INDICATORS

Under right-censoring, the survival indicators $Z_i(t)$ are only partially observed. To account for this, we use inverse probability of censoring weighting (IPCW), which reweights observations according to their probability of remaining uncensored.

Let $O_i = (X_i, Y_i, \delta_i)$ denote the observed data, and as before define the censoring analog of the survival function $G(t \mid x) = \mathbb{P}(C > t \mid X = x)$. We also define the IPCW weight $w(O) = \delta / G(Y \mid X)$. For a fixed time $t$, consider the weighted regression problem

$$f_t = \underset{f:\text{non-increasing}}{\arg\min}\ \mathbb{E}\left[w(O)(f(r(X)) - \mathbf{1}[Y > t])^2\right].$$

Note that the population minimizer is the well-known Hajek estimator of the survival probability.

$$f_t(r(x)) = \frac{\mathbb{E}\left[w(O)\,\mathbf{1}[Y > t] \mid r(X) = r(x)\right]}{\mathbb{E}[w(O) \mid r(X) = r(x)]} = \mathbb{P}(T > t \mid r(X) = r(x)) = S(t \mid x).$$

There are two additional practical considerations. First, neither the risk score $r$ nor the censoring function $G$ is known, so both must be learned on the training data, for example using a Cox model. We then treat these estimates as fixed and solve the corresponding weighted optimization problem on the independent calibration set. Second, we impose a monotonicity constraint in the risk score so that the resulting estimator defines a valid survival curve. The calibration step solves,

$$\widehat{S}^{\text{RW}}_{\text{ISR}}(t \mid x) = f_t(\widehat{r}(x)), \quad \text{where } f_t = \underset{f:\text{non-increasing}}{\arg\min} \sum_{O_i \in \mathcal{D}_{\text{cal}}} \frac{\delta_i (f(\widehat{r}(X_i)) - \mathbf{1}[Y_i > t])^2}{\widehat{G}(Y_i \mid X_i)}. \tag{11}$$

The RW-ISR estimator does not make full use of the available information, since the survival indicator is known not only when $\delta = 1$, but also for censored individuals with $Y > t$. We can modify the IPCW to account for all samples for which $Z_i(t)$ is observed. For a fixed time $t$, consider the weighted regression problem,

$$f_t^+ = \underset{f:\text{non-increasing}}{\arg\min}\ \mathbb{E}\left[w^+(O,t)(f(r(X)) - \mathbf{1}[Y > t])^2\right], \text{ for } w^+(O,t) = \frac{\mathbf{1}[Y > t]}{G(t \mid X)} + \frac{\delta \mathbf{1}[Y \le t]}{G(Y \mid X)},$$

where we now explicitly note the dependence of the weights on the time since they are no longer constant across time.

The survival probability remains the population minimizer.

$$f_t^+(r(x)) = \frac{\mathbb{E}\left[w^+(O,t)\,\mathbf{1}[Y > t] \mid r(X) = r(x)\right]}{\mathbb{E}[w^+(O,t) \mid r(X) = r(x)]} = \mathbb{P}(T > t \mid r(X) = r(x)) = S(t \mid x).$$

Since this estimator improves upon the RW-ISR estimator, we refer to it as RW$^+$-ISR.

Note, however, the weights in the objective are no longer constant across time. As such, in finite samples, the isotonic regression solutions enforcing monotonicity in risk independently at each time are no longer guaranteed to be non-increasing in time. This necessitates explicitly enforcing

non-increasingness in time and solving a 2D isotonic regression problem (for example using Dykstra's algorithm with PAVA) as for the DR-ISR estimator discussed in Sec. 3. Specifically,

$$\widehat{S}^{\text{RW+}}_{\text{ISR}}(t \mid x) = f_t^+(\widehat{r}(x)), \quad \text{where } f_t^+ = \operatorname*{arg\,min}_{f \in \mathcal{F}_\leq} \sum_{t \in \mathcal{T}} \sum_{O_i \in \mathcal{D}_{\text{cal}}} w^+(O_i, t)(f(\widehat{r}(X_i)) - \mathbf{1}[Y_i > t])^2. \tag{12}$$

Intuitively, both RW-ISR and RW$^+$-ISR use the censoring model to determine how to weight the calibration data in order to correct for missing survival indicators induced by censoring.

### C.2 Isotonic projection of pseudo-outcomes

We now describe our second calibration strategy as used by DR-ISR. Rather than incorporating censoring directly into the calibration objective, this approach first constructs pointwise pseudo-outcomes for the conditional survival probability on the calibration set.

For each $O_i \in \mathcal{D}_{\text{cal}}$ and $t \in \mathcal{T}$, let $\tilde{S}(t \mid O_i; \eta)$ denote a pseudo-outcome, where $\eta$ denotes the nuisance parameters. Because these pseudo-outcomes are constructed pointwise, they may not satisfy the structural properties of a valid survival function. We therefore project them onto a class of functions that is monotone in both risk and time. Consider the optimization problem

$$f = \operatorname*{arg\,min}_{f \in \mathcal{F}_\leq} \sum_{t \in \mathcal{T}} \mathbb{E}\left[\left(f(t, r(X)) - \tilde{S}(t \mid O; \eta)\right)^2\right].$$

For conditionally unbiased pseudo-outcomes in the sense that $\mathbb{E}[\tilde{S}(t \mid O; \eta) \mid X = x] = S(t \mid x)$, the population minimizer agrees with the true survival function on the grid $\mathcal{T}$.

In practice, the risk scores and nuisance quantities are first estimated on the training data. In the calibration step, we use the plug-in nuisance estimate $\widehat{\eta}$ and replace the population objective by its empirical analogue on the calibration set. The calibration step solves

$$\widehat{S}(t \mid x) = f(t, \widehat{r}(x)), \quad \text{where } f = \operatorname*{arg\,min}_{f \in \mathcal{F}_\leq} \sum_{t \in \mathcal{T}} \sum_{X_i \in \mathcal{D}_{\text{cal}}} \left[\left(f(t, \widehat{r}(X_i)) - \tilde{S}(t \mid O_i; \widehat{\eta})\right)^2\right]. \tag{13}$$

*Remark.* Survival models such as the Cox model already produce structurally valid survival curves, but their survival probabilities are often not well calibrated in practice. The pseudo-outcomes provide pointwise estimates that may be statistically more efficient than the initial survival model, while the projection step restores the structural constraints required of a valid survival function.

We next describe two ways to construct conditionally unbiased pseudo-outcomes using IPCW and Augmented IPCW (AIPCW).

**Horvitz–Thompson.** The first method uses IPCW in the same spirit as the RW-ISR estimator. For each $t \in \mathcal{T}$ and $O_i \in \mathcal{D}_{\text{cal}}$, we define the pointwise pseudo-outcomes

$$\tilde{S}^{\text{HT}}_{\text{ISR}}(t \mid O_i; \eta) = \frac{\delta_i \mathbf{1}[Y_i > t]}{G(Y_i \mid X_i)}, \qquad \tilde{S}^{\text{HT+}}_{\text{ISR}}(t \mid O_i; \eta) = \frac{\mathbf{1}[Y_i > t]}{G(t \mid X_i)}, \qquad \text{where } \eta = G.$$

The corresponding distributional estimators in Equation 13 are denoted $\widehat{S}_{\text{HT}}$ and $\widehat{S}_{\text{HT+}}$. Because the pseudo-outcomes take the well-known Horvitz–Thompson (HT) form, we refer to them as the HT-ISR and HT$^+$-ISR distributional estimators, respectively. As in the RW-ISR setting, HT$^+$-ISR improves on the initial construction by incorporating censored observations for which the survival indicator is still known. If multiple calibration points share the same covariate value (e.g. if the covariates are discrete), we define a separate pseudo-outcome for each observation.

**Doubly robust.** The RW and Horvitz–Thompson constructions above rely only on inverse probability of censoring weighting, and therefore use only the censoring model to correct for missing survival indicators. A different strategy is to combine the censoring model with an outcome model for the survival function itself. This leads to the augmented IPCW (AIPCW) construction.

$$\tilde{S}^{\text{DR}}_{\text{ISR}}(t \mid O_i; \eta) = S(t \mid X_i) - \int_0^t \frac{S(t \mid X_i)}{S(u \mid X_i) G(u \mid X_i)} \, dM_{T|X}(u \mid X_i), \qquad \text{where } \eta = (S, G).$$

Here, $dM_{T|X}(u \mid x) = dN_T(u) - \mathbf{1}[Y \geq u]\, d\Lambda_{T|X}(u \mid x)$ denotes the event martingale difference process, where $dN_T(u) = \mathbf{1}[Y \in du \wedge \delta = 1]$ is the event increment counting process and $\Lambda_{T|X}(u \mid x) = -\log S(u \mid x)$ is the cumulative hazard function as described in Sec. 3

The AIPCW construction has an attractive theoretical advantage over IPCW: it is conditionally unbiased if either the survival model or the censoring model is correctly specified.

# D ALGORITHMIC IMPLEMENTATION

## D.1 DR-ISR ALGORITHM

For completeness, we present an expanded version of algorithm 1 to fit and make predictions from the DR-ISR estimator. Both algorithm 1 and algorithm 2 contain the exact same operations but algorithm 2 improves legibility by spelling out the operations more clearly. In practice, we also clip the probabilities returned by $\widehat{S}, \widehat{G}$ at some small $\epsilon$ to ensure stability. Finally, to construct the time grid, we use 10000 evenly spaced points between $t = 0$ and the maximum observed time in the calibration dataset. We additionally augment this time grid with all the observed times in the training and calibration datasets.

**Algorithm 2** Construction of DR-ISR calibrated survival curve estimates

**Require:** Survival $\widehat{S}$, censoring $\widehat{G}$, and risk $r_\theta$ models from an initial Deep Cox model
**Require:** Calibration dataset $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i, \delta_i)\}_{i=1}^{n_{\text{cal}}}$ sorted by increasing risk
**Require:** Time grid $\mathcal{T} = \{t_1, \ldots, t_K\}$

▷ Doubly robust point estimation

**for** $j \in [n_{\text{cal}}]$ **do**
  $\widehat{S}(t_0 \mid X_j) \leftarrow 1$
  **for** $i \in [K]$ **do**
    $d\Lambda_{T|X}(t_i \mid X_j) \leftarrow 1 - \widehat{S}(t_i \mid X_j)/\widehat{S}(t_{i-1} \mid X_j)$
    $dM_{T|X}(t_i \mid X_j) \leftarrow \mathbf{1}[Y_j = t_i, \delta_j = 1] - \mathbf{1}[Y_j \geq t_i]\, d\Lambda_{T|X}(t_i \mid X_j)$
    $\tilde{S}^{\text{DR}}[i, j] \leftarrow \widehat{S}(t_i \mid X_j) - \sum_{k=1}^{i} \frac{\widehat{S}(t_i|X_j)}{\widehat{S}(t_k|X_j)\widehat{G}(t_k|X_j)} dM_{T|X}(t_k \mid X_j)$ ▷ Eq. 2
  **end for**
**end for**

▷ Dykstra's algorithm initialization

$S \leftarrow \tilde{S}_{\text{DR}}$
$\epsilon_{\text{risk}} \leftarrow \mathbf{0}^{|\mathcal{T}| \times n_{\text{cal}}}$
$\epsilon_{\text{time}} \leftarrow \mathbf{0}^{|\mathcal{T}| \times n_{\text{cal}}}$
**while** not converged **do**

▷ Monotonicity in risk

  $U \leftarrow S + \epsilon_{\text{time}}$
  **for** $i \in [|\mathcal{T}|]$ **do**
    $S[i, :] \leftarrow \texttt{PAVA}(U[i, :],\ \text{decreasing} = \text{True})$
  **end for**
  $\epsilon_{\text{time}} \leftarrow U - S$

▷ Monotonicity in time

  $V \leftarrow S + \epsilon_{\text{risk}}$
  **for** $j \in [n_{\text{cal}}]$ **do**
    $S[:, j] \leftarrow \texttt{PAVA}(V[:, j],\ \text{decreasing} = \text{True})$
  **end for**
  $\epsilon_{\text{risk}} \leftarrow V - S$
**end while**

▷ Prediction

$S \leftarrow \text{Clamp}(S, \text{min} = 0, \text{max} = 1)$ ▷ Clip probabilities to $[0, 1]$
$\hat{f} \leftarrow \text{Interpolate}(S)$ ▷ Linearly interpolate to form 2D function
$\widehat{S}^{\text{DR}}_{\text{ISR}}(t \mid X_{\text{test}}) \leftarrow \hat{f}(t, r_\theta(X_{\text{test}}))$

### D.2 RW-ISR ESTIMATOR

We present the pseudocode for construction of the RW-ISR estimator below in algorithm 3. As in the DR-ISR case, we also clip the probabilities returned by $\widehat{G}$ at some small $\epsilon$ to ensure stability. To construct the time grid, we use 10000 evenly spaced points between $t = 0$ and the maximum observed time in the calibration dataset. We additionally augment this time grid with all the observed times in the training and calibration datasets.

**Algorithm 3** Construction of RW-ISR calibrated survival curve estimates

**Require:** Survival $\widehat{S}$, censoring $\widehat{G}$, and risk $r_\theta$ models from an initial Deep Cox model
**Require:** Calibration dataset $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i, \delta_i)\}_{i=1}^{n_{\text{cal}}}$ sorted by increasing risk
**Require:** Time grid $\mathcal{T} = \{t_1, \ldots, t_K\}$

▷ Weight estimation

**for** $j \in [n_{\text{cal}}]$ **do**
  **if** $\delta_j = 1$ **then**
    $W[j] \leftarrow \frac{1}{\widehat{G}(Y_j | X_j)}$
  **else**
    $W[j] \leftarrow 0$ ▷ Assign zero weight to censored patients
  **end if**
**end for**

▷ Perform Isotonic Regression at each time

**for** $i \in [K]$ **do**
  **for** $j \in [n_{\text{cal}}]$ **do** ▷ Construct Binary Outcomes
    $Z[j] \leftarrow \mathbf{1}\{Y_j > t_i\}$
  **end for**
  $S[i, :] \leftarrow \texttt{PAVA}(Z,\ \text{weights=}W,\ \text{decreasing=True})$
**end for**

▷ Prediction

$\hat{f} \leftarrow \text{Interpolate}(S)$ ▷ Linearly interpolate to form 2D function
$\widehat{S}^{\text{RW}}_{\text{ISR}}(t \mid X_{\text{test}}) \leftarrow \hat{f}(t, r_\theta(X_{\text{test}}))$

# E EXPERIMENT DETAILS

We elaborate on the three metrics used in the experiments. For AUPIT, note that it is equivalent to the Wasserstein-1 distance between the predicted PIT distribution and the ideal uniform distribution.

## E.1 QUANTILE LOSS DETAILS

We first derive the form of the quantile loss under censoring. Recall the standard pinball loss for a quantile prediction $q$ at level $\tau$ in the absence of censoring is

$$\rho_\tau(T, q) = (1-\tau)(q-T)\mathbf{1}\{T \le q\} + \tau(T-q)\mathbf{1}\{T > q\}.$$

Since survival times are nonnegative, each term can be written as an integral over threshold indicators.

$$\rho_\tau(T, q) = (1-\tau)\int_0^q \mathbf{1}[T \le t]\, dt + \tau \int_q^\infty \mathbf{1}[T > t]\, dt.$$

Under conditional independent censoring, we can estimate the threshold indicators using IPCW.

$$\mathbf{1}[T \le t] \rightsquigarrow \frac{\delta \mathbf{1}[Y \le t]}{\widehat{G}(Y \mid X)}, \qquad \mathbf{1}[T > t] \rightsquigarrow \frac{\mathbf{1}[Y > t]}{\widehat{G}(t \mid X)}.$$

In practice, we also truncate the second integral at the maximum observed time $t_{\max}$. Substituting these terms into the integral representation and using the predicted quantile $\hat{q}_\tau(X_i)$ gives the censored quantile score

$$\mathrm{QS}(\tau) = \frac{1}{|\mathcal{I}_{\text{test}}|} \sum_{i \in \mathcal{I}_{\text{test}}} \left[ (1-\tau) \int_0^{\hat{q}_\tau(X_i)} \frac{\delta_i \mathbf{1}[Y_i \le t]}{\widehat{G}(Y_i \mid X_i)}\, dt + \tau \int_{\hat{q}_\tau(X_i)}^{t_{\max}} \frac{\mathbf{1}[Y_i > t]}{\widehat{G}(t \mid X_i)}\, dt \right].$$

Note that if $\widehat{S}(t \mid X_i)$ does not reach the $1-\tau$ level by $t_{\max}$, then $\widehat{q}_\tau(X_i)$ is ill-defined and hence the $\mathrm{QS}(\tau)$ cannot be calculated.

This stems from a common problem in survival analysis where the survival curves may not reach 0. Intuitively, for a variety of time-to-event outcomes, the event may never happen and hence even the true survival curve may never reach 0. Additionally, when estimating the survival curve, we may have no information regarding how the survival probabilities behave after a maximum time $t_{\max}$. For example, for the Kaplan-Meier (Kaplan and Meier, 1958) estimator, if the largest observed times correspond to censored instances, the estimated survival curve will never reach 0. Cox models, when combined with the Breslow estimator, also suffer from this issue.

When absolutely necessary to obtain survival curve estimates that go to zero (for example to calculate a mean predicted survival time), a variety of extrapolation methods can be applied such as fitting a parametric tail beyond the maximum time or using external data. Unfortunately, however, there is no principled and well-accepted way to extrapolate. See, for example, Latimer (2013) which presents an official report made to the government of the United Kingdom reviewing suitability of different extrapolation techniques when making decisions regarding efficacy of medical treatments. Ultimately, the appropriate extrapolation method depends on the problem context and downstream analysis.

For our estimators too, predicted survival curves may not reach 0 and are constant beyond the maximum observed time in the calibration dataset. Thus, $\widehat{q}_\tau(X_i)$ can often be ill-defined at higher quantiles. When evaluating the quantile score, we simply exclude such instances from evaluation. In particular, if the quantile score is ill-defined for a test point for even one of the methods under comparison, we exclude that instance from evaluation for all methods to allow for fair comparison. Figure 5 reports the proportion of patients that the quantile losses were evaluated over for the real data pathology and RNA models.

## E.2 INTEGRATED BRIER SCORE DETAILS

We discuss the properties of the IBS and then derive its connection to the integrated quantile loss. We will first find it useful to relate the IBS to the continuous ranked probability score (CRPS), which

measures the calibration and sharpness of a probabilistic forecast in the uncensored setting (Hersbach, 2000). For a predictive CDF $F$ of a generic response $Y$, the CRPS is

$$\mathrm{CRPS}(F, Y) = \int_0^\infty \left(F(z) - \mathbf{1}[Y \le z]\right)^2 \, dz.$$

The CRPS is a proper scoring rule, in the sense that its expectation is minimized by the true CDF. In the context of survival analysis, the estimated CDF is $\widehat{F}(t \mid X) = 1 - \widehat{S}(t \mid X)$. Thus, we can express the scoring rule as

$$\mathrm{CRPS}(S, T) = \int_0^\infty \left(\mathbf{1}[T \le t]\widehat{S}(t \mid X)^2 + \mathbf{1}[T > t]\big(1 - \widehat{S}(t \mid X)\big)^2\right) \, dt.$$

Thus, in the absence of censoring, the CRPS is the integrated Brier score over time. Under right censoring, the threshold indicators can be estimated using IPCW.

$$\mathbf{1}[T \le t] \rightsquigarrow \frac{\delta \mathbf{1}[Y \le t]}{\widehat{G}(Y \mid X)}, \qquad \mathbf{1}[T > t] \rightsquigarrow \frac{\mathbf{1}[Y > t]}{\widehat{G}(t \mid X)}.$$

Substituting these IPCW terms gives,

$$\int_0^\infty \left[\frac{\delta_i \mathbf{1}[Y_i \le t]\widehat{S}(t \mid X_i)^2}{\widehat{G}(Y_i \mid X_i)} + \frac{\mathbf{1}[Y_i > t](1 - \widehat{S}(t \mid X_i))^2}{\widehat{G}(t \mid X_i)}\right] dt.$$

Truncating the integral at the maximum observed time $t_{\max}$ and normalizing recovers the IBS.

$$\mathrm{IBS} = \frac{1}{|\mathcal{I}_{\mathrm{test}}|} \sum_{i \in \mathcal{I}_{\mathrm{test}}} \frac{1}{t_{\max}} \int_0^{t_{\max}} \left[\frac{\delta_i \mathbf{1}[Y_i \le t]\widehat{S}(t \mid X_i)^2}{\widehat{G}(Y_i \mid X_i)} + \frac{\mathbf{1}[Y_i > t](1 - \widehat{S}(t \mid X_i))^2}{\widehat{G}(t \mid X_i)}\right] dt.$$

The connection between the CRPS and IBS highlights several shared properties. In particular, it makes clear that IBS assesses both calibration and sharpness of the predicted survival distribution, as noted in the literature (DeGroot and Fienberg, 1983; Qi et al., 2024b).

In the uncensored setting, the CRPS is also known to admit an equivalent representation as an integral of quantile loss over all quantile levels (Gneiting and Ranjan, 2011). We next show that an analogous identity holds for the IBS in the right-censored setting if $\hat{q}_\tau(X_i) \le t_{\max}$ for all $\tau \in [0, 1]$ and all $i \in \mathcal{I}_{\mathrm{test}}$. Integrating over $\tau$ in the quantile loss and exchanging the order of integration gives

$$\int_0^1 \mathrm{QS}(\tau) \, d\tau = \frac{1}{|\mathcal{I}_{\mathrm{test}}|} \sum_{i \in \mathcal{I}_{\mathrm{test}}} \int_0^{t_{\max}} \left[\frac{\delta_i \mathbf{1}[Y_i \le t]}{\widehat{G}(Y_i \mid X_i)} A_i(t) + \frac{\mathbf{1}[Y_i > t]}{\widehat{G}(t \mid X_i)} B_i(t)\right] dt,$$

where

$$A_i(t) = \int_0^1 (1 - \tau)\mathbf{1}[t \le \hat{q}_\tau(X_i)] \, d\tau, \qquad B_i(t) = \int_0^1 \tau \mathbf{1}[\hat{q}_\tau(X_i) \le t] \, d\tau.$$

Since $t \le \hat{q}_\tau(X_i)$ is equivalent to $\tau \ge \widehat{F}(t \mid X_i)$, we have

$$A_i(t) = \int_{\widehat{F}(t|X_i)}^1 (1 - \tau) \, d\tau = \frac{1}{2}\left(1 - \widehat{F}(t \mid X_i)\right)^2 = \frac{1}{2}\widehat{S}(t \mid X_i)^2.$$

Similarly, since $\hat{q}_\tau(X_i) \le t$ is equivalent to $\tau \le \widehat{F}(t \mid X_i)$,

$$B_i(t) = \int_0^{\widehat{F}(t|X_i)} \tau \, d\tau = \frac{1}{2}\widehat{F}(t \mid X_i)^2 = \frac{1}{2}\left(1 - \widehat{S}(t \mid X_i)\right)^2.$$

Substituting these two identities yields

$$\int_0^1 \mathrm{QS}(\tau) \, d\tau = \frac{1}{2|\mathcal{I}_{\mathrm{test}}|} \sum_{i \in \mathcal{I}_{\mathrm{test}}} \int_0^{t_{\max}} \left[\frac{\delta_i \mathbf{1}[Y_i \le t]\widehat{S}(t \mid X_i)^2}{\widehat{G}(Y_i \mid X_i)} + \frac{\mathbf{1}[Y_i > t]\big(1 - \widehat{S}(t \mid X_i)\big)^2}{\widehat{G}(t \mid X_i)}\right] dt.$$

Comparing this expression with the definition of IBS gives

$$\mathrm{IBS} = \frac{2}{t_{\max}} \int_0^1 \mathrm{QS}(\tau) \, d\tau.$$

Thus, IBS is the integrated quantile score over all quantile levels, up to scaling.

### E.3 COMPUTE RESOURCES

All experiments were conducted using on an Ubuntu 22.04.5 Linux cluster. 10 Intel(R) Xeon(R) E5-2670 v2 @ 2.50GHz CPUs were available along with two NVIDIA Tesla V100 GPUs. 200TB of long term storage was available and 300GB of RAM.

Our simulation experiments need only about 2GB of RAM and testing all methods for a single seed for a single setting takes about 2 minutes on a single CPU.

For the real data experiments, training the Deep Cox models for single seed for a single experiment takes under 2 minutes on a single GPU with no more than 30GB of combined GPU memory and CPU RAM needed. Running all calibration procedures for a single seed takes no more than 2 minutes on a single CPU.

To directly compare runtimes across methods, we vary the size of the calibration set and report the mean runtime in seconds (± 2 s.e.) for each method on synthetic setting 5 below.

Table 3: Calibration runtime in seconds (mean $\pm$ 2 standard error over five seeds)

| $n_{\text{cal}}$ | CSD | CiPOT | RW-ISR | DR-ISR |
|---|---|---|---|---|
| 100 | $0.028 \pm 0.001$ | $0.015 \pm 0.001$ | $0.002 \pm 0.001$ | $0.084 \pm 0.004$ |
| 250 | $0.080 \pm 0.001$ | $0.036 \pm 0.001$ | $0.004 \pm 0.001$ | $0.302 \pm 0.034$ |
| 500 | $0.177 \pm 0.001$ | $0.079 \pm 0.001$ | $0.006 \pm 0.001$ | $0.940 \pm 0.006$ |
| 1000 | $0.383 \pm 0.003$ | $0.191 \pm 0.001$ | $0.012 \pm 0.001$ | $2.049 \pm 0.009$ |
| 2500 | $1.154 \pm 0.006$ | $0.719 \pm 0.002$ | $0.038 \pm 0.002$ | $6.608 \pm 0.045$ |
| 5000 | $2.771 \pm 0.015$ | $2.225 \pm 0.007$ | $0.116 \pm 0.005$ | $19.499 \pm 0.076$ |
| 10000 | $7.141 \pm 0.020$ | $7.536 \pm 0.054$ | $0.900 \pm 0.021$ | $87.452 \pm 2.261$ |

As discussed in Section 3, the runtime of DR-ISR depends on the convergence of Dykstra's algorithm for the 2D isotonic regression. Thus, it scales less favorably than the other methods. Nonetheless, our procedure remains lightweight in practice. Even at $n_{\text{cal}} = 10,000$, calibration takes only about 1.5 minutes.

### E.4 SYNTHETIC DATA DETAILS

For completeness, we reproduce details from (Gui et al., 2024) on the data generating process for the synthetic data experiments in Sec. 6.1.

The datapoints $(X_i, Y_i, \delta_i)$ are i.i.d. for all 6 settings. We use 2500 points for each of the training, calibration datasets and 5000 points for the test datasets. For each setting, $P_X \sim \text{Unif}([0,4]^p)$, $P_{T|X} \sim \text{LogNormal}(\mu(x), \sigma^2(x))$, where $p$ is the covariate dimension.

Settings 1–2 are univariate with independent censoring, while Settings 3–4 involve covariate-dependent censoring. Settings 5-6 have high-dimensional covariates ($p = 10$) and covariate-dependent censoring. Table 4 summarizes the parameters for each setup, along with the censoring rates

Table 4: Data generating mechanism for synthetic data experiments. Reproduced from (Gui et al., 2024).

| Setting | $p$ | $\mu(x)$ | $\sigma(x)$ | $P_{C\|X}$ | Censoring Rate |
|---|---|---|---|---|---|
| 1 | 1 | $0.632x$ | 2 | $\text{Exp}(0.1)$ | 41.81% |
| 2 | 1 | $3 \cdot 1\{x > 2\} + x \cdot 1\{x \leq 2\}$ | 0.5 | $\text{Exp}(0.1)$ | 55.67% |
| 3 | 1 | $2 \cdot 1\{x > 2\} + x \cdot 1\{x \leq 2\}$ | 0.5 | $\text{Exp}\left(0.25 + \frac{6+x}{100}\right)$ | 73.26% |
| 4 | 1 | $3 \cdot 1\{x > 2\} + 1.5x \cdot 1\{x \leq 2\}$ | 0.5 | $\text{LogNormal}\left(2 + \frac{2-x}{50}, 0.5\right)$ | 62.63% |
| 5 | 10 | $0.126(x_1 + \sqrt{x_3 x_5}) + 1$ | 1 | $\text{Exp}\left(\frac{x_{10}}{10} + \frac{1}{20}\right)$ | 61.20% |
| 6 | 10 | $0.126(x_1 + \sqrt{x_3 x_5}) + 1$ | $\frac{x_2+2}{4}$ | $\text{Exp}\left(\frac{x_{10}}{10} + \frac{1}{20}\right)$ | 61.32% |

For each setting, we train a linear Cox model to estimate the risk score using the `lifelines` package. We use Breslow's estimate to obtain the probabilities needed for the survival and censoring functions $\widehat{S}$ and $\widehat{G}$ for our calibration algorithms. We clip the probabilities returned from the Cox model if they are below $10^{-4}$ for numerical stability. We repeat the training and calibration procedure for each setting $100$ times using different random seeds for data generation.

### E.5 Real-data foundation models

We use the TCGA (Weinstein et al., 2013) dataset which consists of cancer patients with various forms of data and clinical outcomes.

#### E.5.1 Pathology Reports Model

**Data:** For this experiment, we use text-based pathology reports as provided by (Kefeli and Tatonetti, 2024) to predict the progression free survival (PFS). The PFS times we use were calculated and provided by Liu et al. (2018). After selecting those patients which have both PFS times and pathology reports available, we are left with 9517 patients. For these selected patients, $59.55\%$ of the PFS outcomes were censored. We randomly split the data into $15\%$ for testing, $25\%$ for calibration, and $60\%$ for Deep Cox model training.

**Model Training:** We make use of the Bio + Clinical BERT Model from (Alsentzer et al., 2019) which was trained on all notes from MIMIC III (Johnson et al., 2016), a database containing electronic health records from ICU patients at the Beth Israel Hospital in Boston, MA. We extract 768 dimensional features by extracting the final token embedding from the last layer of the Bio + Clinical BERT Model. We then train a three layer MLP on top of these embeddings with 128 and 64 hidden units. We use ReLU activation and dropout at a rate of 0.1 between layers. We train the model to maximize the Cox partial likelihood as in (Katzman et al., 2018). We use Adam (Kingma and Ba, 2015) as an optimizer with a learning rate of 0.001 and betas=(0.9, 0.999) with a batch size of 256. We train for 25 epochs. We follow exactly this same procedure to train a model to predict the censoring time too. Over 100 seeds controlling the dataset splits and stochasticity in model training, we obtain a mean C-index of $0.682 \pm 0.002$ for the time to event Deep Cox model and $0.678 \pm 0.002$ for the censoring Deep Cox model ($\pm$ 2 standard error reported). We use Breslow's estimator (Breslow, 1972) on top of these Deep Cox models to obtain the probabilities needed for $\widehat{S}, \widehat{G}$ for our calibration algorithms. We clip the probabilities returned from the Deep Cox $\widehat{S}, \widehat{G}$ if they are below $0.02$ for numeric stability.

#### E.5.2 RNA-seq Model

**Data:** We again use the TCGA dataset. For this experiment, we use bulk-RNA samples as provided by (Vivian et al., 2017) to predict the PFS. The PFS times we use were calculated and provided by Liu et al. (2018). After selecting those patients which have both PFS times and RNA-seq samples available, we are left with 10272 patients. For these selected patients, $56.90\%$ of the PFS outcomes were censored. We randomly split the data into $15\%$ for testing, $25\%$ for calibration, and $60\%$ for Deep Cox model training.

**Model Training:** We make use of the BulkFormer 147M Model from (Kang et al., 2025). BulkFormer was pretrained on over 500,000 human bulk transcriptomic profiles. It incorporates a hybrid encoder architecture, combining a graph neural network to capture explicit gene-gene interactions. We extract 643 dimensional features by taking the mean of the final gene embeddings for all the genes that can be handled by BulkFormer and are present in the compiled data from (Vivian et al., 2017) as suggested by (Kang et al., 2025). We then train a three layer MLP on top of these embeddings with 128 and 64 hidden units. We use ReLU activation and dropout at a rate of 0.1 between layers. We train the model to maximize the Cox partial likelihood as in (Katzman et al., 2018). We use Adam (Kingma and Ba, 2015) as an optimizer with a learning rate of 0.001 and betas=(0.9, 0.999) with a batch size of 256. We train for 25 epochs. We follow exactly this same procedure to train a model to predict the censoring time too. Over 100 seeds controlling the dataset splits and stochasticity in model training, we obtain a mean C-index of $0.707 \pm 0.002$ for the time to event Deep Cox model and $0.578 \pm 0.002$ for the censoring Deep Cox model ($\pm$ 2 standard error reported). We use Breslow's

estimator (Breslow, 1972) on top of these Deep Cox models to obtain the probabilities needed for $\widehat{S}, \widehat{G}$ for our calibration algorithms. We clip the probabilities returned from the Deep Cox $\widehat{S}, \widehat{G}$ if they are below $0.02$ for numeric stability.

#### E.5.3 H&E MODEL

**Data:** We again use the TCGA dataset. For this experiment, we use H&E slides as provided by (Weinstein et al., 2013) to predict the PFS. The PFS times we use were calculated and provided by Liu et al. (2018). After selecting those patients which have both PFS times and H&E samples available, we are left with 6589 patients. When there are multiple H&E slides available for a patient, we only use the first one. For these selected patients, $55.27\%$ of the PFS outcomes were censored. We randomly split the data into $15\%$ for testing, $25\%$ for calibration, and $60\%$ for Deep Cox model training.

**Model Training:** We make use of the image-only SPARC model from (Ayed et al., 2026). It works by first splitting H&E slide into patches, extracting embeddings from each patch using the H-Optimus-1 (Bioptimus, 2025) foundation model and then using a linear attention mechanism to aggregate information across patches. SPARC was also trained on the TCGA data, with 5 sets of model weights available from training runs on 5 folds of the entire dataset. To avoid train-test leakage, for each slide, we extract features from the set of model weights for which that slide was in the test fold. We extract 256 dimensional features from the last layers of the model. We then train a three layer MLP on top of these embeddings with 128 and 64 hidden units. We use ReLU activation and dropout at a rate of 0.1 between layers. We train the model to maximize the Cox partial likelihood as in (Katzman et al., 2018). We use Adam (Kingma and Ba, 2015) as an optimizer with a learning rate of 0.001 and betas=(0.9, 0.999) with a batch size of 256. We train for 25 epochs. We follow exactly this same procedure to train a model to predict the censoring time too. Over 100 seeds controlling the dataset splits and stochasticity in model training, we obtain a mean C-index of $0.688 \pm 0.002$ for the time to event Deep Cox model and $0.640 \pm 0.002$ for the censoring Deep Cox model ($\pm$ 2 standard error reported). We use Breslow's estimator (Breslow, 1972) on top of these Deep Cox models to obtain the probabilities needed for $\widehat{S}, \widehat{G}$ for our calibration algorithms. We clip the probabilities returned from the Deep Cox $\widehat{S}, \widehat{G}$ if they are below $0.02$ for numeric stability.

#### E.5.4 CENSORING RSF MODELS

When evaluating performance on the real datasets, we report IPCW adjusted metrics due to the censoring in the test set. Since our models make use of IPC-weights from the censoring Deep Cox models during calibration, to ensure fair evaluation, we obtain IPC-weights for metric calculation from completely independently trained models from an entirely different model class — Random Survival Forests (RSF) (Ishwaran and Lu, 2019).

**Data and Training:** Across all 3 real datasets, we first extract the features from the foundation models. Since these features are extremely high dimensional, we run principal component analysis on these features, reducing the dimensionality to retain 90% of the variance. We then train RSFs on these reduced features to predict the censoring times. We use the `scikit-survival` library, setting the hyperparameter as `n_estimators=100, min_samples_split=10, min_samples_leaf=15`. These RSF models for censoring achieve C-indices of $0.664 \pm 0.002, 0.582 \pm 0.002$ and $0.615 \pm 0.002$ for the text reports, RNA-seq and H&E datasets respectively — comparable to the censoring Deep Cox models. Hence, these models are suitable to provide independent IPCW estimates for evaluating performance on the test instances. We clip the model predictions at $1e-4$ to ensure stability of the IPCW adjusted metrics.

# F ADDITIONAL EXPERIMENTS

## F.1 FULL SIMULATIONS RESULTS

We present the AUPIT and IBS metrics for the expanded set of estimators across the six simulation settings in Figure 3. In Table 5, we further examine quantile loss across levels 10% through 90%. The fraction of patients included in the computation of the quantile loss is shown in Figure 4.

The DR-ISR method continues to perform best empirically on the AUPIT and IBS metrics, followed by RW$^+$-ISR and HT-ISR. Our proposed estimators also improve quantile loss across all quantile levels in Settings 2–4. In the remaining settings, the differences across methods are small, and no estimators offers a significant advantage over the uncalibrated Cox baseline.

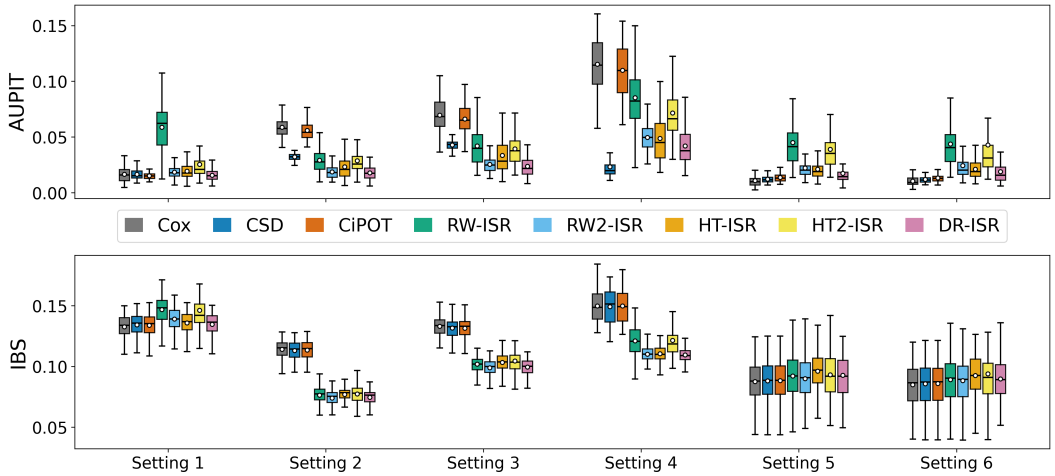



Figure 3: AUPIT and IBS across the six simulation settings for all estimators. Each boxplot summarizes variation over simulation seeds, where the center line indicates the median and the white dot indicates the mean.

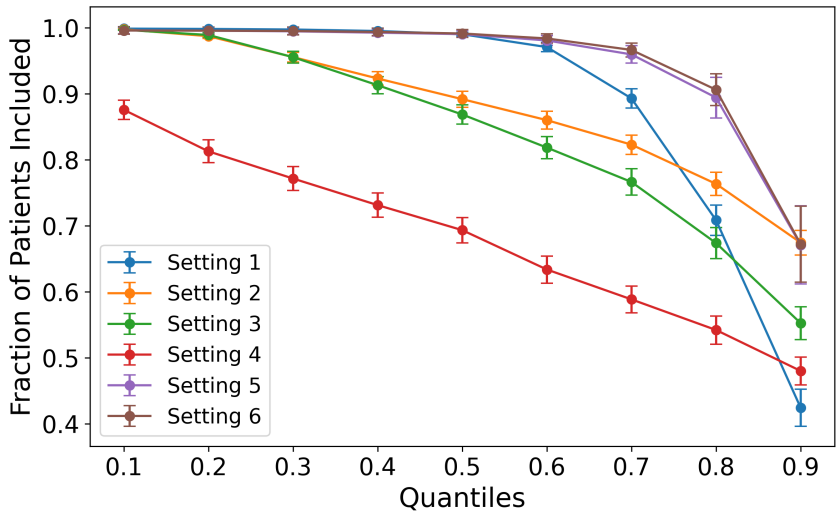



Figure 4: Fraction of total samples included in the computation of the quantile loss across nine different levels. Values shown are the mean over 100 seeds $\pm$ 2 standard error bars.

Table 5: Quantile loss from 10% to 90% on the test set across the six simulation settings. Entries report the mean over 100 seeds $\pm$ 2 standard errors. Bold values indicate the lowest quantile loss for a given setting and quantile when strictly lower than Cox after rounding. For brevity, we remove the ISR suffix on our methods.

| | Method | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% |
|---|---|---|---|---|---|---|---|---|---|---|
| Setting 1 | Cox | 3.39 ± 0.08 | 6.73 ± 0.16 | 9.99 ± 0.24 | 13.1 ± 0.32 | 15.9 ± 0.42 | 17.8 ± 0.58 | 16.4 ± 0.78 | 10.7 ± 0.87 | 4.27 ± 0.53 |
| | CSD | 3.39 ± 0.08 | 6.73 ± 0.16 | 9.99 ± 0.24 | 13.1 ± 0.32 | 15.9 ± 0.42 | 17.8 ± 0.58 | **16.3** ± 0.78 | 10.7 ± 0.86 | 4.28 ± 0.53 |
| | CiPOT | 3.39 ± 0.08 | 6.73 ± 0.16 | 9.99 ± 0.24 | 13.1 ± 0.32 | 15.9 ± 0.42 | 17.8 ± 0.58 | **16.3** ± 0.78 | 10.7 ± 0.86 | 4.27 ± 0.53 |
| | RW | 3.40 ± 0.08 | 6.75 ± 0.16 | 10.1 ± 0.25 | 13.2 ± 0.32 | 16.1 ± 0.42 | 18.0 ± 0.58 | 16.6 ± 0.79 | 10.9 ± 0.88 | 4.42 ± 0.55 |
| | RW$^+$ | 3.39 ± 0.08 | 6.74 ± 0.16 | 10.1 ± 0.24 | 13.2 ± 0.33 | 16.1 ± 0.44 | 17.9 ± 0.59 | 16.6 ± 0.83 | 11.0 ± 0.90 | 4.32 ± 0.54 |
| | HT | 3.39 ± 0.08 | 6.73 ± 0.16 | 10.00 ± 0.24 | 13.1 ± 0.32 | 15.9 ± 0.42 | 17.8 ± 0.58 | 16.4 ± 0.79 | 10.7 ± 0.87 | 4.31 ± 0.53 |
| | HT$^+$ | 3.79 ± 0.26 | 7.08 ± 0.27 | 10.4 ± 0.31 | 13.5 ± 0.36 | 16.3 ± 0.45 | 18.2 ± 0.73 | 16.8 ± 0.85 | 11.0 ± 0.90 | 4.33 ± 0.54 |
| | DR | 3.39 ± 0.08 | 6.73 ± 0.16 | 10.0 ± 0.24 | 13.1 ± 0.32 | 15.9 ± 0.42 | 17.8 ± 0.58 | 16.4 ± 0.78 | 10.7 ± 0.87 | 4.29 ± 0.54 |
| Setting 2 | Cox | 1.28 ± 0.02 | 2.41 ± 0.06 | 2.81 ± 0.09 | 2.88 ± 0.11 | 2.78 ± 0.12 | 2.54 ± 0.11 | 2.18 ± 0.09 | 1.64 ± 0.08 | 0.96 ± 0.07 |
| | CSD | 1.23 ± 0.01 | 2.39 ± 0.05 | 2.80 ± 0.09 | 2.87 ± 0.11 | 2.77 ± 0.12 | 2.54 ± 0.11 | 2.18 ± 0.10 | 1.64 ± 0.08 | 1.02 ± 0.06 |
| | CiPOT | 1.35 ± 0.03 | 2.42 ± 0.05 | 2.76 ± 0.09 | 2.83 ± 0.11 | 2.74 ± 0.12 | 2.52 ± 0.11 | 2.18 ± 0.10 | 1.65 ± 0.08 | 0.96 ± 0.07 |
| | RW | **0.84** ± 0.01 | 1.46 ± 0.01 | 1.84 ± 0.03 | 2.05 ± 0.04 | 2.14 ± 0.06 | 2.08 ± 0.07 | 1.86 ± 0.07 | 1.44 ± 0.07 | 0.84 ± 0.06 |
| | RW$^+$ | **0.84** ± 0.01 | **1.44** ± 0.01 | **1.82** ± 0.03 | **2.03** ± 0.04 | **2.09** ± 0.06 | 2.03 ± 0.07 | 1.82 ± 0.07 | 1.38 ± 0.07 | 0.78 ± 0.05 |
| | HT | **0.84** ± 0.01 | 1.45 ± 0.01 | **1.82** ± 0.03 | **2.03** ± 0.04 | 2.10 ± 0.06 | 2.04 ± 0.07 | 1.83 ± 0.07 | 1.40 ± 0.07 | 0.82 ± 0.06 |
| | HT$^+$ | 0.93 ± 0.04 | 1.51 ± 0.03 | 1.87 ± 0.03 | 2.09 ± 0.06 | 2.15 ± 0.06 | 2.07 ± 0.07 | 1.86 ± 0.07 | 1.40 ± 0.07 | 0.79 ± 0.06 |
| | DR | **0.84** ± 0.01 | 1.45 ± 0.01 | **1.82** ± 0.03 | **2.03** ± 0.04 | 2.10 ± 0.06 | **2.02** ± 0.07 | **1.80** ± 0.07 | **1.37** ± 0.07 | **0.77** ± 0.05 |
| Setting 3 | Cox | 0.52 ± 0.01 | 0.97 ± 0.02 | 1.17 ± 0.03 | 1.20 ± 0.04 | 1.14 ± 0.04 | 1.01 ± 0.04 | 0.85 ± 0.04 | 0.60 ± 0.03 | 0.32 ± 0.02 |
| | CSD | 0.51 ± 0.01 | 0.96 ± 0.01 | 1.18 ± 0.03 | 1.20 ± 0.04 | 1.14 ± 0.04 | 1.02 ± 0.04 | 0.85 ± 0.04 | 0.59 ± 0.03 | 0.33 ± 0.02 |
| | CiPOT | 0.53 ± 0.01 | 0.96 ± 0.02 | 1.15 ± 0.03 | 1.18 ± 0.04 | 1.12 ± 0.04 | 1.00 ± 0.04 | 0.85 ± 0.04 | 0.60 ± 0.03 | 0.32 ± 0.02 |
| | RW | **0.38** ± 0.01 | 0.68 ± 0.01 | 0.86 ± 0.02 | 0.95 ± 0.02 | 0.97 ± 0.02 | 0.93 ± 0.03 | 0.82 ± 0.03 | 0.59 ± 0.03 | 0.32 ± 0.02 |
| | RW$^+$ | **0.38** ± 0.01 | **0.66** ± 0.01 | **0.84** ± 0.01 | 0.94 ± 0.02 | 0.96 ± 0.02 | **0.91** ± 0.03 | **0.80** ± 0.03 | **0.57** ± 0.03 | **0.30** ± 0.02 |
| | HT | **0.38** ± 0.01 | **0.66** ± 0.01 | 0.85 ± 0.01 | 0.94 ± 0.02 | 0.96 ± 0.03 | 0.92 ± 0.03 | 0.81 ± 0.03 | 0.59 ± 0.03 | 0.32 ± 0.02 |
| | HT$^+$ | 0.44 ± 0.04 | 0.71 ± 0.03 | 0.88 ± 0.02 | 0.98 ± 0.03 | 0.99 ± 0.03 | 0.93 ± 0.03 | 0.81 ± 0.03 | 0.59 ± 0.04 | **0.30** ± 0.02 |
| | DR | **0.38** ± 0.01 | **0.66** ± 0.01 | **0.84** ± 0.01 | **0.93** ± 0.02 | **0.95** ± 0.02 | **0.91** ± 0.03 | **0.80** ± 0.03 | **0.57** ± 0.03 | **0.30** ± 0.02 |
| Setting 4 | Cox | 1.30 ± 0.09 | 1.59 ± 0.10 | 1.73 ± 0.10 | 1.73 ± 0.10 | 1.66 ± 0.10 | 1.41 ± 0.10 | 1.17 ± 0.09 | 0.89 ± 0.09 | 0.52 ± 0.07 |
| | CSD | 1.13 ± 0.02 | 1.86 ± 0.04 | 2.18 ± 0.05 | 1.98 ± 0.08 | 1.77 ± 0.09 | 1.47 ± 0.10 | 1.21 ± 0.09 | 0.90 ± 0.09 | 0.52 ± 0.06 |
| | CiPOT | 1.33 ± 0.09 | 1.59 ± 0.10 | 1.71 ± 0.10 | 1.72 ± 0.10 | 1.66 ± 0.10 | 1.41 ± 0.10 | 1.18 ± 0.09 | 0.90 ± 0.09 | 0.53 ± 0.07 |
| | RW | 0.78 ± 0.02 | 1.24 ± 0.05 | 1.52 ± 0.07 | 1.65 ± 0.08 | 1.66 ± 0.09 | 1.46 ± 0.10 | 1.24 ± 0.10 | 0.96 ± 0.10 | 0.58 ± 0.08 |
| | RW$^+$ | **0.77** ± 0.02 | **1.20** ± 0.04 | 1.47 ± 0.06 | 1.59 ± 0.08 | 1.59 ± 0.09 | 1.37 ± 0.09 | 1.15 ± 0.09 | 0.87 ± 0.09 | 0.52 ± 0.07 |
| | HT | **0.77** ± 0.02 | **1.20** ± 0.04 | **1.46** ± 0.06 | **1.58** ± 0.08 | 1.59 ± 0.09 | 1.38 ± 0.10 | 1.15 ± 0.09 | 0.87 ± 0.09 | 0.50 ± 0.06 |
| | HT$^+$ | 0.91 ± 0.07 | 1.30 ± 0.07 | 1.55 ± 0.07 | 1.67 ± 0.08 | 1.63 ± 0.09 | 1.41 ± 0.10 | 1.16 ± 0.09 | 0.88 ± 0.09 | 0.52 ± 0.07 |
| | DR | **0.77** ± 0.02 | **1.20** ± 0.04 | 1.47 ± 0.06 | **1.58** ± 0.07 | **1.58** ± 0.09 | **1.36** ± 0.10 | **1.14** ± 0.09 | **0.85** ± 0.08 | **0.49** ± 0.06 |
| Setting 5 | Cox | 0.65 ± 0.01 | 1.22 ± 0.01 | 1.72 ± 0.01 | 2.14 ± 0.02 | 2.47 ± 0.02 | 2.67 ± 0.03 | 2.67 ± 0.05 | 2.34 ± 0.09 | 1.41 ± 0.13 |
| | CSD | 0.65 ± 0.01 | 1.22 ± 0.01 | 1.72 ± 0.01 | 2.14 ± 0.02 | 2.47 ± 0.02 | 2.67 ± 0.03 | 2.67 ± 0.05 | 2.34 ± 0.09 | **1.40** ± 0.13 |
| | CiPOT | 0.65 ± 0.01 | 1.22 ± 0.01 | 1.72 ± 0.01 | 2.14 ± 0.02 | 2.47 ± 0.02 | 2.67 ± 0.03 | 2.67 ± 0.05 | 2.34 ± 0.09 | **1.40** ± 0.13 |
| | RW | 0.66 ± 0.02 | 1.26 ± 0.04 | 1.77 ± 0.03 | 2.24 ± 0.08 | 2.59 ± 0.07 | 2.77 ± 0.05 | 2.76 ± 0.06 | 2.42 ± 0.10 | 1.50 ± 0.13 |
| | RW$^+$ | 0.65 ± 0.01 | 1.22 ± 0.01 | 1.73 ± 0.02 | 2.16 ± 0.02 | 2.52 ± 0.03 | 2.72 ± 0.04 | 2.75 ± 0.06 | 2.39 ± 0.10 | 1.45 ± 0.13 |
| | HT | 0.65 ± 0.01 | 1.22 ± 0.01 | 1.73 ± 0.01 | 2.15 ± 0.02 | 2.49 ± 0.02 | 2.70 ± 0.03 | 2.73 ± 0.06 | 2.47 ± 0.11 | 1.48 ± 0.13 |
| | HT$^+$ | 0.69 ± 0.03 | 1.29 ± 0.04 | 1.83 ± 0.05 | 2.26 ± 0.06 | 2.58 ± 0.05 | 2.79 ± 0.05 | 2.81 ± 0.07 | 2.40 ± 0.10 | 1.46 ± 0.13 |
| | DR | 0.65 ± 0.01 | 1.22 ± 0.01 | 1.73 ± 0.01 | 2.15 ± 0.02 | 2.53 ± 0.06 | 2.71 ± 0.05 | 2.72 ± 0.06 | 2.42 ± 0.10 | 1.44 ± 0.13 |
| Setting 6 | Cox | 0.72 ± 0.01 | 1.35 ± 0.01 | 1.90 ± 0.02 | 2.38 ± 0.02 | 2.78 ± 0.03 | 3.05 ± 0.04 | 3.15 ± 0.05 | 2.90 ± 0.10 | 1.85 ± 0.17 |
| | CSD | 0.72 ± 0.01 | 1.35 ± 0.01 | 1.90 ± 0.02 | 2.38 ± 0.02 | 2.78 ± 0.03 | 3.05 ± 0.04 | 3.15 ± 0.05 | 2.90 ± 0.10 | **1.84** ± 0.17 |
| | CiPOT | 0.72 ± 0.01 | 1.35 ± 0.01 | 1.90 ± 0.02 | 2.38 ± 0.02 | 2.78 ± 0.03 | 3.05 ± 0.04 | 3.15 ± 0.05 | 2.90 ± 0.10 | 1.85 ± 0.17 |
| | RW | 0.73 ± 0.01 | 1.39 ± 0.04 | 1.96 ± 0.04 | 2.45 ± 0.04 | 2.85 ± 0.04 | 3.12 ± 0.04 | 3.24 ± 0.07 | 2.98 ± 0.11 | 1.95 ± 0.18 |
| | RW$^+$ | 0.72 ± 0.01 | 1.35 ± 0.01 | 1.91 ± 0.02 | 2.42 ± 0.04 | 2.86 ± 0.09 | 3.19 ± 0.09 | 3.28 ± 0.08 | 2.96 ± 0.11 | 1.88 ± 0.17 |
| | HT | 0.72 ± 0.01 | 1.35 ± 0.01 | 1.91 ± 0.02 | 2.39 ± 0.02 | 2.79 ± 0.03 | 3.07 ± 0.04 | 3.19 ± 0.05 | 2.98 ± 0.11 | 1.94 ± 0.18 |
| | HT$^+$ | 0.82 ± 0.10 | 1.45 ± 0.09 | 2.10 ± 0.15 | 2.61 ± 0.15 | 3.02 ± 0.13 | 3.27 ± 0.10 | 3.30 ± 0.08 | 2.97 ± 0.11 | 1.89 ± 0.17 |
| | DR | 0.72 ± 0.01 | 1.35 ± 0.01 | 1.91 ± 0.02 | 2.40 ± 0.03 | 2.81 ± 0.05 | 3.08 ± 0.05 | 3.25 ± 0.09 | 3.02 ± 0.12 | 1.88 ± 0.17 |

### F.2 Ablations on Simulations

We now conduct a variety of ablations on the simulations to systematically test the sensitivity of our method to various properties of the data.

#### F.2.1 Training-calibration data split

First, we examine sensitivity to the train/calibration split in the Tables 6 and 7, using synthetic Setting 3, where DR-ISR shows a clear advantage on AUPIT and IBS at the 50/50 split. We vary the split from 90/10 to 10/90 ($n_{\text{cal}}$ from 500 to 4,500) while holding the test set fixed.

The DR-ISR improvement is stable across most of this range and begins to narrow only at small calibration sets, around the 80/20 split and below.

Table 6: AUPIT across different training/calibration splits for Setting 3 (mean $\pm$ 2 standard error over 10 seeds). Smallest mean values bolded and underlined.

| train/cal split | Cox | CSD | CiPOT | RW-ISR | DR-ISR |
|---|---|---|---|---|---|
| 10/90 | $0.077 \pm 0.021$ | $0.043 \pm 0.004$ | $0.067 \pm 0.012$ | $0.047 \pm 0.015$ | $\mathbf{0.020 \pm 0.006}$ |
| 20/80 | $0.068 \pm 0.013$ | $0.045 \pm 0.003$ | $0.062 \pm 0.010$ | $0.039 \pm 0.011$ | $\mathbf{0.017 \pm 0.002}$ |
| 30/70 | $0.071 \pm 0.009$ | $0.044 \pm 0.002$ | $0.065 \pm 0.006$ | $0.034 \pm 0.008$ | $\mathbf{0.016 \pm 0.003}$ |
| 40/60 | $0.068 \pm 0.008$ | $0.042 \pm 0.003$ | $0.065 \pm 0.005$ | $0.048 \pm 0.013$ | $\mathbf{0.018 \pm 0.004}$ |
| 50/50 | $0.067 \pm 0.008$ | $0.043 \pm 0.003$ | $0.064 \pm 0.007$ | $0.046 \pm 0.018$ | $\mathbf{0.019 \pm 0.003}$ |
| 60/40 | $0.066 \pm 0.007$ | $0.043 \pm 0.003$ | $0.063 \pm 0.007$ | $0.049 \pm 0.014$ | $\mathbf{0.020 \pm 0.004}$ |
| 70/30 | $0.064 \pm 0.007$ | $0.043 \pm 0.003$ | $0.061 \pm 0.006$ | $0.060 \pm 0.013$ | $\mathbf{0.026 \pm 0.006}$ |
| 80/20 | $0.063 \pm 0.006$ | $0.043 \pm 0.003$ | $0.061 \pm 0.006$ | $0.075 \pm 0.014$ | $\mathbf{0.033 \pm 0.009}$ |
| 90/10 | $0.065 \pm 0.005$ | $\mathbf{0.043 \pm 0.002}$ | $0.062 \pm 0.005$ | $0.077 \pm 0.019$ | $0.047 \pm 0.011$ |

Table 7: IBS across different training/calibration splits for Setting 3 (mean $\pm$ 2 standard error over 10 seeds). Smallest mean values bolded and underlined.

| train/cal split | Cox | CSD | CiPOT | RW-ISR | DR-ISR |
|---|---|---|---|---|---|
| 10/90 | $0.139 \pm 0.012$ | $0.138 \pm 0.014$ | $0.135 \pm 0.012$ | $0.100 \pm 0.006$ | $\mathbf{0.099 \pm 0.004}$ |
| 20/80 | $0.132 \pm 0.006$ | $0.132 \pm 0.008$ | $0.130 \pm 0.007$ | $0.099 \pm 0.005$ | $\mathbf{0.098 \pm 0.004}$ |
| 30/70 | $0.133 \pm 0.006$ | $0.133 \pm 0.007$ | $0.132 \pm 0.007$ | $0.099 \pm 0.005$ | $\mathbf{0.098 \pm 0.004}$ |
| 40/60 | $0.133 \pm 0.006$ | $0.132 \pm 0.006$ | $0.132 \pm 0.006$ | $0.102 \pm 0.006$ | $\mathbf{0.098 \pm 0.004}$ |
| 50/50 | $0.133 \pm 0.006$ | $0.132 \pm 0.006$ | $0.132 \pm 0.006$ | $0.104 \pm 0.007$ | $\mathbf{0.098 \pm 0.004}$ |
| 60/40 | $0.131 \pm 0.005$ | $0.130 \pm 0.005$ | $0.130 \pm 0.005$ | $0.104 \pm 0.005$ | $\mathbf{0.099 \pm 0.003}$ |
| 70/30 | $0.131 \pm 0.005$ | $0.130 \pm 0.006$ | $0.130 \pm 0.006$ | $0.103 \pm 0.006$ | $\mathbf{0.100 \pm 0.004}$ |
| 80/20 | $0.130 \pm 0.004$ | $0.129 \pm 0.005$ | $0.129 \pm 0.005$ | $0.107 \pm 0.007$ | $\mathbf{0.102 \pm 0.006}$ |
| 90/10 | $0.131 \pm 0.005$ | $0.131 \pm 0.005$ | $0.131 \pm 0.005$ | $0.110 \pm 0.011$ | $\mathbf{0.106 \pm 0.005}$ |

### F.2.2 DATASET SIZE

Next, we examine sensitivity to the overall dataset size in the Tables 8 and 9, using synthetic Setting 3, where DR-ISR showed a clear advantage on AUPIT and IBS at the 50/50 split with 5000 data points. We vary the dataset size from 500 to 10000 while holding the test set fixed at 5000 and always using half the data for training and half for calibration.

The advantage of DR-ISR over the baselines narrows at smaller calibration sizes, due to the higher variance of the doubly robust pseudo-outcomes. For AUPIT, we find DR-ISR becomes worse than the CSD baseline at small calibration set sizes

Table 8: AUPIT across different sample sizes for Setting 3 (mean $\pm$ 2 standard error over 10 seeds). Smallest mean values bolded and underlined.

| $n$ | Cox | CSD | CiPOT | DR-ISR |
|---|---|---|---|---|
| 500 | $0.087 \pm 0.021$ | $\mathbf{0.046 \pm 0.007}$ | $0.077 \pm 0.016$ | $0.082 \pm 0.022$ |
| 1000 | $0.070 \pm 0.010$ | $\mathbf{0.040 \pm 0.004}$ | $0.063 \pm 0.009$ | $0.059 \pm 0.013$ |
| 2500 | $0.065 \pm 0.006$ | $0.046 \pm 0.005$ | $0.060 \pm 0.005$ | $\mathbf{0.029 \pm 0.004}$ |
| 5000 | $0.069 \pm 0.006$ | $0.042 \pm 0.002$ | $0.066 \pm 0.005$ | $\mathbf{0.024 \pm 0.004}$ |
| 10000 | $0.068 \pm 0.004$ | $0.044 \pm 0.001$ | $0.066 \pm 0.003$ | $\mathbf{0.016 \pm 0.003}$ |

Table 9: IBS across different sample sizes for Setting 3 (mean $\pm$ 2 standard error over 10 seeds). Smallest mean values bolded and underlined.

| $n$ | Cox | CSD | CiPOT | DR-ISR |
|---|---|---|---|---|
| 500 | $0.141 \pm 0.009$ | $0.142 \pm 0.011$ | $\mathbf{0.139 \pm 0.009}$ | $\mathbf{0.139 \pm 0.012}$ |
| 1000 | $0.137 \pm 0.008$ | $0.136 \pm 0.009$ | $0.136 \pm 0.008$ | $\mathbf{0.120 \pm 0.007}$ |
| 2500 | $0.133 \pm 0.002$ | $0.132 \pm 0.003$ | $0.132 \pm 0.002$ | $\mathbf{0.107 \pm 0.004}$ |
| 5000 | $0.133 \pm 0.004$ | $0.132 \pm 0.004$ | $0.132 \pm 0.004$ | $\mathbf{0.099 \pm 0.003}$ |
| 10000 | $0.132 \pm 0.004$ | $0.131 \pm 0.004$ | $0.131 \pm 0.004$ | $\mathbf{0.096 \pm 0.003}$ |

### F.2.3 Effect of Censoring Rate

Once again, on setting 3, we hold the outcome mechanism fixed and vary the shape of the censoring distribution to produce five settings with censoring rates of approximately 10%, 30%, 50%, 70%, and 90% (Tables 10 and 11).

The improvement of DR-ISR over the other methods is relatively stable across all settings except under very high (≈90%) censoring.

Table 10: AUPIT across different censoring rates $\eta$ modifying Setting 3 (mean $\pm$ 2 standard error over 20 seeds). Smallest mean values bolded and underlined

| Setting<br>$\eta$ | Very Low<br>$\approx 10\%$ | Low<br>$\approx 30\%$ | Medium<br>$\approx 50\%$ | High<br>$\approx 70\%$ | Very High<br>$\approx 90\%$ |
|---|---|---|---|---|---|
| Cox | $0.028 \pm 0.001$ | $0.031 \pm 0.002$ | $0.041 \pm 0.004$ | $0.061 \pm 0.004$ | $0.157 \pm 0.019$ |
| CSD | $0.010 \pm 0.001$ | $0.016 \pm 0.001$ | $0.026 \pm 0.001$ | $0.040 \pm 0.002$ | $\mathbf{0.049 \pm 0.005}$ |
| CiPOT | $0.011 \pm 0.002$ | $0.019 \pm 0.002$ | $0.035 \pm 0.004$ | $0.058 \pm 0.004$ | $0.136 \pm 0.018$ |
| RW-ISR | $\mathbf{0.009 \pm 0.002}$ | $0.010 \pm 0.002$ | $0.014 \pm 0.002$ | $0.029 \pm 0.005$ | $0.146 \pm 0.021$ |
| DR-ISR | $\mathbf{0.009 \pm 0.002}$ | $\mathbf{0.008 \pm 0.001}$ | $\mathbf{0.011 \pm 0.002}$ | $\mathbf{0.018 \pm 0.002}$ | $0.082 \pm 0.014$ |

Table 11: IBS across different censoring rates $\eta$ modifying Setting 3 (mean $\pm$ 2 standard error over 20 seeds). Smallest mean values bolded and underlined

| Setting<br>$\eta$ | Very Low<br>$\approx 10\%$ | Low<br>$\approx 30\%$ | Medium<br>$\approx 50\%$ | High<br>$\approx 70\%$ | Very High<br>$\approx 90\%$ |
|---|---|---|---|---|---|
| Cox | $0.054 \pm 0.004$ | $0.068 \pm 0.004$ | $0.090 \pm 0.005$ | $0.125 \pm 0.003$ | $0.142 \pm 0.006$ |
| CSD | $0.054 \pm 0.004$ | $0.067 \pm 0.004$ | $0.089 \pm 0.005$ | $0.123 \pm 0.004$ | $0.164 \pm 0.006$ |
| CiPOT | $0.053 \pm 0.004$ | $0.067 \pm 0.004$ | $0.089 \pm 0.005$ | $0.123 \pm 0.004$ | $0.141 \pm 0.006$ |
| RW-ISR | $\mathbf{0.045 \pm 0.003}$ | $\mathbf{0.055 \pm 0.003}$ | $\mathbf{0.069 \pm 0.003}$ | $0.092 \pm 0.004$ | $0.158 \pm 0.008$ |
| DR-ISR | $\mathbf{0.045 \pm 0.003}$ | $\mathbf{0.055 \pm 0.003}$ | $\mathbf{0.069 \pm 0.003}$ | $\mathbf{0.090 \pm 0.004}$ | $\mathbf{0.128 \pm 0.005}$ |

### F.2.4 Crossing Survival Curves

Lastly, we test the sensitivity of our method when we explicitly violate Assumption A2, i.e. when the survival curves corresponding to two different risk scores cross. In such a scenario, a patient with a higher risk may have lower survival probability at an earlier time but higher survival probability at a later time when compared to a patient with lower risk.

We probe this case by modifying synthetic Settings 3 and 5 so that the scale of the standard deviation of the log-normal distribution also depends on $X$. This induces crossings in the survival curves such that no risk score can order correctly across all time points. We report the calibration metrics for these settings in Table 12 and 13 below.

The DR-ISR results are robust to this violation. In Setting 3, DR-ISR continues to perform strongly after A2 is violated. In Setting 5 where DR-ISR was slightly worse but competitive, the relative performance remains similar even after the violation. In both cases, inducing crossing survival curves does not significantly degrade the calibration of DR-ISR relative to the baselines.

Table 12: AUPIT across modified synthetic settings where true survival curves cross. We report the mean ($\pm 2$ s.e.) over 20 seeds. Smallest mean values are bolded and underlined.

| Model | Setting 3 (modified) | Setting 5 (modified) |
|---|---|---|
| Cox | $0.048 \pm 0.005$ | $\underline{\mathbf{0.016 \pm 0.003}}$ |
| CSD | $0.049 \pm 0.002$ | $0.017 \pm 0.002$ |
| CiPOT | $0.046 \pm 0.004$ | $0.018 \pm 0.002$ |
| RW-ISR | $0.046 \pm 0.007$ | $0.073 \pm 0.011$ |
| DR-ISR | $\underline{\mathbf{0.023 \pm 0.004}}$ | $0.019 \pm 0.004$ |

Table 13: IBS across modified synthetic settings where true survival curves cross. We report the mean ($\pm 2$ s.e.) over 20 seeds. Smallest mean values are bolded and underlined.

| Model | Setting 3 (modified) | Setting 5 (modified) |
|---|---|---|
| Cox | $0.139 \pm 0.003$ | $\underline{\mathbf{0.137 \pm 0.006}}$ |
| CSD | $0.139 \pm 0.003$ | $0.140 \pm 0.007$ |
| CiPOT | $0.139 \pm 0.003$ | $0.140 \pm 0.006$ |
| RW-ISR | $0.106 \pm 0.004$ | $0.150 \pm 0.008$ |
| DR-ISR | $\underline{\mathbf{0.101 \pm 0.003}}$ | $0.139 \pm 0.007$ |

### F.3 Full real-data results

For completeness, we now present complete results on the real data for IPCW evaluation of the metrics as described in Sec. 6. We also include the RW$^+$-ISR, HT-ISR and HT$^+$-ISR methods in our results. Table 14 lists the IPCW estimates of AUPIT and IBS. Figure 5 visualizes the quantile losses for all models. As described in Sec. E.1, when evaluating quantile loss, if the predicted survival curve does not reach the target quantile by the maximum valid prediction time (maximum observed time across train and calibration splits for the ISR methods) for any one method, we exclude that patient from evaluation for all methods.

As discussed in Sec. 6.2, DR-ISR exhibits the strongest performance for the IPCW estimates, even improving upon the baselines CSD and CiPOT which explicitly target PIT calibration on the AUPIT metric. RW$^+$-ISR also exhibits strong performance, which is expected by improving upon RW-ISR in terms of data efficiency, as discussed in Sec. C, and also achieves lower IBS compared to DR-ISR. On extreme quantiles performance of all methods is similar with RW$^+$-ISR clearly improving near the middle on all datasets and DR-ISR improving on the text reports and RNA-seq data. While DR-ISR has the strongest theoretical guarantees, we find that RW$^+$-ISR also performs well on the real data. This may because RW$^+$-ISR is less sensitive to violations of the assumptions or due to the IPCW evaluation.

Table 14: IPCW estimates of AUPIT and IBS over the test set for the Text Reports, RNA-seq, and H&E models. Values are mean over 100 seeds with p-value in brackets (calculated by taking the maximum across paired one sided t-test p-values comparing pairwise with all other methods). Entries are bolded if $p < 0.05$. Note that the $p$-values differ from those reported in Table 2 due to comparison with the additional estimators.

| | Method | AUPIT | IBS |
|---|---|---|---|
| Text Reports | Deep Cox | 0.0970 (1.00) | 0.125 (1.00) |
| | CSD | 0.1256 (1.00) | 0.125 (1.00) |
| | CiPOT | 0.0948 (1.00) | 0.125 (1.00) |
| | RW (ours) | 0.1394 (1.00) | 0.123 (1.00) |
| | HT (ours) | 0.0935 (1.00) | 0.124 (1.00) |
| | RW+ (ours) | 0.0839 (0.98) | **0.111 (<0.001)** |
| | HT+ (ours) | 0.0999 (1.00) | 0.118 (1.00) |
| | DR (ours) | **0.0819 (0.02)** | 0.122 (1.00) |
| RNA-seq | Deep Cox | 0.0546 (1.00) | 0.102 (1.00) |
| | CSD | 0.0826 (1.00) | 0.102 (1.00) |
| | CiPOT | 0.0544 (1.00) | 0.102 (1.00) |
| | RW (ours) | 0.0923 (1.00) | 0.092 (1.00) |
| | HT (ours) | 0.0566 (1.00) | 0.103 (1.00) |
| | RW+ (ours) | 0.0532 (1.00) | **0.088 (<0.001)** |
| | HT+ (ours) | 0.0590 (1.00) | 0.090 (1.00) |
| | DR (ours) | **0.0517 (<0.001)** | 0.098 (1.00) |
| H&E Slides | Deep Cox | 0.1007 (1.00) | 0.113 (1.00) |
| | CSD | 0.0787 (1.00) | 0.110 (1.00) |
| | CiPOT | 0.0800 (1.00) | 0.111 (1.00) |
| | RW (ours) | 0.0889 (1.00) | 0.096 (1.00) |
| | HT (ours) | 0.0972 (1.00) | 0.110 (1.00) |
| | RW+ (ours) | 0.0664 (1.00) | **0.094 (0.001)** |
| | HT+ (ours) | 0.1363 (1.00) | 0.117 (1.00) |
| | DR (ours) | **0.0568 (<0.001)** | 0.110 (1.00) |

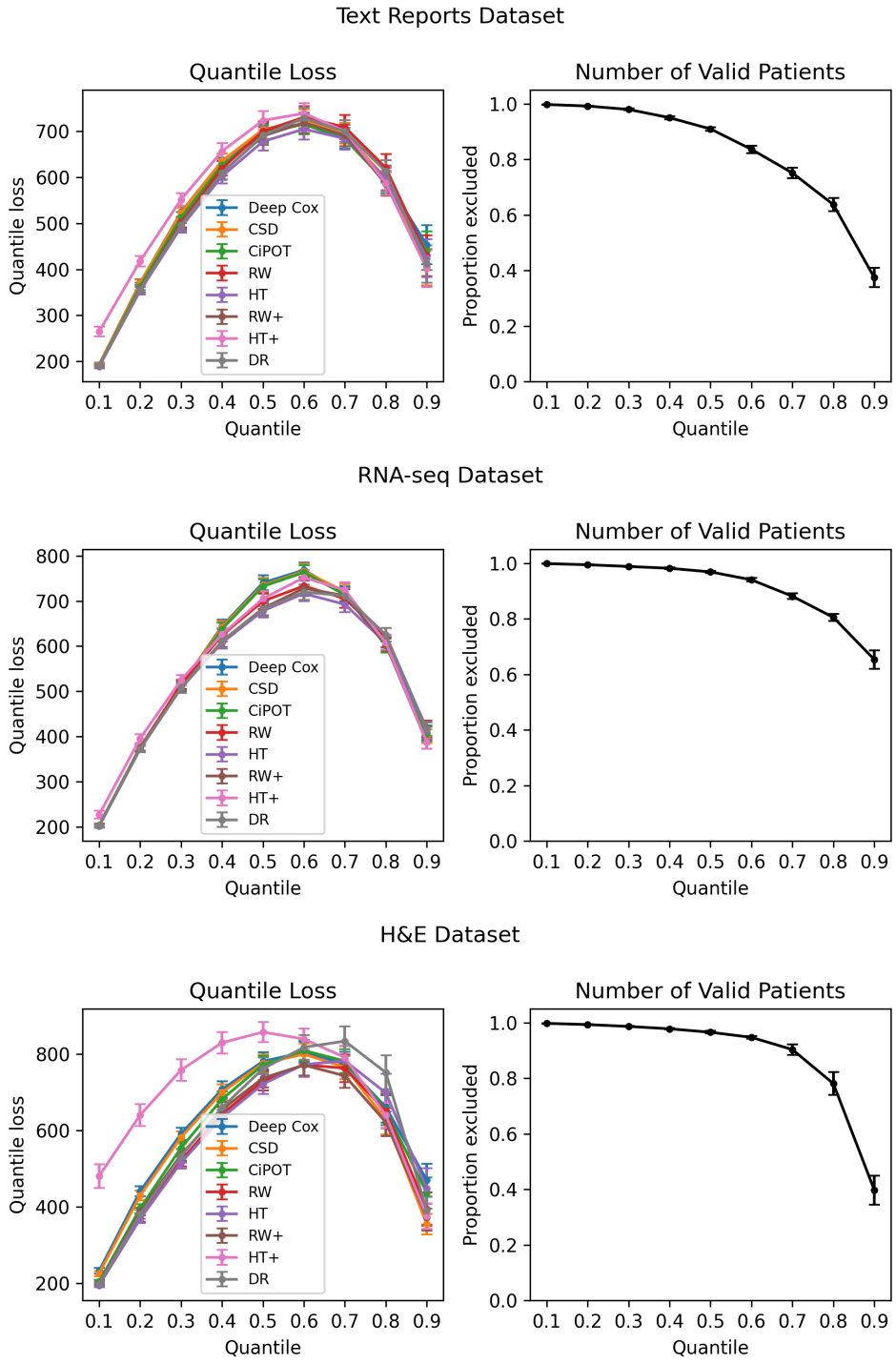



Figure 5: Quantile losses for the text reports (top), RNA-seq (middle) and H&E datasets. Right panel visualizes proportion of patients for which the survival curve reached the desired quantile across all methods and hence were included in evaluation. See Sec. E.1 for details. Values shown are mean over 100 seeds $\pm 2$ standard error bars.

## G EXPERIMENTAL ASSETS

The TCGA patient outcomes were obtained from Liu et al. (2018) which is publicly available under the CC BY-NC-ND license. The pathology reports were obtained from Kefeli and Tatonetti (2024) available under the CC-BY license. The BioClinical BERT model from Alsentzer et al. (2019) is available under the MIT license. The RNA-seq samples were obtained from Vivian et al. (2017) available under the Apache 2.0 license. The BulkFormer model from Kang et al. (2025) is available under the MIT license. The SPARC model from Ayed et al. (2026) is available under the CC-BY-NC 4.0 license. Lastly, the H-optimus-1 model from Bioptimus (2025) is available under the CC BY-NC-ND 4.0 license.

The code for the baseline methods was sourced from the official GitHub repositories, available under the MIT license. For training the linear Cox models, we use the official `lifelines` package, available under the MIT license. For training the Random Survival Forests, we use the `scikit-survival` library, available under the GNU General Public License v3.0 .