Proper Scoring Rules for Survival Analysis

Hiroki Yanagisawa [1]

Abstract

Survival analysis is the problem of estimating probability distributions for future event times, which can be seen as a problem in uncertainty quantification. Although there are fundamental theories on strictly proper scoring rules for uncertainty quantification, little is known about those for survival analysis. In this paper, we investigate extensions of four major strictly proper scoring rules for survival analysis and we prove that these extensions are proper under certain conditions, which arise from the discretization of the estimation of probability distributions. We also compare the estimation performances of these extended scoring rules by using real datasets, and the extensions of the logarithmic score and the Brier score performed the best.

1. Introduction

The theory of *scoring rules* is a fundamental theory in statistical analysis, and it is widely used in uncertainty quantification (see, e.g., (Mura et al., 2008; Parmigiani & Inoue, 2009; Benedetti, 2010; Schlag et al., 2015)). Suppose that there is a random variable Y whose cumulative distribution function (CDF) is F_Y. Given an estimation $\hat{F}_Y$ of F_Y and a single sample y obtained from Y, a scoring rule $S(\hat{F}_Y, y)$ is a function that returns an evaluation score for $\hat{F}_Y$ based on y. Since $\hat{F}_Y$ is a CDF and y is a single sample of Y, it is not straightforward to choose an appropriate scoring rule $S(\hat{F}_Y, y)$. The theory of scoring rules suggests how to choose an appropriate one. This theory proves that a certain class of scoring rules satisfies this natural property: the average evaluation score $S(\hat{F}_Y, y)$ over $y \sim Y$ is minimized only by the true CDF F_Y. A scoring rule that satisfies this property is called *strictly proper* in this theory. Examples of strictly proper scoring rules include the pinball loss, the logarithmic score, the Brier score, and the ranked probability score (see, e.g., (Gneiting & Raftery, 2007) for the definitions of these scoring rules). In uncertainty quantification, it is standard to use a strictly proper scoring rule both for a loss function to train machine learning models and for an evaluation metric to evaluate the models. Note that, if we use a non-proper scoring rule $S(\hat{F}_Y, y)$ as a loss function, a prediction model (e.g., a neural network model) might find an estimation $\hat{F}_Y$ such that $S(\hat{F}_Y, y) < S(F_Y, y)$ holds for $y \sim Y$ on average and such $\hat{F}_Y$ could be very different from true F_Y.

Survival analysis, which is also known as *time-to-event analysis*, can be seen as a problem in uncertainty quantification. Despite the long history of research from the seminal work (Cox, 1972) on survival analysis (see, e.g., (Wang et al., 2019) for a comprehensive survey), little is known about the strictly proper scoring rules for survival analysis. Therefore, we investigate extensions of these strictly proper scoring rules for survival analysis.

In survival analysis, the time between a well-defined starting point and the occurrence of an event is called the *survival time* or *event time*, and the goal of survival analysis is to estimate the probability distribution of event times. In healthcare applications, an event usually corresponds to an undesirable event for a patient (e.g., a death or the onset of disease). Survival analysis has important applications in many fields such as credit scoring (Dirick et al., 2017) and fraud detection (Zheng et al., 2019) as well as healthcare.

Datasets for survival analysis are *censored*, which means that events of interest might not be observed for a number of data points. This may be due to either a limited observation time window or missing traces caused by other irrelevant events. In this paper, we consider only *right censored* data, which is a widely studied problem setting in survival analysis. The exact event time of a right censored data point is unknown; we know only that the event had not happened up to a certain time for the data point. The time between a well-defined starting point and the last observation time of a right censored data point is called the *censoring time*.

Many neural network models have been proposed for survival analysis (e.g., (Avati et al., 2019; Kamran & Wiens, 2021; Pearce et al., 2022)). A common problem with these models is that they define their own custom loss functions, and they use these loss functions without proving that they

[1]IBM Research - Tokyo, Tokyo, Japan. Correspondence to: Hiroki Yanagisawa <yanagis@jp.ibm.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

are strictly proper in terms of the theory of scoring rules. Indeed, Rindt et al. (2022) show that the loss functions used in (Avati et al., 2019; Kamran & Wiens, 2021) are not proper. Moreover, survival models have been evaluated by custom evaluation metrics without proving that these metrics are proper in terms of the theory of scoring rules. Popular metrics used for survival analysis include the integrated Brier score (Graf et al., 1999) and variants of C-index (Antolini et al., 2005; Uno et al., 2011). However, all of them are not proper (Blanche et al., 2018; Rindt et al., 2022). We also note that Sonabend et al. (2022) discuss the problems of using these variants of C-index as evaluation metrics in survival analysis.

The only exception to the above argument is (Rindt et al., 2022). This paper shows a rigorous proof that an extension of the logarithmic score for survival analysis is strictly proper. Note that this paper is not the first one that uses this extension of the logarithmic score (e.g., (Lee et al., 2018; Ren et al., 2019; Tjandra et al., 2021)). However, it is usually used in *part* of the loss functions of the proposed models, and these loss functions are used without the proof of properness.

Our contributions. We analyze survival analysis through the lens of the theory of scoring rules. First, we prove that Portnoy's estimator (Portnoy, 2003), which is an extension of the pinball loss for survival analysis, is proper under certain conditions. This is the first proof for the properness for Portnoy's estimator. In addition, we show such conditions are due to the discretization of the estimation of a probability distribution and we explain why such conditions are required to be proper scoring rules for survival analysis. Second, we show that the proof of strict properness of the extension of the logarithmic score (Rindt et al., 2022) is based on implicit assumptions by showing its alternative proof. Third, we show two new proper scoring rules for survival analysis under certain conditions by extending the Brier score and the ranked probability score. These scoring rules are the first scoring rules with rigorous proofs of properness as extensions of the Brier score and the ranked probability score. Finally, we compare these four extensions of the scoring rules by using real datasets, and the results show that the extensions of the logarithmic score and Brier score performed the best.

2. Related Work

Survival analysis has been traditionally studied under the *proportional hazard assumption*. Its seminal work is the Cox model (Cox, 1972), and many other prediction models have been proposed under this strong assumption. Since outputs of these models are scalar values called *hazard rates* and are not CDFs, we use different types of loss functions and evaluation metrics in traditional survival analysis. One of the popular evaluation metrics is the concordance index (C-index) (Harrell et al., 1982), which is a generalization of the Kendall rank correlation coefficient. See, e.g., (Wang et al., 2019) for a comprehensive survey on survival analysis based on this assumption. In this paper, we focus on survival analysis *without* this assumption.

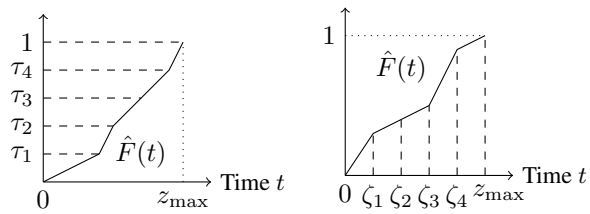
We note that there are many loss functions used in survival models that can be seen as variants of known scoring rules.

- **Pinball loss.** Portnoy's estimator (Portnoy, 2003), which is an extension of the pinball loss, has been used in quantile regression-based survival analysis (Portnoy, 2003; Neocleous et al., 2006; Pearce et al., 2022). It was unknown if this estimator is proper or not in terms of the theory of scoring rules, and we are the first to prove that this estimator is proper under a certain condition.
- **Brier score.** The IPCW Brier score (Graf et al., 1999) and integrated Brier score (Graf et al., 1999) are widely used in survival analysis (e.g., (Kvamme et al., 2019; Haider et al., 2020; Han et al., 2021; Zhong et al., 2021)) as variants of the Brier score. However, Rindt et al. (2022) show that neither of them is proper in terms of the theory of scoring rules.
- **Ranked probability score.** Variants of the ranked probability score have been proposed in (Avati et al., 2019; Kamran & Wiens, 2021), but Rindt et al. (2022) show that they are not proper in terms of the theory of scoring rules.

3. Preliminaries

Given a feature vector $x \in X$, let T_x and C_x be random variables for the event time and censoring time of x, respectively. Unless otherwise stated, we consider a fixed x, and we denote them by T and C by omitting the subscript x from T_x and C_x, respectively.

Let $t \sim T$ and $c \sim C$ be samples obtained from T and C, respectively. We assume that t and c are positive real values (i.e., $t \in \mathbb{R}^+$ and $c \in \mathbb{R}^+$). In survival analysis, we can observe only the minimum $z = \min\{t, c\}$, and we use $\delta = \mathbb{1}(t \leq c)$ to indicate whether z represents the true event time (i.e., $\delta = 1$ means z is uncensored and $z = t$) or z represents the censoring time (i.e., $\delta = 0$ means z is censored and $z = c$). In this paper, a pair of samples (t, c) is often represented as a pair of values (z, δ) to emphasize that we can observe only one of t and c. We assume that we have prior knowledge that z is at most $z_{\max}$ (i.e., $0 < z \leq z_{\max}$ holds for any z). Let $F(t)$ be the CDF of T, which is defined as $F(t) = \Pr(T \leq t)$. By the definition of $F(t)$, we have $F(0) = 0$, and we can represent



(a) Quantile regression (b) Distribution regression

Figure 1. Two types of discretization of probability distribution $\hat{F}(t)$ with $B = 5$.

the probability that the true event time is between t_1 and t_2 as $\Pr(t_1 < T \leq t_2) = F(t_2) - F(t_1)$.

Survival analysis is the problem of estimating the $\hat{F}(t)$ of the true CDF $F(t)$. For simplicity, we assume that both $F(t)$ and $\hat{F}(t)$ are monotonically increasing continuous functions. This means that $F(t_1) < F(t_2)$ holds if and only if $0 \leq t_1 < t_2 < \infty$. This assumption enables us to calculate $F(t)$ for any time $0 \leq t < \infty$ and to calculate $F^{-1}(\tau)$ for any quantile level $0 \leq \tau \leq 1$. When we estimate $\hat{F}(t)$ by using a prediction model (e.g., a neural network), we usually discretize $p = \hat{F}(t)$ along the p-axis or the t-axis as shown in Fig. 1. In quantile regression-based survival analysis, $p = \hat{F}(t)$ is discretized along the p-axis, $\hat{F}^{-1}(\tau_i)$ is estimated for $0 = \tau_0 < \tau_1 < \cdots < \tau_{B-1} < \tau_B = 1$, and we assume that $\hat{F}^{-1}(\tau_0) = 0$ and $\hat{F}^{-1}(\tau_B) = z_{\max}$. In distribution regression-based survival analysis, $p = \hat{F}(t)$ is discretized along the t-axis, $\hat{F}(\zeta_i)$ is estimated for $0 = \zeta_0 < \zeta_1 < \cdots < \zeta_{B-1} < \zeta_B = z_{\max}$, and we assume that $\hat{F}(\zeta_0) = 0$ and $\hat{F}(\zeta_B) = 1$.

Throughout this paper, we assume that the censoring time and the event time are independent of each other given a feature vector x. This assumption is usually represented as follows.

Assumption 3.1. $T \perp\!\!\!\perp C | X$.

This assumption is widely used in survival analysis (e.g., the classical Kaplan-Meier estimator (Kaplan & Meier, 1958) and the calibration metric D-calibration (Haider et al., 2020)). We can find examples of the other stronger assumptions (e.g., unconditionally random right censoring) used in survival analysis in (Peng, 2021).

4. Proper Scoring Rules for Survival Analysis

We briefly review the theory of scoring rules for uncertainty quantification. Let Y be a random variable, and let $F_Y(y)$ be its CDF, which is defined as $F_Y(y) = \Pr(Y \leq y)$. A *scoring rule* is a function $S(\hat{F}_Y, y)$ that returns a real value (i.e., an evaluation score) for inputs $\hat{F}_Y$ and y, where $\hat{F}_Y$ is an estimation of F_Y and y is a sample obtained from Y. In this paper, we consider *negatively-oriented* scoring rules, which means that a smaller score is better. We can interpret the scoring rule $S(\hat{F}_Y, y)$ as a penalty function for the misestimation of $\hat{F}_Y$ for a sample y.

The *proper* and *strictly proper* scoring rules are defined as follows.

Definition 4.1. A scoring rule $S(\hat{F}_Y, y)$ is *proper* if

$$\mathop{\mathbb{E}}_{y \sim Y}[S(\hat{F}_Y, y)] \geq \mathop{\mathbb{E}}_{y \sim Y}[S(F_Y, y)].$$

Definition 4.2. A scoring rule $S(\hat{F}_Y, y)$ is *strictly proper* if

$$\mathop{\mathbb{E}}_{y \sim Y}[S(\hat{F}_Y, y)] \geq \mathop{\mathbb{E}}_{y \sim Y}[S(F_Y, y)]$$

holds and the equality holds only when $\hat{F}_Y = F_Y$.

These definitions are based on a natural property that any scoring rule should satisfy. Definition 4.2 means that we can recover the true F_Y by minimizing the average evaluation score $S(\hat{F}_Y, y)$ over $y \sim Y$ for a strictly proper scoring rule $S(\cdot, \cdot)$.

We extend these definitions of the proper and strictly proper scoring rules for survival analysis. We define the *proper* and *strictly proper* scoring rules for survival analysis by changing the inputs of a scoring rule $S(\hat{F}, (z, \delta))$ from F_Y and y to F and (z, δ).

Definition 4.3. A scoring rule $S(\hat{F}, (z, \delta))$ is *proper* if

$$\mathop{\mathbb{E}}_{(t,c) \sim (T,C)}[S(\hat{F}, (z, \delta))] \geq \mathop{\mathbb{E}}_{(t,c) \sim (T,C)}[S(F, (z, \delta))].$$

Definition 4.4. A scoring rule $S(\hat{F}, (z, \delta))$ is *strictly proper* if

$$\mathop{\mathbb{E}}_{(t,c) \sim (T,C)}[S(\hat{F}, (z, \delta))] \geq \mathop{\mathbb{E}}_{(t,c) \sim (T,C)}[S(F, (z, \delta))]$$

holds and the equality holds only when $\hat{F} = F$.

Following the standard approach of using a strictly proper scoring rule in uncertainty quantification (Bengs et al., 2022), we explain how to use a scoring rule $S(\hat{F}, (z, \delta))$ as a loss function in survival analysis. Given a training dataset $\{(x^{(i)}, z^{(i)}, \delta^{(i)})\}_{i=1}^n$, we formulate survival analysis as minimizing the empirical loss

$$\sum_{i=1}^{n} S(\hat{F}_{x^{(i)}}, (z^{(i)}, \delta^{(i)})),$$

where $\hat{F}_{x^{(i)}}$ is an estimation of the true CDF $F_{x^{(i)}}$ of random variable $T_{x^{(i)}}$ for $x^{(i)} \in X$. This formulation assumes that each $x^{(i)} \in X$ has an underlying random variable $T_{x^{(i)}}$ for event times, and our task is to find a good estimation $\hat{F}_{x^{(i)}}$ for each $x^{(i)}$.

In this paper, we investigate the extensions of the scoring rules for survival analysis. In Sec. 4.1, we consider quantile regression and survival analysis based on quantile regression. In Secs. 4.2–4.4, we consider distribution regression and survival analysis based on distribution regression.

4.1. Extension of Pinball Loss

We first review quantile regression (Koenker & Bassett, 1978; Koenker & Hallock, 2001). Let Y be a real-valued random variable and F_Y be its CDF. In quantile regression, we estimate the τ-th quantile of Y, which can be written as

$$F_Y^{-1}(\tau) = \inf\{y \mid F_Y(y) \geq \tau\}.$$

The *pinball loss* (Koenker & Bassett, 1978), which is also known as the *check function*, is a widely used scoring rule. The pinball loss for an estimation $\hat{F}_Y$ of F_Y and a quantile level $\tau \in [0, 1]$ is defined as

$$\begin{aligned} & S_{\mathrm{Pinball}}(\hat{F}_Y, y; \tau) \\ &= \rho_\tau(\hat{F}_Y^{-1}(\tau), y) \\ &= \begin{cases} (1-\tau)(\hat{F}_Y^{-1}(\tau) - y) & \text{if } \hat{F}_Y^{-1}(\tau) \geq y, \\ \tau(y - \hat{F}_Y^{-1}(\tau)) & \text{if } \hat{F}_Y^{-1}(\tau) < y. \end{cases} \end{aligned} \tag{1}$$

Note that the pinball loss with $\tau = 0.5$ is equivalent to the mean absolute error (MAE), and it can be used to estimate the *median* (i.e., 0.5-th quantile) of Y. This means that the pinball loss is a generalization of MAE for any quantile level $\tau \in [0, 1]$. Note also that we include the quantile level τ in the notation $S_{\mathrm{Pinball}}(\hat{F}_Y^{-1}, y; \tau)$ to clarify that this scoring rule receives τ as an input.

It is known that the pinball loss is strictly proper (see e.g., (Gneiting & Raftery, 2007)), which means that we have

$$\mathop{\mathbb{E}}_{y \sim Y}[S_{\mathrm{Pinball}}(\hat{F}_Y, y; \tau)] \geq \mathop{\mathbb{E}}_{y \sim Y}[S_{\mathrm{Pinball}}(F_Y, y; \tau)],$$

and the equality holds only when $\hat{F}_Y^{-1}(\tau) = F_Y^{-1}(\tau)$ by Definition 4.2. Therefore, it is standard to use the pinball loss both for a loss function and an evaluation metric in quantile regression.

Portnoy's estimator (2003) is an extension of the pinball loss for quantile regression-based survival analysis, which is defined as

$$\begin{aligned} & S_{\mathrm{Portnoy}}(\hat{F}, (z, \delta); w, \tau) \\ &= \begin{cases} \rho_\tau(\hat{F}^{-1}(\tau), z) & \text{if } \delta = 1, \\ w\rho_\tau(\hat{F}^{-1}(\tau), z) & \\ \quad +(1-w)\rho_\tau(\hat{F}^{-1}(\tau), z_\infty) & \text{if } \delta = 0, \end{cases} \end{aligned} \tag{2}$$

where ρ_τ is the pinball loss defined in Eq. (1), w is a weight parameter to control the balance between two pinball loss terms, and z_∞ is any constant such that $z_\infty > z_{\max}$. In Portnoy's estimator, we can set an arbitrary constant $0 \leq w \leq 1$ for the parameter w if $\tau_c > \tau$, where $\tau_c = \Pr(t \leq c) = F(c)$, but we have to set $w = \Pr(F(c) < F(t) \leq \tau | t > c) = (\tau - \tau_c)/(1 - \tau_c)$ otherwise (i.e., $\tau_c \leq \tau$). Since we do not know the true value $\tau_c = F(c)$, we have to resolve this problem to use this estimator. Before showing how to resolve this problem, we prove that this estimator is proper under the condition that w is correct. Note that this is the first result for the quantile regression-based survival analysis in terms of the theory of scoring rules.

Theorem 4.5. *Portnoy's estimator is proper under the condition that w is correct.*

Proof. We give a proof in Appendix A.1. □

This theorem means that the crucial part of Portnoy's estimator is to set an appropriate value for w, and this theorem ensures that we can recover the true probability distribution F by minimizing Eq. (2) if w is correct.

Now, we emphasize that we cannot avoid the dependence on unknown parameters such as $F(c)$ in the definition of *any* of the scoring rules for survival analysis due to the discretization of $\hat{F}$. In the case of Portnoy's estimator, even if we know the true value $F^{-1}(\tau_i)$ for all $\{\tau_i\}_{i=0}^B$, we cannot compute $F(c)$ because c is not always contained in $\{F^{-1}(\tau_i)\}_{i=0}^B$. The best we can do is to find quantile levels τ_i and τ_{i+1} such that $F^{-1}(\tau_i) < c \leq F^{-1}(\tau_{i+1})$ by using the assumption that F is a monotonically increasing function. This means that $F(c)$ is between τ_i and τ_{i+1}. Even if we could find such τ_i and τ_{i+1}, we would not be able to calculate some important probabilities such as $\Pr(c < t \leq F^{-1}(\tau_{i+1})) = \tau_{i+1} - F(c)$. Therefore, we usually mitigate this problem by using a large B, which enables us to assume, for example, $F^{-1}(\tau_{i+1}) - F^{-1}(\tau_i) \approx 0$ for all i.

Even if we use a large B to assume that we can find the quantile level τ'_c such that $c \approx F^{-1}(\tau'_c)$ for any c, the problem that we do not know the true F^{-1} remains. One of the approaches to tackling this problem is the grid search algorithm (Portnoy, 2003; Neocleous et al., 2006). In this algorithm, we use a sufficiently large B, and we estimate $\hat{F}^{-1}(\tau_i)$ of $F^{-1}(\tau_i)$ in the increasing order of $i = 0, 1, \ldots, B$. Suppose that we have estimated $\{\hat{F}^{-1}(\tau_i)\}_{i=0}^{j-1}$ and we are going to estimate $\hat{F}^{-1}(\tau_j)$. The key idea of this algorithm is that we can find $\tau'_c \in \{\tau_i\}_{i=0}^{j-1}$ such that $c \approx \hat{F}^{-1}(\tau'_c)$ if $\tau_c = F(c) < \tau_j$. If we can find such τ'_c, we estimate w by using $\tau'_c \approx \tau_c$. If we cannot find such τ'_c, this algorithm assumes that $\tau_c > \tau_j$ and we use an arbitrary constant $0 \leq w \leq 1$. Portnoy (2003) discusses that this algorithm is analogous to the Kaplan-Meier estimator (Kaplan & Meier, 1958), and their theoretical

analysis (Portnoy, 2003; Neocleous et al., 2006) proves that the estimation model combining Portnoy's estimator, linear regression, and the grid search algorithm can recover the true probability distribution F if there is a sufficient number of data points.

As for another approach, Pearce et al. (2022) propose the CQRNN algorithm, which we call an *iterative reweighting (IR)* algorithm. Unlike the grid search algorithm, this algorithm estimates $\{\hat{F}^{-1}(\tau_i)\}_{i=0}^{B}$ simultaneously by using a neural network. This algorithm starts with an arbitrary initial estimation $\hat{F}$, and it estimates $\hat{w}$ of the true w by using $\hat{F}$. Then, it updates $\hat{F}$ by using $\hat{w}$, and it repeats this iterative procedure of estimating $\hat{F}$ and $\hat{w}$ until these values converge. This IR algorithm is similar to the expectation-maximization (EM) algorithm, and the relationship between this algorithm and the EM algorithm is discussed in (Pearce et al., 2022). Note that this IR algorithm can be implemented for "free" according to (Pearce et al., 2022), which means that we can implement it easily in the computation of the loss function of a neural network training algorithm, and we do not need to construct two separate neural network models for estimating $\hat{F}$ and $\hat{w}$. The experimental evaluation in (Pearce et al., 2022) shows that the IR algorithm performs the best among the quantile regression-based survival analysis models.

4.2. Extension of Logarithmic Score

While we estimate $\{\hat{F}_Y^{-1}(\tau_i)\}_{i=0}^{B}$ in quantile regression, we consider distribution regression, in which we estimate $\{\hat{F}_Y(\zeta_i)\}_{i=0}^{B}$. For distribution regression, the logarithmic score (Good, 1952) is known as a strictly proper scoring rule, and it is defined as

$$\begin{aligned} & S_{\mathrm{Log}}(\hat{F}_Y, y; \{\zeta_i\}_{i=0}^{B}) \\ &= -\sum_{i=0}^{B-1} \mathbb{1}(\zeta_i < y \le \zeta_{i+1}) \log(\hat{F}_Y(\zeta_{i+1}) - \hat{F}_Y(\zeta_i)) \\ &= -\sum_{i=0}^{B-1} \mathbb{1}(\zeta_i < y \le \zeta_{i+1}) \log \hat{f}_i, \end{aligned} \tag{3}$$

where $\hat{f}_i = \hat{F}_Y(\zeta_{i+1}) - \hat{F}_Y(\zeta_i)$ for $i = 0, 1, \ldots, B-1$.

We extend this logarithmic score for distribution regression-based survival analysis as

$$\begin{aligned} & S_{\mathrm{Cen-log}}(\hat{F}, (z, \delta); \{w_i\}_{i=0}^{B-1}, \{\zeta_i\}_{i=0}^{B}) \\ &= -\sum_{i=0}^{B-1} \mathbb{1}(\zeta_i < z \le \zeta_{i+1}) g(i, \delta, w_i), \end{aligned} \tag{4}$$

where

$$g(i, \delta, w_i) = \begin{cases} \log \hat{f}_i & \text{if } \delta = 1, \\ w_i \log \hat{f}_i + (1 - w_i) \log(1 - \hat{F}(\zeta_{i+1})) & \text{if } \delta = 0, \end{cases}$$

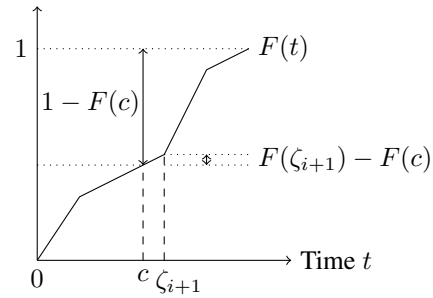


Figure 2. Illustration of computation of weight $w_i = (F(\zeta_{i+1}) - F(c))/(1 - F(c))$ for scoring rule $S_{\mathrm{Cen-log}}$.

$\hat{f}_i = \hat{F}(\zeta_{i+1}) - \hat{F}(\zeta_i)$, and $w_i = \Pr(c < t \le \zeta_{i+1} | t > c) = (F(\zeta_{i+1}) - F(c))/(1 - F(c))$. Note that this scoring rule is equivalent to Eq. (3) if $\delta = 1$. Similar to Portnoy's estimator, we cannot set the parameter w_i of this scoring rule because we do not know $F(\zeta_{i+1})$ and $F(c)$.

Even though we do not know the correct $\{w_i\}_{i=0}^{B-1}$, we prove that this scoring rule is proper if the set of parameters $\{w_i\}_{i=0}^{B-1}$ is correct.

Theorem 4.6. *$S_{\mathrm{Cen-Log}}(\hat{F}, (z, \delta); \{w_i\}_{i=0}^{B-1}, \{\zeta_i\}_{i=0}^{B})$ is a proper scoring rule under the condition that w_i is correct for all i.*

Proof. We give a proof in Appendix A.2. □

Similar to Portnoy's estimator, we can use both the grid-search algorithm and the IR algorithm to estimate $\{w_i\}_{i=0}^{B-1}$.

In addition, we show another simpler approach by assuming that $w_i \approx 0$ for all i if B is large. If B is large, $1 - F(c)$ is usually much larger than $F(\zeta_{i+1}) - F(c)$ (see Fig. 2), and hence, we have $w_i = (F(\zeta_{i+1}) - F(c))/(1 - F(c)) \approx 0$. Therefore, we can obtain a simpler variant of $S_{\mathrm{Cen-log}}$ by setting $w_i = 0$ for all i:

$$\begin{aligned} & S_{\mathrm{Cen-log-simple}}(\hat{F}, (z, \delta); \{\zeta_i\}_{i=0}^{B}) \\ &= -\sum_{i=0}^{B-1} \mathbb{1}(\zeta_i < z \le \zeta_{i+1}) g(i, \delta, 0) \\ &= -\delta \sum_{i=0}^{B-1} \mathbb{1}(\zeta_i < z \le \zeta_{i+1}) \log \hat{f}_i \\ &\quad -(1 - \delta) \sum_{i=0}^{B-1} \mathbb{1}(\zeta_i < z \le \zeta_{i+1}) \log(1 - \hat{F}(\zeta_{i+1})). \end{aligned} \tag{5}$$

Furthermore, by increasing B to infinity (i.e., $B \to \infty$), we obtain the continuous version of this scoring rule:

$$\begin{aligned} & S_{\mathrm{Cen-cont-log}}(\hat{F}, (z, \delta)) \\ &= -\delta \log \frac{d\hat{F}}{dt}(z) - (1 - \delta) \log(1 - \hat{F}(z)), \end{aligned} \tag{6}$$

which is equal to the extension of the logarithmic score that is proven to be strictly proper in (Rindt et al., 2022).

Remarks. This simplification clarifies that the proof in (Rindt et al., 2022) implicitly assumes that B is sufficiently large. This means that we should set B large enough in practice. Moreover, strictly speaking, the relation $w_i = (F(\zeta_{i+1}) - F(c))/(1 - F(c)) \approx 0$ may not hold if $1 - F(c) \approx 0$. Therefore, we recommend $S_{\text{Cen-log}}$ (Eq. (4)) rather than $S_{\text{Cen-log-simple}}$ (Eq. (5)) and $S_{\text{Cen-cont-log}}$ (Eq. (6)) whenever possible.

4.3. Extension of Brier Score

In distribution regression, the Brier score (Brier, 1950) is also known as a strictly proper scoring rule, which is defined as

$$S_{\text{Brier}}(\hat{F}_Y, y; \{\zeta_i\}_{i=0}^{B}) = \sum_{i=0}^{B-1} (\mathbb{1}(\zeta_i < y \leq \zeta_{i+1}) - \hat{f}_i)^2, \quad (7)$$

where $\hat{f}_i = \hat{F}_Y(\zeta_{i+1}) - \hat{F}_Y(\zeta_i)$ for $i = 0, 1, \ldots, B-1$.

We extend this Brier score for distribution regression-based survival analysis as

$$S_{\text{Cen-Brier}}(\hat{F}, (z, \delta); \{w_i\}_{i=0}^{B-1}, \{\zeta_i\}_{i=0}^{B}) = \sum_{i=0}^{B-1} \left(w_i (1 - \hat{f}_i)^2 + (1 - w_i) \hat{f}_i^2 \right), \quad (8)$$

where

$$w_i = \begin{cases} 0 & \text{if } \delta = 1 \text{ and } \zeta_{i+1} < z = t, \\ 1 & \text{if } \delta = 1 \text{ and } \zeta_i < z = t \leq \zeta_{i+1}, \\ 0 & \text{if } z \leq \zeta_i, \\ \frac{F(\zeta_{i+1}) - F(c)}{1 - F(c)} & \text{if } \delta = 0 \text{ and } \zeta_i < z = c \leq \zeta_{i+1}, \\ \frac{F(\zeta_{i+1}) - F(\zeta_i)}{1 - F(c)} & \text{if } \delta = 0 \text{ and } \zeta_{i+1} < z = c. \end{cases}$$

If $\delta = 1$, it is easy to see that Eq. (8) is equivalent to Eq. (7).

We prove that this scoring rule is proper if the set of parameters $\{w_i\}_{i=0}^{B-1}$ is correct.

Theorem 4.7. *$S_{\text{Cen-Brier}}(\hat{F}, (z, \delta); \{w_i\}_{i=0}^{B-1}, \{\zeta_i\}_{i=0}^{B})$ is a proper scoring rule under the condition that w_i is correct for all i.*

Proof. We give a proof in Appendix A.3. □

We can use the IR algorithm to estimate w_i. However, unlike Portnoy's estimator and the extension of the logarithmic score, we cannot use the grid-search algorithm in this extension of the Brier score because we need to estimate w_i for all $i = 0, 1, \ldots, B-1$.

Note that each w_i in this scoring rule is close to zero if B is large and $\delta = 0$. However, since w_is are designed to satisfy $\sum_i w_i = 1$, we cannot use the approximation $w_i \approx 0$ for this scoring rule.

4.4. Extension of Ranked Probability Score

The ranked probability score (RPS) is also known as a strictly proper scoring rule in distribution regression (see, e.g., (Gneiting & Raftery, 2007)). It is defined as

$$S_{\text{RPS}}(\hat{F}_Y, y) = \sum_{i=1}^{B-1} S_{\text{Binary-Brier}}(\hat{F}_Y, y; \zeta_i),$$

where $S_{\text{Binary-Brier}}$ is the binary version of S_{Brier} (Eq. (7)) with a single threshold ζ:

$$S_{\text{Binary-Brier}}(\hat{F}_Y, y; \zeta) = (\mathbb{1}(y \leq \zeta) - \hat{F}_Y(\zeta))^2.$$

We extend this scoring rule for survival analysis:

$$S_{\text{Cen-RPS}}(\hat{F}, (z, \delta); \{w_i\}_{i=1}^{B-1}, \{\zeta_i\}_{i=1}^{B-1}) = \sum_{i=1}^{B-1} S_{\text{Cen-Binary-Brier}}(\hat{F}, (z, \delta); w_i, \zeta_i), \quad (9)$$

where $S_{\text{Cen-Binary-Brier}}$ is the binary version of $S_{\text{Cen-Brier}}$ (Eq. (8)) with a single threshold ζ:

$$S_{\text{Cen-Binary-Brier}}(\hat{F}, (z, \delta); w, \zeta) = \begin{cases} \hat{F}(\zeta)^2 & \text{if } z > \zeta, \\ (1 - \hat{F}(\zeta))^2 & \text{if } \delta = 1 \text{ and } z = t \leq \zeta, \\ w(1 - \hat{F}(\zeta))^2 & \\ \quad + (1 - w)\hat{F}(\zeta)^2 & \text{if } \delta = 0 \text{ and } z = c \leq \zeta, \end{cases}$$

where $w = (F(\zeta) - F(c))/(1 - F(c))$.

Since this scoring rule is just the sum of the binary version of Brier scores for survival analysis, it is straightforward to prove this theorem.

Theorem 4.8. *$S_{\text{Cen-RPS}}(\hat{F}, (z, \delta); \{w_i\}_{i=1}^{B-1}, \{\zeta_i\}_{i=1}^{B-1})$ is a proper scoring rule under the condition that w_i is correct for all i.*

Note that the scoring rule $S_{\text{Cen-Binary-Brier}}$ is analogous to Portnoy's estimator. The scoring rule $S_{\text{Cen-Binary-Brier}}$ is designed to estimate $\hat{F}(\zeta)$, where ζ is an input, and we use $F(c)$ and ζ to set w, whereas Portnoy's estimator is designed to estimate $\hat{F}^{-1}(\tau)$, where τ is an input, and we use $F(c)$ and τ to set w. As these two scoring rules are similar, we can use both the grid-search algorithm and the IR algorithm for $S_{\text{Cen-RPS}}$.

Unlike $S_{\text{Cen-log}}$ defined in Eq. (4), the parameter w of the scoring rule $S_{\text{Cen-Binary-Brier}}$ is usually not close to zero because ζ and c are usually not close to each other as shown in Fig. 3. We note that the parameter w of Portnoy's estimator is also not close to zero for a similar reason.

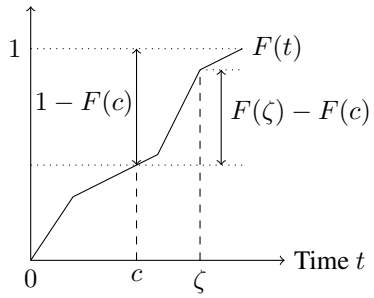


Figure 3. Illustration of computations of weight $w_i = (F(\zeta) - F(c))/(1 - F(c))$ for scoring rule $S_{\mathrm{Cen-Binary-Brier}}$.

5. Evaluation Metrics for Survival Analysis

While we have discussed the extensions of the scoring rules as loss functions, we should use strictly proper scoring rules also for evaluation metrics. However, among the extensions of the scoring rules for survival analysis, we can use only $S_{\mathrm{Cen-log-simple}}$ (Eq. (5)) as an evaluation metric because the other scoring rules depend on the parameter w or $\{w_i\}_{i=0}^{B-1}$. Note that we can use $S_{\mathrm{Cen-log-simple}}$ only when B is sufficiently large. In Appendix B, we conducted experiments on choosing an appropriate B, and the results suggested using $B > 16$.

While we can use $S_{\mathrm{Cen-log-simple}}$ as a discrimination metric for survival analysis, we note that there is a calibration metric, D-calibration (Haider et al., 2020), for survival analysis. D-calibration is widely used in survival analysis, but we propose another calibration metric, *KM-calibration*. Let $\kappa(t)$ be the survival function estimated by the Kaplan-Meier estimator (Kaplan & Meier, 1958). This function $\kappa(t)$ represents the survival rate (i.e., the probability that the event time is less than t) over the entire dataset rather than individual feature vector x. By definition, $\kappa(0) = 1$ and $\kappa(t)$ is a monotonically decreasing function. Assuming that $\kappa(t)$ is correct, $\kappa(t) = 1 - \hat{F}_{\mathrm{avg}}(t)$ must hold, where $\hat{F}_{\mathrm{avg}}(t)$ is the average of $\hat{F}(t)$ over all data points in the test dataset. Therefore, we define our KM-calibration as the Kullback-Leibler divergence between $\kappa(t)$ and $1 - \hat{F}_{\mathrm{avg}}(t)$:

$$\begin{aligned} d_{\mathrm{KM-cal}}(\kappa, \hat{F}_{\mathrm{avg}}) &= d_{\mathrm{KL}}(\kappa || 1 - \hat{F}_{\mathrm{avg}}) \\ &= \sum_{i=0}^{B-1} (p_i \log p_i - p_i \log q_i), \end{aligned}$$

where $p_i = \kappa(\zeta_{i+1}) - \kappa(\zeta_i)$, $q_i = (1 - \hat{F}_{\mathrm{avg}}(\zeta_{i+1})) - (1 - \hat{F}_{\mathrm{avg}}(\zeta_i))$, and we assume here that $\kappa(\zeta_B) = 0$. This metric is based on the observation that the model's predicted number of events within any time interval should be similar to the observed number (Goldstein et al., 2020). We note that there is another calibration metric (Chapfuwa et al., 2020) based on the Kaplan-Meier estimator. Whereas this calibration metric uses the absolute difference, our KM-calibration uses the Kullback-Leibler divergence.

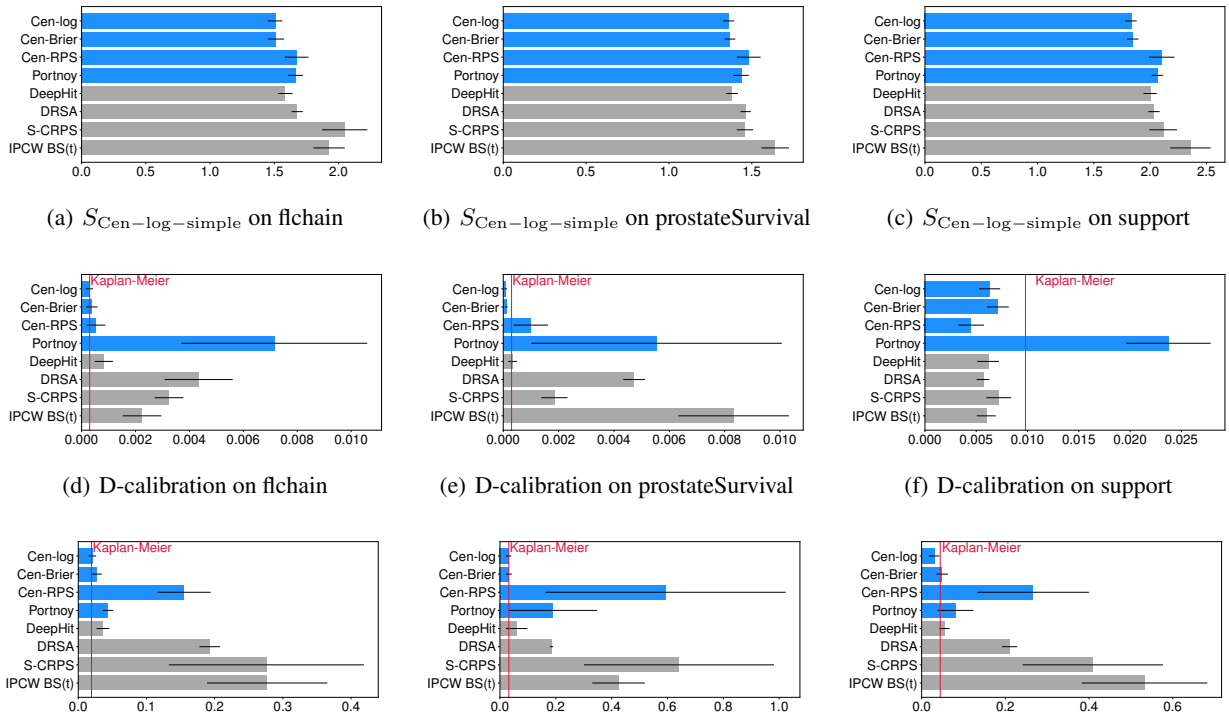
6. Experiments

In our experiments, we compared practical prediction performances of various loss functions on real datasets. We used three datasets for the survival analysis from the packages in R (R Core Team, 2016): the flchain dataset (Dispenzieri et al., 2012), which was obtained from the "survival" package and contains 7874 data points (69.9% of which are censored), the prostateSurvival dataset (Lu-Yao et al., 2009), which was obtained from the "asaur" package and contains 14294 data points (71.7% of which are censored), and the support dataset (Knaus et al., 1995), which was obtained from the "casebase" package and contains 9104 data points (31.9% of which are censored). For each dataset, we divided the time interval $[0, z_{\max} + \epsilon)$, where $\epsilon = 10^{-3}$, into B equal-length intervals to get the thresholds $\{\zeta_i\}_{i=0}^{B}$ for distribution regression-based survival analysis, and we divided the unit interval $[0, 1]$ into B equal-length intervals to get the quantile levels $\{\tau_i\}_{i=0}^{B}$ for quantile regression-based survival analysis. Unless otherwise stated, we set $B = 32$.

All our experiments were conducted on a virtual machine with an Intel Xeon CPU (3.30 GHz) processor without any GPU and 64 GB of memory running Red Hat Enterprise Linux Server 7.6. We used Python 3.7.4 and PyTorch 1.7.1 for the implementation.

We estimated $\hat{F}(t)$ by combining a multi-layer perceptron (MLP) and the IR algorithm (see Sec. 4.1) to estimate w or $\{w_i\}_{i=0}^{B-1}$. The MLP consists of three hidden layers containing 128 neurons, and the number of outputs was B. The type of activation function after the hidden layer was the rectified linear unit (ReLU), and the activation function at the output node was softmax. The softmax function is used to satisfy the assumption that $\hat{F}(t)$ is a monotonically increasing continuous function. In distribution regression-based survival analysis, each output of MLP estimates $\hat{f}_i = \hat{F}(\zeta_{i+1}) - \hat{F}(\zeta_i)$ for $i = 0, 1, \ldots, B-1$. By using these outputs $\{\hat{f}_i\}_{i=0}^{B-1}$, we can calculate $\{\hat{F}(\zeta_i)\}_{i=0}^{B}$ and we can represent the function $\hat{F}(t)$ as a piecewise linear function connecting the values $\{\hat{F}(\zeta_i)\}_{i=0}^{B}$. Since $\hat{f}_i > 0$ holds for all i, $\hat{F}(t)$ estimated in this way is a monotonically increasing continuous function. We can estimate $\hat{F}$ for quantile regression-based survival analysis by using a similar way.

For the training of the neural network, we used the Adam optimizer (Kingma & Ba, 2015) with the learning rate 0.001, and the other parameters were set to their default values. We ran training for 300 epochs for our neural network models. Our implementation of the scoring rules are available at `https://github.com/IBM/dqs`.



(a) $S_{\text{Cen-log-simple}}$ on flchain (b) $S_{\text{Cen-log-simple}}$ on prostateSurvival (c) $S_{\text{Cen-log-simple}}$ on support

(d) D-calibration on flchain (e) D-calibration on prostateSurvival (f) D-calibration on support

(g) KM-calibration on flchain (h) KM-calibration on prostateSurvival (i) KM-calibration on support

Figure 4. Prediction performance (lower is better) comparison on three datasets with $S_{\text{Cen-log-simple}}$, KM-calibration, and D-calibration.

We compared the prediction performances of various scoring rules (i.e., loss functions), and Fig. 4 shows the results. In these experiments, we split the data points into training (60%), validation (20%), and test (20%), and each bar shows the mean of the measurements on the test data of five random splits together with the error bar, which represents the standard deviation. We used $S_{\text{Cen-log-simple}}$ (Eq. (5)) as a metric for discrimination performance and D-calibration (Haider et al., 2020) and KM-calibration (see Sec. 5) as calibration metrics, where we used 20 bins of equal length for D-calibration. For the calibration metrics, we added the mean D-calibration and mean KM-calibration of the Kaplan-Meier estimator (Kaplan & Meier, 1958) as a red line in each graph. Since the Kaplan-Meier estimator is calibrated in theory, the values of the D-calibration and the KM-calibration of this estimator should be regarded as close to zeros. In this figure, the four scoring rules Cen-log ($S_{\text{Cen-log}}$ defined in Eq. (4)), Cen-Brier ($S_{\text{Cen-Brier}}$ defined in Eq. (8)), Cen-RPS ($S_{\text{Cen-RPS}}$ defined in Eq. (9)), and Portnoy (S_{Portnoy} defined in Eq. (2)) are proved to be conditionally proper in this paper. Note that Cen-log is similar to the scoring rule (Eq. (6)) that is proved to be strictly proper in (Rindt et al., 2022) and Portnoy is proposed in (Portnoy, 2003). This figure also contains the results for other scoring rules in the state-of-the-art models for survival analysis: DeepHit (Lee et al., 2018) with parameter $\alpha = 1$, DRSA (Ren et al., 2019) with parameter $\alpha = 0.25$, S-CRPS (Avati et al., 2019), and IPCW BS(t) game model (Han et al., 2021). These four scoring rules are not proved to be proper.

Figure 4 shows that the prediction performances of the four extended scoring rules (Cen-log, Cen-Brier, Cen-RPS, and Portnoy) were not similar, even though we prove that these four scoring rules are conditionally proper and the outputs are expected to be similar if the parameters $\hat{w}$ and $\{\hat{w}_i\}_{i=0}^{B-1}$ are correct. The scoring rules Cen-log and Cen-Brier outperformed the scoring rules Cen-RPS and Portnoy in discrimination performance $S_{\text{Cen-log-simple}}$. These results indicate that the accuracy of the estimated parameters $\hat{w}$ and $\{\hat{w}_i\}_{i=0}^{B-1}$ by the IR algorithm are important when we use these scoring rules in practice. The major difference between these scoring rules are that, whereas the set of parameters $\{w_i\}_{i=0}^{B-1}$ in Cen-log and Cen-Brier usually satisfies $w_i \approx 0$ or $w_i = 1$, the set of parameters $\{w_i\}_{i=0}^{B}$ in Cen-RPS can take an arbitrary value $0 \leq w_i \leq 1$. The parameter w in Portnoy can also take an arbitrary value $0 \leq w \leq 1$. Therefore, Cen-log and Cen-Brier seem less sensitive to the accuracy of the parameters than Cen-RPS and Portnoy. This figure also shows that the other four scoring rules (DeepHit, DRSA, S-CRPS, and IPCW BS(t)) performed worse than Cen-log and Cen-Brier. Note that IPCW BS(t) model is

Table 1. Prediction performances of DeepHit (lower is better) with various α for $B = 32$.

Metric	Model	flchain	prostateSurvival	support
$S_{\mathrm{Cen-log-simple}}$	DeepHit ($\alpha = 0$)	1.5059 ± 0.0513	1.3609 ± 0.0301	1.8296 ± 0.0446
	DeepHit ($\alpha = 0.1$)	1.5200 ± 0.0398	1.3644 ± 0.0293	1.8481 ± 0.0453
	DeepHit ($\alpha = 1$)	1.5858 ± 0.0495	1.3813 ± 0.0318	1.9996 ± 0.0525
	DeepHit ($\alpha = 10$)	2.0313 ± 0.1648	1.5688 ± 0.0823	2.3657 ± 0.0441
D-calibration	DeepHit ($\alpha = 0$)	0.0003 ± 0.0001	0.0001 ± 0.0000	0.0062 ± 0.0012
	DeepHit ($\alpha = 0.1$)	0.0005 ± 0.0002	0.0001 ± 0.0000	0.0056 ± 0.0009
	DeepHit ($\alpha = 1$)	0.0008 ± 0.0003	0.0003 ± 0.0001	0.0062 ± 0.0010
	DeepHit ($\alpha = 10$)	0.0138 ± 0.0046	0.0064 ± 0.0035	0.0179 ± 0.0053
KM-calibration	DeepHit ($\alpha = 0$)	0.0213 ± 0.0049	0.0343 ± 0.0102	0.0288 ± 0.0127
	DeepHit ($\alpha = 0.1$)	0.0264 ± 0.0071	0.0418 ± 0.0139	0.0249 ± 0.0067
	DeepHit ($\alpha = 1$)	0.0362 ± 0.0084	0.0599 ± 0.0341	0.0545 ± 0.0110
	DeepHit ($\alpha = 10$)	0.2077 ± 0.0543	0.4937 ± 0.1772	0.4273 ± 0.1188

similar to the IR algorithm in that both of the algorithms are used to estimate unknown parameters, but the loss function of IPCW BS(t) model is not proved to be proper in terms of the theory of scoring rules. With respect to the calibration metrics, Cen-log and Cen-Brier showed comparable performance with the Kaplan-Meier estimator. However, the other scoring rules showed worse calibration performances for at least one of D-calibration and KM-calibration.

Regarding the parameter α of DeepHit (Lee et al., 2018), we conducted additional experiments by changing this parameter. The loss function of DeepHit consists of two terms. The first term is equal to the extension of the logarithmic score $S_{\mathrm{Cen-log-simple}}$, and the second term is used to improve a ranking metric (i.e., a variant of C-index). The parameter α is used to control the balance between these two terms, and the weight for the second term is increased by using a large α. Note that the scoring rule $S_{\mathrm{Cen-log-simple}}$ is equivalent to DeepHit with $\alpha = 0$. Table 1 shows the results for $\alpha \in \{0, 0.1, 1, 10\}$. This table shows that the prediction performances of DeepHit became worse as α increases. This means that we should set $\alpha = 0$ when we use DeepHit.

7. Conclusion

We discussed extensions of four scoring rules for survival analysis, and we proved that these extensions are proper if the parameter w or $\{w_i\}_{i=0}^{B-1}$ is correct. These proofs reduce the problem of estimating $\hat{F}$ to the problem of estimating the parameter w or $\{w_i\}_{i=0}^{B-1}$ in proper scoring rules. We also demonstrated that the models with $S_{\mathrm{Cen-log}}$ and $S_{\mathrm{Cen-Brier}}$ as loss functions performed the best in our experiments. These results indicate that it is better to use a proper scoring rule that has low sensitivity on the parameter. In addition, we clarified the hidden assumption in the proof of the properness for $S_{\mathrm{Cen-cont-log}}$ (Rindt et al., 2022). This suggests us to use a sufficiently large B when we use it, and we demonstrated that such B can be found by comparing the prediction performances of $S_{\mathrm{Cen-log-simple}}$ and $S_{\mathrm{Cen-log}}$ with various B.

References

Antolini, L., Boracchi, P., and Biganzoli, E. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005.

Avati, A., Duan, T., Zhou, S., Jung, K., Shah, N. H., and Ng, A. Y. Countdown regression: Sharp and calibrated survival predictions. In *Proceedings of UAI 2019*, pp. 145–155, 2019.

Benedetti, R. Scoring rules for forecast verification. *American Meteorological Society*, 138(1):203–211, 2010.

Bengs, V., Hüllermeier, E., and Waegeman, W. Pitfalls of epistemic uncertainty quantification through loss minimisation. In *Proceedings of NeurIPS 2022*, 2022.

Blanche, P., Kattan, M. W., and Gerds, T. A. The c-index is not proper for the evaluation of t-year predicted risks. *Biostatistics*, 20(2):347–357, 2018.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Chapfuwa, P., Tao, C., Li, C., Khan, I., Chandross, K. J., Pencina, M. J., Carin, L., and Henao, R. Calibration and uncertainty in neural time-to-event modeling. *IEEE Transactions on Neural Networks and Learning Systems*, 34(4):1666–1680, 2020.

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

Dirick, L., Claeskens, G., and Baesens, B. Time to default in credit scoring using survival analysis: a benchmark study. *Journal of the Operational Research Society*, 68 (6):652–665, 2017.

Dispenzieri, A., Katzmann, J. A., Kyle, R. A., Larson, D. R., Therneau, T. M., Colby, C. L., Clark, R. J., Mead, G. P., Kumar, S., III, L. J. M., and Rajkumar, S. V. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. *Mayo Clinic Proceedings*, 87(6):517–523, 2012.

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

Goldstein, M., Han, X., Puli, A. M., Perotte, A., and Ranganath, R. X-CAL: Explicit calibration for survival analysis. In *Proceedings of NeurIPS 2020*, pp. 18296–18307, 2020.

Good, I. J. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1):107–114, 1952.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18 (17–18):2529–2545, 1999.

Haider, H., Hoehn, B., Davis, S., and Greiner, R. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 21(85): 1–63, 2020.

Han, X., Goldstein, M., Puli, A., Wies, T., Perotte, A., and Ranganath, R. Inverse-weighted survival games. In *Proceedings of NeurIPS 2021*, pp. 2160–2172, 2021.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247(18):2543–2546, 1982.

Kamran, F. and Wiens, J. Estimating calibrated individualized survival curves with deep learning. In *Proceedings of AAAI 2021*, pp. 240–248, 2021.

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

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *Proceedings of ICLR 2015*, 2015.

Knaus, W. A., Harrell, Jr., F. E., Lynn, J., Goldman, L., Phillips, R. S., Connors, Jr., A. F., Dawson, N. V., Fulkerson, Jr., W. J., Califf, R. M., Desbiens, N., Layde, P., Oye, R. K., Bellamy, P. E., Hakim, R. B., and Wagner, D. P. The SUPPORT prognostic model. Objective estimates of survival for seriously ill hospitalized adults. Study to understand prognoses and preferences for outcomes and risks of treatments. *Annals of Internal Medicine*, 122(3): 191–203, 1995.

Koenker, R. and Bassett, Jr., B. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

Koenker, R. and Hallock, K. F. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001.

Kvamme, H., Borgan, O., and Scheel, I. Time-to-event prediction with neural networks and Cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019.

Lee, C., Zame, W. R., Yoon, J., and van der Schaar, M. DeepHit: A deep learning approach to survival analysis with competing risks. In *Proceedings of AAAI-18*, pp. 2314–2321, 2018.

Lu-Yao, G. L., Albertsen, P. C., Moore, D. F., Shih, W., Lin, Y., DiPaola, R. S., Barry, M. J., Zietman, A., O'Leary, M., Walker-Corkery, E., and Yao, S.-L. Outcomes of localized prostate cancer following conservative management. *Journal of the American Medical Association*, 302(11): 1202–1209, 2009.

Mura, A., Galavotti, M., Hykel, H., and de Finetti, B. *Philosophical Lectures on Probability: collected, edited, and annotated by Alberto Mura*. Synthese Library. Springer Netherlands, 2008.

Neocleous, T., Branden, K. V., and Portnoy, S. Correction to censored regression quantiles by S. Portnoy, 98 (2003), 1001–1012. *Journal of the American Statistical Association*, 101(474):860–861, 2006.

Parmigiani, G. and Inoue, L. *Decision Theory: Principles and Approaches*. Wiley Series in Probability and Statistics. Wiley, 2009.

Pearce, T., Jeong, J.-H., Jia, Y., and Zhu, J. Censored quantile regression neural networks. In *Proceedings of NeurIPS 2022*, 2022.

Peng, L. Quantile regression for survival data. *Annual Review of Statistics and Its Application*, 8:413–437, 2021.

Portnoy, S. Censored regression quantiles. *Journal of the American Statistical Association*, 98(464):1001–1012, 2003.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/.

Ren, K., Qin, J., Zheng, L., Yang, Z., Zhang, W., Qiu, L., and Yu, Y. Deep recurrent survival analysis. In *Proceedings of AAAI-19*, pp. 4798–4805, 2019.

Rindt, D., Hu, R., Steinsaltz, D., and Sejdinovic, D. Survival regression with proper scoring rules and monotonic neural networks. In *Proceedings of AISTATS 2022*, 2022.

Schlag, K. H., Tremewan, J., and van der Weele, J. J. A penny for your thoughts: a survey of methods for eliciting beliefs. *Experimental Economics*, 18:457–490, 2015.

Sonabend, R., Bender, A., and Vollmer, S. Avoiding c-hacking when evaluating survival distribution predictions with discrimination measures. *Bioinformatics*, 38(17): 4178–4184, 2022.

Tjandra, D. E., He, Y., and Wiens, J. A hierarchical approach to multi-event survival analysis. In *Proceedings of AAAI 2021*, pp. 591–599, 2021.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117, 2011.

Wang, P., Li, Y., and Reddy, C. K. Machine learning for survival analysis: A survey. *ACM Computing Surveys*, 51 (6):1–36, 2019.

Zheng, P., Yuan, S., and Wu, X. SAFE: A neural survival analysis model for fraud early detection. In *Proceedings of AAAI-19*, pp. 1278–1285, 2019.

Zhong, Q., Mueller, J., and Wang, J.-L. Deep extended hazard models for survival analysis. In *Proceedings of NeurIPS 2021*, 2021.

A. Proofs of Theorems

We give proofs of the theorems, which are omitted from the main body of this paper.

A.1. Portnoy's Estimator

We show a proof of Theorem 4.5.

Proof. We consider a fixed $c \sim C$, and we prove

$$\mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\text{Portnoy}}(\hat{F},(z,\delta);w,\tau)] \geq \mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\text{Portnoy}}(F,(z,\delta);w,\tau)] \tag{10}$$

for these four cases separately.

- Case 1: $c \leq \min\{F^{-1}(\tau), \hat{F}^{-1}(\tau)\}$.
- Case 2: $\max\{F^{-1}(\tau), \hat{F}^{-1}(\tau)\} < c$.
- Case 3: $F^{-1}(\tau) < c \leq \hat{F}^{-1}(\tau)$.
- Case 4: $\hat{F}^{-1}(\tau) < c \leq F^{-1}(\tau)$.

Note that, if Inequality (10) holds for any $c \sim C$, we can marginalize the inequality with respect to C, and we have

$$\mathop{\mathbb{E}}_{t\sim T,c\sim C}[S_{\text{Portnoy}}(\hat{F},(z,\delta);w,\tau)] \geq \mathop{\mathbb{E}}_{t\sim T,c\sim C}[S_{\text{Portnoy}}(F,(z,\delta);w,\tau)],$$

which means that $S_{\text{Portnoy}}(\hat{F},(z,\delta);w,\tau)$ is proper. Therefore, we prove Inequality (10) for the four cases.

Case 1. We prove the case for $c \leq \min\{F^{-1}(\tau), \hat{F}^{-1}(\tau)\}$. This means that $\tau_c \leq \tau$ and $w = (\tau - \tau_c)/(1-\tau_c)$. Hence, we have

$$\begin{aligned}
S_{\text{Portnoy}}(\hat{F},(z,\delta);w,\tau) &= \begin{cases} \rho_\tau(\hat{F}^{-1}(\tau),t) & \text{if } t \leq c, \\ w\rho_\tau(\hat{F}^{-1}(\tau),c) + (1-w)\rho_\tau(\hat{F}^{-1}(\tau),z_\infty) & \text{if } t > c, \end{cases} \\
&= \begin{cases} (1-\tau)(\hat{F}^{-1}(\tau)-t) & \text{if } t \leq c, \\ \frac{\tau-\tau_c}{1-\tau_c}(1-\tau)(\hat{F}^{-1}(\tau)-c) + \frac{1-\tau}{1-\tau_c}\tau(z_\infty - \hat{F}^{-1}(\tau)) & \text{if } t > c, \end{cases} \\
&= \begin{cases} (1-\tau)(\hat{F}^{-1}(\tau)-t) & \text{if } t \leq c, \\ \frac{-\tau_c(1-\tau)}{1-\tau_c}\hat{F}^{-1}(\tau) - \frac{(\tau-\tau_c)(1-\tau)}{1-\tau_c}c + \frac{(1-\tau)\tau}{1-\tau_c}z_\infty & \text{if } t > c. \end{cases}
\end{aligned}$$

By Assumption 3.1, we have $\Pr(t \leq c|C=c) = \Pr(t \leq c) = \tau_c$ and $\Pr(t > c|C=c) = 1-\tau_c$. Hence, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\text{Portnoy}}(\hat{F},(z,\delta);w,\tau)] &= \Pr(t \leq c|C=c)(1-\tau)\hat{F}^{-1}(\tau) - (1-\tau)\mathop{\mathbb{E}}_{t\sim T|C=c,t\leq c}[t] \\
&\quad -\Pr(t > c|C=c)\frac{\tau_c(1-\tau)}{1-\tau_c}\hat{F}^{-1}(\tau) - \frac{(\tau-\tau_c)(1-\tau)}{1-\tau_c}c + \frac{(1-\tau)\tau}{1-\tau_c}z_\infty \\
&= -(1-\tau)\mathop{\mathbb{E}}_{t\sim T|C=c,t\leq c}[t] - \frac{(\tau-\tau_c)(1-\tau)}{1-\tau_c}c + \frac{(1-\tau)\tau}{1-\tau_c}z_\infty.
\end{aligned}$$

Since this value is the same for $S_{\text{Portnoy}}(\hat{F},(z,\delta);w,\tau)$ and $S_{\text{Portnoy}}(F,(z,\delta);w,\tau)$, we have

$$\mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\text{Portnoy}}(\hat{F},(z,\delta);w,\tau)] = \mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\text{Portnoy}}(F,(z,\delta);w,\tau)].$$

Case 2. We prove the case for $\max\{F^{-1}(\tau), \hat{F}^{-1}(\tau)\} < c$.

If $F^{-1}(\tau) \leq \hat{F}^{-1}(\tau) < c$, then we have

$$\begin{aligned}
S_{\text{Portnoy}}(\hat{F}, (z, \delta); w, \tau) &= \begin{cases} \rho_\tau(\hat{F}^{-1}(\tau), t) & \text{if } t \leq c, \\ w\rho_\tau(\hat{F}^{-1}(\tau), c) + (1-w)\rho_\tau(\hat{F}^{-1}(\tau), z_\infty) & \text{if } t > c, \end{cases} \\
&= \begin{cases} (1-\tau)(\hat{F}^{-1}(\tau) - t) & \text{if } t \leq \hat{F}^{-1}(\tau), \\ -\tau(\hat{F}^{-1}(\tau) - t) & \text{if } \hat{F}^{-1}(\tau) < t \leq c, \\ -w\tau(\hat{F}^{-1}(\tau) - c) - (1-w)\tau(\hat{F}^{-1}(\tau) - z_\infty) & \text{if } t > c, \end{cases} \\
&\geq \begin{cases} (1-\tau)(\hat{F}^{-1}(\tau) - t) & \text{if } t \leq F^{-1}(\tau), \\ -\tau(\hat{F}^{-1}(\tau) - t) & \text{if } F^{-1}(\tau) < t \leq c, \\ -\tau\hat{F}^{-1}(\tau) + w\tau c + (1-w)z_\infty & \text{if } t > c, \end{cases}
\end{aligned}$$

where we used $-\tau(\hat{F}^{-1}(\tau) - t) \leq (1-\tau)(\hat{F}^{-1}(\tau) - t)$ when $F^{-1}(\tau) < t \leq \hat{F}^{-1}(\tau)$ for the inequality.

If $\hat{F}^{-1}(\tau) \leq F^{-1}(\tau) < c$, then we have

$$\begin{aligned}
S_{\text{Portnoy}}(\hat{F}, (z, \delta); w, \tau) &= \begin{cases} \rho_\tau(\hat{F}^{-1}(\tau), t) & \text{if } t \leq c, \\ w\rho_\tau(\hat{F}^{-1}(\tau), c) + (1-w)\rho_\tau(\hat{F}^{-1}(\tau), z_\infty) & \text{if } t > c, \end{cases} \\
&= \begin{cases} (1-\tau)(\hat{F}^{-1}(\tau) - t) & \text{if } t \leq \hat{F}^{-1}(\tau), \\ -\tau(\hat{F}^{-1}(\tau) - t) & \text{if } \hat{F}^{-1}(\tau) < t \leq c, \\ -w\tau(\hat{F}^{-1}(\tau) - c) - (1-w)\tau(\hat{F}^{-1}(\tau) - z_\infty) & \text{if } t > c, \end{cases} \\
&> \begin{cases} (1-\tau)(\hat{F}^{-1}(\tau) - t) & \text{if } t \leq F^{-1}(\tau), \\ -\tau(\hat{F}^{-1}(\tau) - t) & \text{if } F^{-1}(\tau) < t \leq c, \\ -\tau\hat{F}^{-1}(\tau) + w\tau c + (1-w)z_\infty & \text{if } t > c, \end{cases}
\end{aligned}$$

where we used $-\tau(\hat{F}^{-1}(\tau) - t) > (1-\tau)(\hat{F}^{-1}(\tau) - t)$ when $\hat{F}^{-1}(\tau) < t \leq F^{-1}(\tau)$ for the inequality.

By Assumption 3.1, we have $\Pr(t \leq F^{-1}(\tau)|C = c) = \Pr(t \leq F^{-1}(\tau)) = \tau$, $\Pr(F^{-1}(\tau) < t|C = c) = 1 - \tau$, and $\Pr(c < t|C = c) = 1 - \tau_c$. Hence, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\text{Portnoy}}(\hat{F}, (z, \delta); w, \tau)] &\geq \Pr(t \leq F^{-1}(\tau)|C = c)(1-\tau)\hat{F}^{-1}(\tau) - (1-\tau)\mathop{\mathbb{E}}_{t\sim T|C=c, t\leq F^{-1}(\tau)}[t] \\
&\quad -\Pr(F^{-1}(\tau) < t|C = c)\tau\hat{F}^{-1}(\tau) \\
&\quad +\Pr(c < t|C = c)(w\tau c + (1-w)z_\infty) \\
&= -(1-\tau)\mathop{\mathbb{E}}_{t\sim T|C=c, t\leq F^{-1}(\tau)}[t] + (1-\tau_c)(w\tau c + (1-w)z_\infty).
\end{aligned}$$

By using a similar argument, we have

$$\mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\text{Portnoy}}(F, (z, \delta); w, \tau)] = -(1-\tau)\mathop{\mathbb{E}}_{t\sim T|C=c, t\leq F^{-1}(\tau)}[t] + (1-\tau_c)(w\tau c + (1-w)z_\infty).$$

Note that this equation holds with equality.

Hence, we have

$$\mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\text{Portnoy}}(\hat{F}, (z, \delta); w, \tau)] \geq \mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\text{Portnoy}}(F, (z, \delta); w, \tau)].$$

Case 3. We prove the case for $F^{-1}(\tau) < c \leq \hat{F}^{-1}(\tau)$.

We have

$$
\begin{aligned}
S_{\text{Portnoy}}(\hat{F},(z,\delta);w,\tau) &= \begin{cases} \rho_\tau(\hat{F}^{-1}(\tau),t) & \text{if } t \le c, \\ w\rho_\tau(\hat{F}^{-1}(\tau),c) + (1-w)\rho_\tau(\hat{F}^{-1}(\tau),z_\infty) & \text{if } t > c, \end{cases} \\
&= \begin{cases} (1-\tau)(\hat{F}^{-1}(\tau)-t) & \text{if } t \le c, \\ w(1-\tau)(\hat{F}^{-1}(\tau)-c) - (1-w)\tau(\hat{F}^{-1}(\tau)-z_\infty) & \text{if } t > c, \end{cases} \\
&\ge \begin{cases} (1-\tau)(\hat{F}^{-1}(\tau)-t) & \text{if } t \le F^{-1}(\tau), \\ -\tau(\hat{F}^{-1}(\tau)-t) & \text{if } F^{-1}(\tau) < t \le c, \\ -w\tau(\hat{F}^{-1}(\tau)-c) - (1-w)\tau(\hat{F}^{-1}(\tau)-z_\infty) & \text{if } t > c, \end{cases}
\end{aligned}
$$

where we used $(1-\tau)(\hat{F}^{-1}(\tau)-t) \ge -\tau(\hat{F}^{-1}(\tau)-t)$ when $F^{-1}(\tau) < t \le c$ and $w(1-\tau)(\hat{F}^{-1}(\tau)-c) \ge -w\tau(\hat{F}^{-1}(\tau)-c)$ when $t > c$.

By using a similar argument, we have

$$
\begin{aligned}
S_{\text{Portnoy}}(F,(z,\delta);w,\tau) &= \begin{cases} \rho_\tau(F^{-1}(\tau),t) & \text{if } t \le c, \\ w\rho_\tau(F^{-1}(\tau),c) + (1-w)\rho_\tau(F^{-1}(\tau),z_\infty) & \text{if } t > c, \end{cases} \\
&= \begin{cases} (1-\tau)(F^{-1}(\tau)-t) & \text{if } t \le F^{-1}(\tau), \\ -\tau(F^{-1}(\tau)-t) & \text{if } F^{-1}(\tau) < t \le c, \\ -w\tau(F^{-1}(\tau)-c) - (1-w)\tau(F^{-1}(\tau)-z_\infty) & \text{if } t > c, \end{cases}
\end{aligned}
$$

Note that this equation holds with equality.

Hence, we have

$$
\mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\text{Portnoy}}(\hat{F},(z,\delta);w,\tau)] \ge \mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\text{Portnoy}}(F,(z,\delta);w,\tau)].
$$

Case 4. We prove the case for $\hat{F}^{-1}(\tau) < c \le F^{-1}(\tau)$.

Regarding $\hat{F}$, we have

$$
\begin{aligned}
S_{\text{Portnoy}}(\hat{F},(z,\delta);w,\tau) &= \begin{cases} \rho_\tau(\hat{F}^{-1}(\tau),t) & \text{if } t \le c, \\ w\rho_\tau(\hat{F}^{-1}(\tau),c) + (1-w)\rho_\tau(\hat{F}^{-1}(\tau),z_\infty) & \text{if } t > c, \end{cases} \\
&= \begin{cases} (1-\tau)(\hat{F}^{-1}(\tau)-t) & \text{if } t \le \hat{F}^{-1}(\tau), \\ -\tau(\hat{F}^{-1}(\tau)-t) & \text{if } \hat{F}^{-1}(\tau) < t \le c, \\ -w\tau(\hat{F}^{-1}(\tau)-c) - (1-w)\tau(\hat{F}^{-1}(\tau)-z_\infty) & \text{if } t > c, \end{cases} \\
&> \begin{cases} (1-\tau)(\hat{F}^{-1}(\tau)-t) & \text{if } t \le c, \\ -\tau\hat{F}^{-1}(\tau) + w\tau c + (1-w)\tau z_\infty & \text{if } t > c, \end{cases}
\end{aligned}
$$

where we used $-\tau(\hat{F}^{-1}(\tau)-t) > (1-\tau)(\hat{F}^{-1}(\tau)-t)$ when $\hat{F}^{-1}(\tau) < t \le c$ for the inequality. By Assumption 3.1, we have $\Pr(t \le c|C=c) = \Pr(t \le c) = \tau_c$ and $\Pr(t > c|C=c) = 1-\tau_c$. Hence, we have

$$
\begin{aligned}
\mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\text{Portnoy}}(\hat{F},(z,\delta);w,\tau)] &> \Pr(t \le c|C=c)(1-\tau)\hat{F}^{-1}(\tau) - (1-\tau)\mathop{\mathbb{E}}_{t\sim T|C=c,t\le c}[t] \\
&\quad + \Pr(t > c|C=c)(-\tau\hat{F}^{-1}(\tau) + w\tau c + (1-w)\tau z_\infty) \\
&> \tau_c(1-\tau)\hat{F}^{-1}(\tau) - (1-\tau)\mathop{\mathbb{E}}_{t\sim T|C=c,t\le c}[t] \\
&\quad -(1-\tau_c)\tau\hat{F}^{-1}(\tau) + (1-\tau_c)(w\tau c + (1-w)\tau z_\infty) \\
&> (\tau_c-\tau)\hat{F}^{-1}(\tau) - (1-\tau)\mathop{\mathbb{E}}_{t\sim T|C=c,t\le c}[t] + (1-\tau_c)(w\tau c + (1-w)\tau z_\infty).
\end{aligned}
$$

Regarding F, since $w = (\tau - \tau_c)/(1 - \tau_c)$, we have

$$\begin{aligned}
S_{\mathrm{Portnoy}}(F, (z, \delta); w, \tau) &= \begin{cases} \rho_\tau(F^{-1}(\tau), t) & \text{if } t \leq c, \\ w\rho_\tau(F^{-1}(\tau), c) + (1 - w)\rho_\tau(F^{-1}(\tau), z_\infty) & \text{if } t > c, \end{cases} \\
&= \begin{cases} (1 - \tau)(F^{-1}(\tau) - t) & \text{if } t \leq c, \\ w(1 - \tau)(F^{-1}(\tau) - c) - (1 - w)\tau(F^{-1}(\tau) - z_\infty) & \text{if } t > c, \end{cases} \\
&= \begin{cases} (1 - \tau)(F^{-1}(\tau) - t) & \text{if } t \leq c, \\ -\frac{\tau_c(1-\tau)}{1-\tau_c} F^{-1}(\tau) - w(1 - \tau)c + (1 - w)\tau z_\infty & \text{if } t > c, \end{cases}
\end{aligned}$$

By Assumption 3.1, we have $\Pr(t \leq c|C = c) = \Pr(t \leq c) = \tau_c$ and $\Pr(t > c|C = c) = 1 - \tau_c$. Hence, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\mathrm{Portnoy}}(\hat{F}, (z, \delta); w, \tau)] &= \Pr(t \leq c|C = c)(1 - \tau)\hat{F}^{-1}(\tau) - (1 - \tau) \mathop{\mathbb{E}}_{t\sim T|C=c,t\leq c}[t] \\
&\quad + \Pr(t > c|C = c)(-\frac{\tau_c(1 - \tau)}{1 - \tau_c} F^{-1}(\tau) - w(1 - \tau)c + (1 - w)\tau z_\infty) \\
&= \tau_c(1 - \tau)\hat{F}^{-1}(\tau) - (1 - \tau) \mathop{\mathbb{E}}_{t\sim T|C=c,t\leq c}[t] \\
&\quad - \tau_c(1 - \tau)\hat{F}^{-1}(\tau) + (1 - \tau_c)(-w(1 - \tau)c + (1 - w)\tau z_\infty) \\
&= -(1 - \tau) \mathop{\mathbb{E}}_{t\sim T|C=c,t\leq c}[t] + (1 - \tau_c)(-w(1 - \tau)c + (1 - w)\tau z_\infty).
\end{aligned}$$

Therefore, since $\tau_c \leq \tau$ and $w = (\tau - \tau_c)/(1 - \tau_c)$, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\mathrm{Portnoy}}(\hat{F}, (z, \delta); w, \tau)] - \mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\mathrm{Portnoy}}(F, (z, \delta); w, \tau)] \\
&> ((\tau_c - \tau)\hat{F}^{-1}(\tau) + (1 - \tau_c)w\tau c) + (1 - \tau_c)w(1 - \tau)c \\
&= (\tau_c - \tau)(\hat{F}^{-1}(\tau) - c) \\
&\geq 0.
\end{aligned}$$

□

A.2. Extension of Logarithmic Score

We show a proof of Theorem 4.6.

Proof. We consider a fixed $c \sim C$, and let t be a sample obtained from T. Let i be the index such that $\zeta_i \leq c < \zeta_{i+1}$. Since Assumption 3.1 holds, we have $\Pr(\zeta_j < t \leq \zeta_{j+1}|C = c) = \Pr(\zeta_j < t \leq \zeta_{j+1}) = F(\zeta_{j+1}) - F(\zeta_j) = f_j$ for any $j < i$, $\Pr(\zeta_i < t \leq c|C = c) = F(c) - F(\zeta_i)$, and $\Pr(c < t|C = c) = \Pr(c < t) = 1 - F(c)$. Hence, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\mathrm{Cen-log}}(\hat{F}, (z, \delta); \{w_k\}_{k=0}^{B-1}, \{\zeta_k\}_{k=0}^{B})] &= -\sum_{j<i} \Pr(\zeta_j < t \leq \zeta_{j+1}|C = c) \log \hat{f}_j \\
&\quad - \Pr(\zeta_i < t \leq c|C = c) \log \hat{f}_i \\
&\quad - \Pr(c < t|C = c) \left(w_i \log \hat{f}_i + (1 - w_i) \log(1 - \hat{F}(\zeta_{i+1})) \right) \\
&= -\sum_{j<i} f_j \log \hat{f}_j \\
&\quad - (F(c) - F(\zeta_i)) \log \hat{f}_i \\
&\quad - (1 - F(c)) \left(w_i \log \hat{f}_i + (1 - w_i) \log(1 - \hat{F}(\zeta_{i+1})) \right) \\
&= -\sum_{j\leq i} f_j \log \hat{f}_j - (1 - F(\zeta_{i+1})) \log(1 - \hat{F}(\zeta_{i+1})),
\end{aligned}$$

where we used $w_i = (F(\zeta_{i+1}) - F(c))/(1 - F(c))$ for the last equality.

Hence, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\mathrm{Cen-log}}(\hat{F},(z,\delta);\{w_k\}_{k=0}^{B-1},\{\zeta_k\}_{k=0}^{B})] - \mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\mathrm{Cen-log}}(F,(z,\delta);\{w_k\}_{k=0}^{B-1},\{\zeta_k\}_{k=0}^{B})] \\
&= -\sum_{j\leq i} f_j(\log \hat{f}_j - \log f_j) - (1-F(\zeta_{i+1}))(\log(1-\hat{F}(\zeta_{i+1})) - \log(1-F(\zeta_{i+1}))) \\
&\geq 0, \qquad (11)
\end{aligned}$$

where we used the fact that the Kullback-Leibler divergence between two probability distributions is non-negative for the inequality. This means that the inequality

$$-\sum_k p_k(\log \hat{p}_k - \log p_k) \geq 0$$

holds for any two probability distributions p_k and $\hat{p}_k$ and the equality holds only if $p_k = \hat{p}_k$ for all k. Here, we use an $(i+2)$-dimensional vector $\boldsymbol{p} = (p_0, p_1, \ldots, p_{i+1})$; we set $p_k = f_k$ for all $k \leq i$ and we set $p_{i+1} = 1 - F(\zeta_{i+1})$. Note that the vectors $\boldsymbol{p}$ and $\hat{\boldsymbol{p}}$ constructed in this way are a probability distribution (i.e., $\sum_k p_k = 1$).

Since Inequality (11) holds for any $c \sim C$, we marginalize the inequality with respect to C, and we have

$$\mathop{\mathbb{E}}_{t\sim T,c\sim C}[S_{\mathrm{Cen-log}}(\hat{F},(z,\delta);\{w_i\}_{i=0}^{B-1},\{\zeta_i\}_{i=0}^{B})] \geq \mathop{\mathbb{E}}_{t\sim T,c\sim C}[S_{\mathrm{Cen-log}}(F,(z,\delta);\{w_i\}_{i=0}^{B-1},\{\zeta_i\}_{i=0}^{B})],$$

which means that $S_{\mathrm{Cen-log}}(\hat{F},(z,\delta);\{w_i\}_{i=0}^{B-1},\{\zeta_i\}_{i=0}^{B})$ is proper. □

A.3. Extension of Brier Score

We show a proof of Theorem 4.7.

Proof. We consider a fixed $c \sim C$, and let t be a sample obtained from T. Let j be the index such that $\zeta_j < c \leq \zeta_{j+1}$. Since Assumption 3.1 holds, we have $\Pr(\zeta_i < t \leq \zeta_{i+1}|C=c) = \Pr(\zeta_i < t \leq \zeta_{i+1}) = F(\zeta_{i+1}) - F(\zeta_i) = f_i$ for any $i < j$, $\Pr(\zeta_j < t \leq c|C=c) = F(c) - F(\zeta_j)$, and $\Pr(c < t|C=c) = \Pr(c < t) = 1 - F(c)$. Hence, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{t\sim T|C=c}[S_{\mathrm{Cen-Brier}}(\hat{F},(z,\delta);\{w_k\}_{k=0}^{B-1},\{\zeta_k\}_{k=0}^{B})] \\
&= \sum_{i<j}\Pr(\zeta_i < t \leq \zeta_{i+1}|C=c)\left((1-\hat{f}_i)^2 + \sum_{k\neq i}\hat{f}_k^2\right) \\
&\quad +\Pr(\zeta_j < t \leq c|C=c)\left((1-\hat{f}_j)^2 + \sum_{k\neq j}\hat{f}_k^2\right) \\
&\quad +\Pr(c < t|C=c)\left(w_j(1-\hat{f}_j)^2 + (1-w_j)\hat{f}_j^2 + \sum_{i<j}\hat{f}_i^2 + \sum_{i>j}\left(w_i(1-\hat{f}_i)^2 + (1-w_i)\hat{f}_i^2\right)\right) \\
&= \sum_{i<j} f_i\left((1-\hat{f}_i)^2 + \sum_{k\neq i}\hat{f}_k^2\right) + (F(c) - F(\zeta_j))\left((1-\hat{f}_j)^2 + \sum_{k\neq j}\hat{f}_k^2\right) \\
&\quad +(1-F(c))\left(w_j(1-\hat{f}_j)^2 + (1-w_j)\hat{f}_j^2 + \sum_{i<j}\hat{f}_i^2 + \sum_{i>j}\left(w_i(1-\hat{f}_i)^2 + (1-w_i)\hat{f}_i^2\right)\right) \\
&= \sum_i\left(f_i(1-\hat{f}_i)^2 + (1-f_i)\hat{f}_i^2\right) \\
&= \sum_i(\hat{f}_i^2 - 2f_i\hat{f}_i + f_i),
\end{aligned}$$

where we used

$$w_i = \begin{cases} 0 & \text{if } \delta = 1 \text{ and } \zeta_{i+1} < z = t, \\ 1 & \text{if } \delta = 1 \text{ and } \zeta_i < z = t \leq \zeta_{i+1}, \\ 0 & \text{if } z \leq \zeta_i \end{cases}$$

for the first equality and

$$w_i = \begin{cases} (F(\zeta_{i+1}) - F(c))/(1 - F(c)) & \text{if } \delta = 0 \text{ and } i = j, \\ f_i/(1 - F(c)) & \text{if } \delta = 0 \text{ and } i > j \end{cases}$$

for the last equality.

Hence we have

$$\begin{aligned} & \mathop{\mathbb{E}}_{t \sim T|C=c}[S_{\text{Cen-Brier}}(\hat{F}, (z, \delta); \{w_i\}_{i=0}^{B-1}, \{\zeta_i\}_{i=0}^{B})] - \mathop{\mathbb{E}}_{t \sim T|C=c}[S_{\text{Cen-Brier}}(F, (z, \delta); \{w_i\}_{i=0}^{B-1}, \{\zeta_i\}_{i=0}^{B})] \\ = & \sum_i (\hat{f}_i^2 - f_i^2 - 2f_i(\hat{f}_i - f_i)) \\ = & \sum_i (\hat{f}_i - f_i)^2 \\ \geq & \ 0. \end{aligned} \tag{12}$$

Note that the equality holds only if $\hat{f}_i = f_i$ holds for all i.

Since Inequality (12) holds for any $c \sim C$, we have

$$\mathop{\mathbb{E}}_{t \sim T, c \sim C}[S_{\text{Cen-Brier}}(\hat{F}, (z, \delta); \{w_i\}_{i=0}^{B-1}, \{\zeta_i\}_{i=0}^{B})] \geq \mathop{\mathbb{E}}_{t \sim T, c \sim C}[S_{\text{Cen-Brier}}(F, (z, \delta); \{w_i\}_{i=0}^{B-1}, \{\zeta_i\}_{i=0}^{B})],$$

which means that $S_{\text{Cen-Brier}}(\hat{F}, (z, \delta); \{w_i\}_{i=0}^{B-1}, \{\zeta_i\}_{i=0}^{B})$ is proper. □

B. Additional Experiments

We investigated the differences of the prediction performances between $S_{\text{Cen-log}}$ (defined in Eq. (4)) and $S_{\text{Cen-log-simple}}$ (defined in Eq. (5)) by using $S_{\text{Cen-log-simple}}$, D-calibration, and KM-calibration as metrics to determine the parameter B. Tables 2– 4 show the results for $B = 8, 16, 32$, respectively, where each number shows the mean and variance of the values obtained by five random runs and the bold numbers were used to emphasize the difference between two scoring rules. These results show that the prediction performances of these two scoring rules were similar for the prostateSurvival and support datasets even for $B = 8$. However, they showed different prediction performances for the flchain dataset for $B = 8$ and $B = 16$, but the performance difference was negligible for $B = 32$. Therefore, we used $B = 32$ in the other experiments in this paper.

Table 2. Comparison between two extensions of logarithmic score for $B = 8$.

Metric	Loss Function	flchain	prostateSurvival	support
$S_{\text{Cen-log-simple}}$	$S_{\text{Cen-log}}$	6.4618 ± 0.1204	1.3460 ± 0.0476	1.5422 ± 0.0704
	$S_{\text{Cen-log-simple}}$	6.4176 ± 0.1266	1.3447 ± 0.0451	1.5368 ± 0.0701
D-calibration	$S_{\text{Cen-log}}$	$\mathbf{0.0045} \pm 0.0004$	0.0002 ± 0.0000	0.0370 ± 0.0032
	$S_{\text{Cen-log-simple}}$	$\mathbf{0.0127} \pm 0.0013$	0.0002 ± 0.0001	0.0349 ± 0.0024
KM-calibration	$S_{\text{Cen-log}}$	$\mathbf{0.0048} \pm 0.0026$	0.0048 ± 0.0028	0.0057 ± 0.0027
	$S_{\text{Cen-log-simple}}$	$\mathbf{0.0614} \pm 0.0081$	0.0083 ± 0.0024	0.0061 ± 0.0033

Table 3. Comparison between two extensions of logarithmic score for $B = 16$.

Metric	Loss Function	flchain	prostateSurvival	support
$S_{\text{Cen-log-simple}}$	$S_{\text{Cen-log}}$	3.6774 ± 0.0386	1.2880 ± 0.0247	1.6017 ± 0.0733
	$S_{\text{Cen-log-simple}}$	3.6676 ± 0.0424	1.3447 ± 0.0451	1.6008 ± 0.0731
D-calibration	$S_{\text{Cen-log}}$	$\mathbf{0.0005} \pm 0.0002$	0.0001 ± 0.0000	0.0147 ± 0.0020
	$S_{\text{Cen-log-simple}}$	$\mathbf{0.0013} \pm 0.0004$	0.0002 ± 0.0000	0.0143 ± 0.0021
KM-calibration	$S_{\text{Cen-log}}$	0.0117 ± 0.0046	0.0142 ± 0.0036	0.0149 ± 0.0080
	$S_{\text{Cen-log-simple}}$	0.0162 ± 0.0049	0.0158 ± 0.0063	0.0158 ± 0.0100

Table 4. Comparison between two extensions of logarithmic score for $B = 32$.

Metric	Loss Function	flchain	prostateSurvival	support
$S_{\text{Cen-log-simple}}$	$S_{\text{Cen-log}}$	1.5054 ± 0.0508	1.3608 ± 0.0295	1.8307 ± 0.0452
	$S_{\text{Cen-log-simple}}$	1.5059 ± 0.0513	1.3609 ± 0.0301	1.8296 ± 0.0446
D-calibration	$S_{\text{Cen-log}}$	0.0003 ± 0.0001	0.0001 ± 0.0000	0.0063 ± 0.0009
	$S_{\text{Cen-log-simple}}$	0.0003 ± 0.0001	0.0001 ± 0.0000	0.0062 ± 0.0012
KM-calibration	$S_{\text{Cen-log}}$	0.0206 ± 0.0049	0.0312 ± 0.0084	0.0299 ± 0.0115
	$S_{\text{Cen-log-simple}}$	0.0213 ± 0.0049	0.0343 ± 0.0102	0.0288 ± 0.0127