Survival Models: Proper Scoring Rule and Stochastic Optimization with Competing Risks

Julie Alberge, Vincent Maladière, Olivier Grisel, Judith Abécassis, Gaël Varoquaux

Survival Models: Proper Scoring Rule and Stochastic Optimization with Competing Risks

Julie Alberge[1], **Vincent Maladière**[2], **Olivier Grisel**[2], **Judith Abécassis**[1], **Gaël Varoquaux**[1]
[1] SODA Team, Inria Saclay, Palaiseau France
[2] :probabl., Paris France
julie.alberge@inria.fr, vincent@probabl.ai

Abstract

When dealing with right-censored data, where some outcomes are missing due to a limited observation period, survival analysis —known as *time-to-event analysis*— focuses on predicting the time until an event of interest occurs. Multiple classes of outcomes lead to a classification variant: predicting the most likely event, a less explored area known as *competing risks*. Classic competing risks models couple architecture and loss, limiting scalability.

To address these issues, we design a strictly proper censoring-adjusted separable scoring rule, allowing optimization on a subset of the data as each observation is evaluated independently. The loss estimates outcome probabilities and enables stochastic optimization for competing risks, which we use for efficient gradient boosting trees. **SurvivalBoost** not only outperforms 12 state-of-the-art models across several metrics on 4 real-life datasets, both in competing risks and survival settings, but also provides great calibration, the ability to predict across any time horizon, and computation times faster than existing methods.

1 INTRODUCTION

We all die at some point. Some applications call for predicting not *if* but *when* an event of interest is likely to occur. In such a setting of *time-to-event regression*, samples often have unobserved outcomes, *e.g.* individuals that have not been followed long enough for the event of interest to occur. Limiting the analysis to fully observed samples creates a censoring bias. To address this, *survival analysis* models use dedicated corrections for censorship. These have long been central to health applications [Zhu et al., 2016, Chaddad et al., 2016, Gaynor et al., 1993]. Nowadays, survival analysis is also used in diverse fields, such as predictive maintenance [Rith et al., 2018, Susto et al., 2015], or user-engagement studies [Maystre and Russo]. Survival analysis has led to many dedicated models, such as the Kaplan and Meier [1958] estimator or the Cox [1972] proportional hazard model.

Competing risks analysis generalizes survival analysis to multiple events, determining which will happen first [Susto et al., 2015, Gaynor et al., 1993]. For instance, if a breast-cancer patient dies from a different cause, it is impossible to determine when they would have succumbed to cancer, regardless of the duration of the observation period. The caregiver may also want to adapt the treatment if it is predicted that the patient will die of a competing event, such as a heart attack, sooner than from cancer. As the risks of the various events are seldom independent–for example, cancer and cardiovascular disease share inflammation or age risk factors [Koene et al., 2016]–competing risks cannot be solved by running a survival model for each event [Wolbers et al., 2009]. The estimated risk of an event of interest will be biased if the competing risks are not included. Hence, adequate models for these risks are critical for decision-making [Ramspek et al., 2022, Koller et al., 2012, van Walraven and McAlister, 2016].

Survival models have traditionally been developed with *ad hoc* adjustments for censoring. The most common approach is to design a likelihood using the probability of censoring per unit time–*i.e.* the time-derivative of the risk–which either comes with strong parametric assumptions [Cox, 1972] or *ad hoc* corrections [Wang and Sun, 2022]. Given that the risk, which is the probability of the outcome at a specific time, is crucial for various applications, it is preferable to use proper scoring rules, that directly control probabilities, as

Preliminary work. Under review. Do not distribute.

developed by Graf et al. [1999], Rindt et al. [2022]. However, no metric (or loss) has been shown to control probabilities in the competing risks setting.

In application domains typical of survival analysis and competing risks –health, predictive maintenance, insurance, marketing– the data are mostly tabular with categorical variables, where tree-based models shine [Grinsztajn et al., 2022]. Existing survival and competing risks models do not fit well with these requirements. In particular, the proper scoring rule in Rindt et al. [2022] requires a time derivative of the risk, typically via an auto-diff operator in a neural architecture. This approach is challenging to adapt to tree-based algorithms. In addition, the ever-growing volume of data calls for computationally efficient algorithms.

Contributions Here, we provide a general theoretical framework for learning a competing risks algorithm using a strictly proper scoring rule. This scoring rule yields a loss function easy to plug into any multiclass estimator to create a competing risks algorithm, providing the individual risk of each event at any given horizon. An interesting property of this new loss is that it can be optimized on a subset of the training data. Hence, it allows stochastic optimization, enabling computationally efficient learning.
With that, we propose an algorithm called SurvivalBoost, based on Stochastic Gradient Boosting Trees. We benchmark our algorithm on a synthetic dataset and 4 real-world datasets - both in the competing risks and the survival analysis setting - with several ranking and calibration metrics and show that it outperforms 12 state-of-the-art (SOTA) baselines in both settings.

2 RELATED WORK

Survival settings Various survival models have been developed, ranging from approaches like the Kaplan and Meier [1958] estimator, which estimates the general survival curve for an entire population, to models that account for covariates. The Cox [1972] Proportional Hazards Model, a linear model of the *hazards*, which represents the instantaneous probability of an event, *i.e.*, the logarithmic derivative of outcome probabilities over time. More complex models have been adapted to the survival setting: Support Vector Machines [Van Belle et al., 2011], survival games [Han et al., 2021] and Neural networks with DeepSurv [Katzman et al., 2018] or PCHazard [Kvamme and Borgan, 2019b]. While these models do not control risks, more recent neural networks employ appropriate loss functions: DQS [Yanagisawa, 2023, though relying on a piecewise constant hazard], SumoNet [Rindt et al., 2022] which requires differentiable models.

Competing risks Competing risks, involving multiple possible outcomes, require new methods that can naturally adapt to the simpler survival analysis setting. Derived from the Kaplan and Meier [1958] estimator, the Nelson [1972]-Aalen et al. [2008] estimator is an unbiased marginal model for competing risks.
The linear Fine and Gray [1999] estimator, inspired by the Cox [1972] estimator in survival analysis, is the most popular model in clinical research. Recently, machine learning models have been adapted to competing risks settings, including tree-based approaches such as the Random Survival Forests [Ishwaran et al., 2008, Kretowska, 2018, Bellot and Schaar, 2018], boosting approaches [Bellot and van der Schaar, 2018], and neural networks approaches such as DeepHit and Gaussian mixtures approaches [Lee et al., 2018, Aala and van der Schaar, 2017, Danks and Yau, 2022, Nagpal et al., 2021]. Tranformer-based approaches with SurvTRACE [Wang and Sun, 2022] using a loss corrected to predict rare competing events, independently forecasts all events but do not ensure that probabilities sum to one.
For a comprehensive review of competing risks models, refer to Monterrubio-Gómez et al. [2022].

Evaluation for such models Prediction evaluation in survival or competing risks settings requires adapted metrics to account for right-censored data points [Harrell et al., 1982], such as the C-index, which is an adaptation of the Area Under the ROC Curve (AUC) used in classification tasks. However, the C-index only evaluates the ranking of samples, *i.e.* which samples are likely to experience the event of interest first. It is also dependent on the censoring distribution, which can introduce bias in the evaluation [Blanche et al., 2019]. In fact, the score may be inflated for distributions that differ from the oracle-censoring distributionRindt et al. [2022]. Alternative methods have been proposed, such as the *time-dependent* C-index, C_ζ [Antolini et al., 2005], which is the same metric but computed at a specific time horizon ζ. The C-index ranking metric has also been extended to competing risks [Uno et al., 2011], but, as in the survival setting, it only evaluates relative risks between pairs of individuals and does not assess the absolute risk for a given individual. Other time-dependent adaptations of the ROC curve have been developed, though these also measure discriminative power rather than the actual risks or probabilities [Blanche et al., 2013]. Yet, controlling risk is crucial for decision making [Van Calster et al., 2019]. Proper scoring rules offer an alternative to overcome the limitations of existing metrics, as they capture more aspects of the problem. Additionally, they can be used for both the training and evaluating probabilistic predictive models.

Proper Scoring Rules (PSR) Scoring rules are cost functions of observations and a candidate probability distribution. When *proper*, they target the oracle probability distribution (Definition 3.2). Crucially, they give machine-learning losses that recover probabilities of outcomes. For classification, where discrete events are observed rather than probabilities, the Brier score and the log loss give proper scoring rules, with relative merits [Benedetti, 2010, Merkle and Steyvers, 2013].

Graf et al. [1999] adapt the Brier score to survival analysis, with a strong assumption of independence of the covariates in the censoring distribution. Yet, this assumption is often violated [Kvamme and Borgan, 2019a], leading to bias [Rindt et al., 2022]. Rindt et al. [2022] show that the likelihood of the survival function yields a proper scoring rule, but requires both the density function and the survival function, which is a time-wise derivative of outcome probabilities (Definition 3.2). For quantile regression, Yanagisawa [2023] adapt the Pinball loss to a proper scoring rule for survival analysis, but requiring an oracle parameter. Han et al. [2021] introduce a double optimization problem, where the stationary point corresponds to the oracle distributions.

For competing risks, Schoop et al. [2011] extend the Brier score to a proper scoring rule. However, the Brier score does not capture the uncertainty as effectively as the log loss [Benedetti, 2010].

3 PROBLEM FORMULATION

Notations We write oracle quantities as a^* and estimates as $\hat{a}$, vectors in bold, $\mathbf{a}$, random variables in upper case, A, observations in lower cases a, and distributions in calligraphic style $\mathcal{A}$.

3.1 Problem Setting

We consider $K \in \mathbb{N}^*$ competing events. For $k \in [\![1, K]\!]$, we denote $T_k^* \in \mathbb{R}_+$ the event time of the event k, depending on the covariates $\mathbf{X} \sim \mathcal{X}$. We also denote $T^* \in \mathbb{R}_+$, the first event of interest that occurs, $T^* = \min_{k \in [\![1,K]\!]} (T_k^*)$. We observe $(\mathbf{X}, T, \Delta) \sim \mathcal{D}$, with $T = \min(T^*, C)$ where $C \in \mathbb{R}_+$ is the censoring time, which can depend on $\mathbf{X}$, and $\Delta \in [\![0, K]\!], \Delta = \arg\min_{k \in [\![0,K]\!]}(T_k^*)$, where 0 denotes a censored observation.
However, we are primarily interested in the distribution of the uncensored data, $(\mathbf{X}, T^*, \Delta) \sim \mathcal{D}^*$, particularly the joint distribution of $T^*, \Delta | \mathbf{X} = \mathbf{x}$.
Given a data set of n individuals, we denote each individual i by its associated covariates $\mathbf{x}_i$. The outcome is represented by (t_i, δ_i), where t_i is the observed time, and $\delta_i \in [\![0, K]\!]$ is the event indicator. $\delta_i = k$ indicates that the event of interest k was observed at time t_i, while $\delta_i = 0$ indicates that the observation was censored at time t_i. This paper aims to predict an unbiased estimate of all cause-specific Cumulative Incidence functions (CIFs) at any time horizon ζ (Definition 3.1).

Definition 3.1 (*Quantities of interest*)**.**

Survival Function to any event:

$$S^*(\zeta|\mathbf{x}) = \mathbb{P}(T^* > \zeta|\mathbf{X} = \mathbf{x})$$

CIF (Cumulative Incidence Function):

$$F^*(\zeta|\mathbf{x}) = \mathbb{P}(T^* \leq \zeta|\mathbf{X} = \mathbf{x}) = 1 - S^*(\zeta|\mathbf{X} = \mathbf{x})$$

CIF of the k^{th} event:

$$F_k^*(\zeta|\mathbf{x}) = \mathbb{P}(T^* \leq \zeta \cap \Delta = k|\mathbf{X} = \mathbf{x})$$

Censoring Function:

$$G^*(\zeta|\mathbf{x}) = \mathbb{P}(C > \zeta|\mathbf{X} = \mathbf{x})$$

Assumption 3.1 (*Non-informative censoring*)**.** We make the classic assumption in survival analysis that censoring is non-informative with respect to covariates:

$$T^* \perp\!\!\!\perp C \,|\, \mathbf{X}$$

Assumption 3.1 is essential for most theoretical results in survival analysis [Rindt et al., 2022, Yanagisawa, 2023, Han et al., 2021]. It shows that single-event survival analysis becomes invalid in the presence of competing risks: if some observations are censored due to other events that share unobserved risk factors with the event of interest, this assumption is violated.

3.2 CIF Scoring Rule

Proper Scoring Rule A scoring rule ℓ evaluates a distribution $\mathcal{P}$ on an observation Y, producing a corresponding score $\ell(\mathcal{P}, Y)$. The higher the score, the better the model fits the observation. For a proper scoring rule, the score reflects the model's ability to predict the oracle distribution [for more on scoring rules, see Gneiting and Raftery, 2007, Ovcharov, 2018, Merkle and Steyvers, 2013].

Definition 3.2 (*Proper Scoring Rule*)**.** A scoring rule ℓ is considered proper if

$$\forall \mathcal{P}, \mathcal{Q}, \text{distributions} \quad \mathbb{E}_{Y \sim \mathcal{Q}}[\ell(\mathcal{P}, Y)] \leq \mathbb{E}_{Y \sim \mathcal{Q}}[\ell(\mathcal{Q}, Y)].$$

If the equality holds if and only if $\mathcal{P} = \mathcal{Q}$, in which case the scoring rule is *strictly proper*.

Proper scoring rule for the Global CIF We denote L_ζ a scoring rule for the global CIF at time ζ.

Definition 3.3 (*PSR for competing risks settings*)**.** In competing risks settings, where censoring is present, a scoring rule L_ζ for the CIF at time ζ for an observation $(\mathbf{X}, T, \Delta)$ is proper if and only if:

$$\forall \zeta, (\mathbf{X}, T, \Delta) \sim \mathcal{D}, \forall (\hat{F}_1, ..., \hat{F}_K, \hat{S}),$$

$$\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}[L_\zeta(\ (\hat{F}_1(\zeta|\mathbf{x}), ..., \hat{F}_K(\zeta|\mathbf{x}), \hat{S}(\zeta|\mathbf{x}))\ , (T, \Delta))]$$

Estimated distributions

$$\leq$$

$$\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}[L_\zeta(\ (F_1^*(\zeta|\mathbf{x}), ..., F_K^*(\zeta|\mathbf{x}), S^*(\zeta|\mathbf{x}))\ , (T, \Delta))] \quad (1)$$

Oracle distributions

When equality is achieved *only* for the oracle distributions, the scoring rule is *strictly proper*.

4 A STRICTLY PROPER SCORING RULE FOR COMPETING RISKS

We prove that the negative log-likelihood, re-weighted by the censoring distribution (IPCW: Inverse Probabilities of Censoring Weights), is strictly proper.

Definition 4.1 (Competitive Weights Negative LogLoss)**.** We introduce the multiclass negative log-likelihood, re-weighted with the censoring distribution. The different classes represent the loss for all the cumulative incidence functions and the survival function.

$$\forall \zeta, (\mathbf{x}, t, \delta) \sim \mathcal{D},$$
$$\mathrm{L}_\zeta((\hat{F}_1(\zeta|\mathbf{x}), ..., \hat{F}_K(\zeta|\mathbf{x}), \hat{S}(\zeta|\mathbf{x})), (t, \delta)) \stackrel{\text{def}}{=}$$
$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\frac{\mathbb{1}_{t_i\leq\zeta,\delta_i=k}\ \log\left(\hat{F}_k(\zeta|\mathbf{x}_i)\right)}{G^*(t_i|\mathbf{x}_i)}\right.$$

Probability of remaining censor-free at t_i

$$\left. + \frac{\mathbb{1}_{t_i>\zeta}\ \log\left(\hat{S}(\zeta|\mathbf{x}_i)\right)}{G^*(\zeta|\mathbf{x}_i)}\right) \quad (2)$$

Probability of remaining censor-free at ζ (1 - probability of censoring)

Eq.2 is a standard log-loss (also known as cross-entropy), reweighted by appropriate sample weights —the inverse probabilities, or IPCW. Therefore, it can easily be added to most multiclass estimators.

Lemma 4.1. *Accounting for the time horizon ζ, the expectation of the above scoring rule can be written as:* $\forall \zeta, (\mathbf{X}, T, \Delta) \sim \mathcal{D}$,

$$\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathrm{L}_\zeta\left((\hat{F}_1(\zeta|\mathbf{x}), ..., \hat{F}_K(\zeta|\mathbf{x}), \hat{S}(\zeta|\mathbf{x})), (T, \Delta)\right)\right]$$
$$= \sum_{k=1}^{K}\log\left(\hat{F}_k(\zeta|\mathbf{x})\right)F_k^*(\zeta|\mathbf{x}) + \log\left(\hat{S}(\zeta|\mathbf{x})\right)S^*(\zeta|\mathbf{x}) \quad (3)$$

Proof sketch. The weights allow us to transition from the observed distribution T to the uncensored distribution T^*, which is crucial for demonstrating properness. The full proof can be found in Appendix B. □

Theorem 1 (Properness of the scoring rule)**.** *Under the assumption that the weights are appropriately chosen, $L_\zeta : \mathbb{R}^{K+1} \times \mathcal{D} \to \mathbb{R}$ is a strictly proper scoring rule for the global CIF on a fixed time horizon $\zeta \in \mathbb{R}_+$.*

Proof sketch. Using the previous result, the properties of the negative log-likelihood, and Definition 3.3, we conclude that the loss is strictly proper. Full proof in Appendix B. □

5 SURVIVALBOOST: GRADIENT BOOSTING COMPETING RISKS

While eq.2 can be used as a loss in any multiclass machine learning algorithm, we choose Gradient Boosting Trees due to their strong performance on tabular data [Grinsztajn et al., 2022] and their compatibility with stochastic optimization. Gradient boosting methods approximate complex functions by combining weak learners (or base learners). At each iteration m, the algorithm focuses on the residuals of the loss function and builds a base learner h_m that minimizes these residuals. For gradient boosting trees, the estimator typically takes the form $H_m(x) = H_{m-1}(x) + \nu h_m(x)$ where ν represents a chosen learning rate. For more on gradient boosting, refer to Friedman [1999].

Most survival or competing risk loss functions cannot be used with tree-based models, as they require time derivates and thus smoothness. To address this, we introduce an algorithm called SURVIVALBOOST, which predicts all CIFs for each competing event as well as the global survival function. By predicting these jointly, we ensure that the stability of the probabilities is maintained, as the outputs of the classification models naturally sum to one. This ensures that $\mathbb{P}(T^* \leq \zeta|\mathbf{X} = \mathbf{x}) + \mathbb{P}(T^* > \zeta|\mathbf{X} = \mathbf{x}) = 1$, meaning the model's outputs are consistent and sum to one:

$$\sum_{k=1}^{K}\underbrace{\mathbb{P}(T^* \leq \zeta \cap \Delta^* = k|\mathbf{X} = \mathbf{x})}_{k^{th}\text{CIF}} + \underbrace{\mathbb{P}(T^* > \zeta|\mathbf{X} = \mathbf{x})}_{\text{Survival Probability}} = 1$$

Using the loss in eq.3, we can directly predict the CIF instead of predicting the hazard function (the derivative of the CIF), as is often done —for example, in DeepHit [Lee et al., 2018] or SurvTRACE [Wang and Sun, 2022]. This approach allows us to drop the constant-hazard assumption present in [Yanagisawa, 2023, Kvamme and Borgan, 2019b, Wang and Sun, 2022, Rindt et al., 2022].

Our algorithm utilizes two classifiers (here, gradient-boosted trees), one for censoring, trained on binary censored/non-censored labels (i.e., for time ζ, $\mathbb{P}(C > \zeta|\mathbf{X} = \mathbf{x})$), and one for multiple events. Both the censoring and event models are adjusted using IPCW weights. To compute these IPCW weights, we iterate

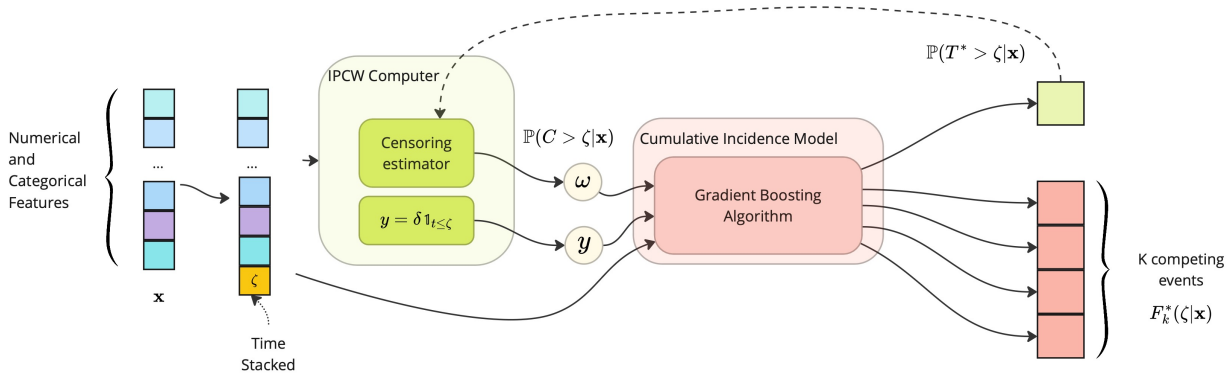

Figure 1: **SurvivalBoost Algorithm with its Feedback Loop.** After providing input to the algorithm, a random time is assigned, and the corresponding weights and target are computed. After each iteration, the feedback loop updates the censoring probability, $G^\star$ as defined in eq.2.

Algorithm 1 SurvivalBoost Algorithm - m^{th} Iteration

Input: $\mathbf{x}, \delta, t$
for $i = 1$ **to** $n_{samples}$ **do**
  $\zeta_i \sim \mathcal{U}(0, t_{max})$
end for
$\zeta \leftarrow (\zeta_i)_{1 \leq i \leq n_{samples}}$ ▷Sample a time horizon
$\tilde{\mathbf{x}} \leftarrow (\mathbf{x}, \zeta)$ ▷Stacking the time to the features
$y, w \leftarrow \text{ipcwComputer}(\mathbf{x}, \delta, t, \hat{G})$ ▷See Alg 2
$L \leftarrow \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \left(\mathbb{1}_{y_i=k}\ y_i\ w_i\ \log\left(\hat{F}_k(\zeta_i|\mathbf{x}_i)\right)\right)$
$\quad + \mathbb{1}_{y_i=0}\ y_i\ w_i \log\left(\hat{S}(\zeta_i|\mathbf{x}_i)\right)$
$h_m(\tilde{\mathbf{x}}) \leftarrow$ Train one iteration of Gradient Boost with L as the loss ▷h_m is the m^{th} weak learner
$H_m(\tilde{\mathbf{x}}) \leftarrow \nu h_m(\tilde{\mathbf{x}}) + H_{m-1}(\tilde{\mathbf{x}})$ ▷H_m is the estimator at the m^{th} iteration, ν the learning rate
$(\hat{S}(\zeta|\mathbf{X} = \mathbf{x}), (\hat{F}_k(\zeta|\mathbf{X} = \mathbf{x})_{1 \leq k \leq K}) \leftarrow H_m(\tilde{\mathbf{x}})$
$\hat{G} \leftarrow$ Train one iteration of the Censoring-Feedback-Loop with $\hat{S}(\zeta|\mathbf{X} = \mathbf{x})$ ▷See Alg 3

the training using a feedback loop similar to boosting. First, we compute a survival censoring model. Then, using these probabilities, we initialize our SurvivalBoost algorithm. After several iterations, we apply a feedback loop to retrain the censoring model.
To capture complex temporal dependencies, we uniformly sample a time point for each observation and include it as an additional feature. Multiple time points can be sampled per iteration for each observation, generating a richer dataset where the targets vary based on the specific times sampled, thus providing a broader range of temporal information. This is enabled by our separable loss function. An additional benefit is that we can predict the CIF at any time, unlike models optimized for a limited number of time points that require interpolation for other times.

Figure 1 illustrates an iteration: we compute the weights w_i and targets y_i based on the sampled times for each individual (eq. 2). Specifically, for censored samples, the corresponding weight is set to 0. A target $y_i \in [\![1, K]\!]$ indicates that the event of interest occurred before ζ and when $y_i = 0$, the individual has survived without experiencing any event. Algorithm 1 gives pseudocode.

6 COMPETING RISKS EXPERIMENTS

6.1 Evaluation Metrics For Competing Risks

The evaluation is mainly performed on two metrics[1].

Evaluating the predicted probability We extend the method proposed by Graf et al. [1999] and Schoop et al. [2011]. The formula and a formal proof of the properness of the loss can be found in Appendix C. To avoid potential circularity with the loss function that we optimized, we apply this evaluation metric to the Brier Score rather than the log-loss. To evaluate the model across all time points, we sum the Brier Score over time, resulting in the *Integrated Brier Score* (IBS).

Prediction accuracy in time In many applications, such as predictive maintenance or medicine, it is crucial to determine the first event a subject is likely to encounter. We use a validation metric to check, for each sample, whether the observed event is predicted as the most likely at given times, selected as before using quantiles. For example, for an individual who encounters event 2 at time t, the probability of surviving until t should be the highest compared to the proba-

[1]We do not focus on the C-index over time, as this metric is biased [Blanche et al., 2019, Rindt et al., 2022]

bilities of encountering any other event. Additionally, the probability of encountering event 2 after t should be the highest. To measure this, we adapt Multi-Class accuracy to different time points:

Definition 6.1 (Prediction accuracy at time ζ)**.** For a fixed time horizon ζ, and denoting survival to any event as index 0, define $\hat{y} = \arg\max_{k \in [0,K]} \hat{F}_k(\zeta | \mathbf{X} = \mathbf{x})$, the most probable event at ζ, and $y_\zeta = \mathbb{1}_{t \leq \zeta} \delta$. We remove censored individuals, and n_{nc} represents the number of uncensored individuals at ζ.

$$Acc(\zeta) = \frac{1}{n_{nc}} \sum_{i=1}^{n} \mathbb{1}_{\hat{y}_i = y_{i,\zeta}} \, \mathbb{1}_{\overline{\delta_i = 0, t_i \leq \zeta}} \tag{4}$$

6.2 Experimental Settings

Synthetic Dataset We design a synthetic dataset with linear relations between features and targets, as well as dependencies between the censoring distribution and the features (Appendix S4). To create the synthetic dataset, for each sample, we draw $2n_{events}$ parameters from a normal distribution. We then generate the event durations from a Weibull distribution based on those parameters. The observation is determined by the minimum duration and its associated event. The censoring event is computed using the same method.

SEER Dataset This dataset tracks 470,000 breast cancer patients for up to ten years, with mortality due to various diseases as the outcomes. The censoring rate is approximately 63%, and Figure S3 shows the distribution of events. Unlike Lee et al. [2018] (DeepHit) and Wang and Sun [2022] (SurvTRACE), which focus on the two most prevalent events and censor the others (undermining the competing risk framework), we consider three competing events, aggregating the remaining events into a third class. We also remove some features following Wang and Sun [2022].

Baselines We compare our approach with 7 other competing risks models from simpler models with Aalen et al. [2008]'s global estimator and the Fine and Gray [1999] linear model to more complex methods.
We benchmark against tree-based approach - Random Survival Forests (RSF) [Ishwaran et al., 2008] -, often criticized for its memory limitations. In our comparison, we also include several neural network-based models. This includes DeepHit [Lee et al., 2018] which is trained with a ranking loss that combines the C-index with a negative log-likelihood, Deep Survival Machines (DSM) [Nagpal et al., 2021] which employ a graphical method for feature encoding and DeSurv [Danks and Yau, 2022] solves Ordinal Differential Equations for continuous time predictions. Finally, we include a transformer-based model, SurvTRACE [Wang and Sun, 2022] which is trained at three-time horizons (based on quantiles of observed event times) and at time 0.
To compute the Integrated Brier Score over time, other methods require linear interpolation of their trained times. For times beyond their trained intervals, we assume the incidence remains constant. In contrast, our method is trained on uniformly sampled time horizons, allowing for predictions at any time.
For fair model comparison, we use the same hyperparameter-tuning time budget (grid in Appendix S11).

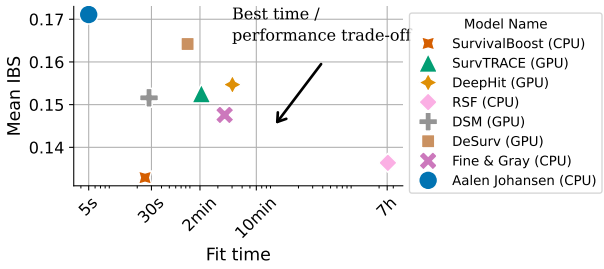

Figure 2: **Prediction performance / training time trade-off for competing risk on the synthetic dataset.** Average IBS compared the fitting time for each model on 20k training data points, with a censoring rate of approximately 50% and a dependant censoring across 6 features.

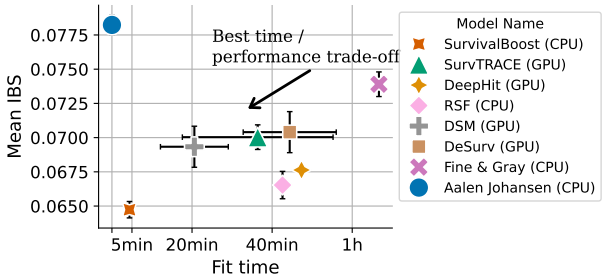

Figure 3: **Prediction performance / training time trade-off for competing risks on SEER dataset**. Average IBS versus fitting time for each model, with a maximum of 330k training points, except for Fine & Gray (50k) and RSF (100k). Table S2 provides the IBS values for each event.

6.3 Results: Competing Risks

Synthetic dataset Figure 2 illustrates the trade-off between statistical performance and training time for each model. Using the synthetic dataset, we are able to compute an oracle IBS. SurvivalBoost performs best in terms of IBS and is the fastest to train.

Results on SEER Dataset On the real-life dataset, we keep 30% of the data for testing the models. Figure 3

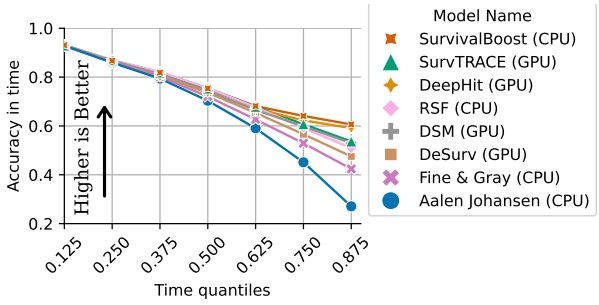

Figure 4: **Prediction accuracy at time** ζ Accuracy of the Argmax of the Cumulative Incidence Functions across different time quantiles on the SEER dataset.

compares the models using the Integrated Brier Score (with Kaplan-Meier weights from Graf et al. [1999] due to the absence of an oracle). SurvivalBoost achieves both the best score and the shortest training time. Random Survival Forest struggles with larger sample sizes (100k) and requires more than 50 GB of RAM. SurvivalBoost also maintains a significant lead with less training samples (Appendix G.3).

Event and time-specific C-indexes are presented in Table S3, but they do not capture the models' ability to predict which event is more likely to occur at a given time horizon. This capability is measured by the accuracy in time, shown in Figure 4, where SurvivalBoost demonstrates the best performance. The advantage increases as time progresses, indicating that SurvivalBoost interpolates more effectively over time.

7 USAGE IN SURVIVAL ANALYSIS

7.1 Survival Experiments

Real-life Datasets As our model can also handle survival analysis, we conducted experiments on three real-life survival datasets.

METABRIC [Curtis et al., 2012] The Molecular Taxonomy of Breast Cancer International Consortium dataset contains gene expression data with approximately 2,000 data points.

SUPPORT [Knaus et al., 1995] Study to Understand Prognoses Preferences Outcomes and Risks of Treatment dataset includes survival times for hospital patients, with more than 8,000 data points.

KKBOX The Churn Prediction Challenge 2017 hosted on Kaggle, which features administrative censoring and 2.5M data points. We trained the models over 100k, 1M, and 2M data points to assess scalability (see Appendix, Fig. S1).

Evaluation We use various metrics to evaluate models: the Integrated Brier Score (detailed in Appendix C) and another metric from Yanagisawa [2023], called $S_{Cen-log-simple} \overset{\text{def}}{=} S_{C-l-s}$ (detailed in Appendix E). Although this metric approximates the proper scoring metric from Rindt et al. [2022], it is not exactly proper (see Appendix E). It can be applied to any model as it does not require the density of the CIFs.

Baselines We benchmark our method against the most performant competing risks and SOTA survival models. This includes neural networks such as DeepHit [Lee et al., 2018] and PCHazard [Kvamme and Borgan, 2019b], as well as those trained with proper survival analysis scoring rules, such as SumoNet [Rindt et al., 2022], and DQS [Yanagisawa, 2023]. We also evaluate transformer methods with SurvTRACE [Wang and Sun, 2022], survival games [Han et al., 2021], and tree-based methods with Random Survival Forests (RSF) [Ishwaran et al., 2008] and Gradient Boosting Survival Analysis (GBS) - from Scikit-survival [Pölsterl, 2020].

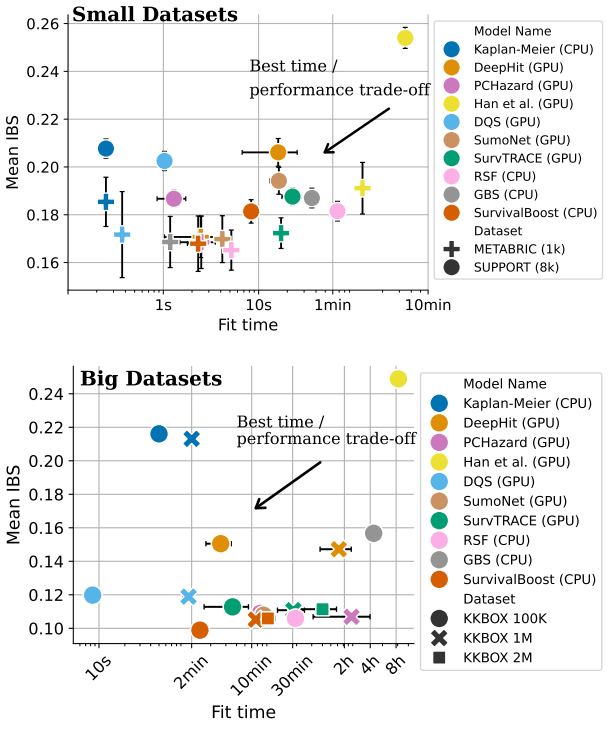

Figure 5: **Prediction performance / training time trade-off in survival analysis** IBS (Integrated Brier score) function of fit time for each model on real-life datasets. For the big datasets, some algorithms exceeded computing resources.

Table 1: **Survival datasets**: Integrated Brier Score and S_{C-l-s} (Lower is Better) depending on the size of each dataset. The ✗ indicates models that could not handle the data volume due to memory limitations.

Dataset	METABRIC (1k)		SUPPORT (8k)		KKBOX (1M)	
Model	IBS	S_{C-l-s}	IBS	S_{C-l-s}	IBS	S_{C-l-s}
Kaplan-Meier	.185±.010	2.039±.218	.208±.004	1.617±.268	.213±.001	1.723±.002
DeepHit	.171±.009	2.039±.001	.207±.004	1.771±.000	.147±.001	1.609±.002
PCHazard	.169±.011	1.980±.086	.187±.004	1.673±.004	.107±.002	1.286±.002
Han et al.	.196±.004	2.665±.036	.253±.002	3.223±.005	✗	✗
DQS8	.172±.018	2.200±.000	.202±.004	2.764±.12	.119±.001	3.791±.027
SuMo net	.170±.010	2.197±.000	.194±.006	1.818±.000	✗	✗
SurvTRACE	.172±.006	1.987±.088	.188±.004	1.606±.003	.111±.002	1.270±.008
RSF	**.165±.025**	**1.937±.227**	.182±.004	1.942±.023	✗	✗
GBS	.169±.011	1.974±.404	.187±.004	1.575±.001	.157±.001	1.511±.001
SurvivalBoost	.168±.019	2.027±.159	**.181±.005**	**1.569±0.341**	**.105±.001**	**1.183±.029**

7.2 Results: Survival Analysis

Figure 5 shows the trade-off between training time and performance in terms of IBS, where SurvivalBoost excels, being the top model in statistical performance and one of the fastest on the datasets with enough data (SUPPORT and KKBOX) while being one of the best models for smaller datasets (METABRIC). Appendix G.2 provides a similar figure for the $S_{Cen-log-simple}$ metric, where SurvivalBoost achieves an excellent trade-off rivaled only by SumoNet, which has comparable performance on the $S_{Cen-log-simple}$ loss. Varying the sample size from 100k to 2M on the KKBOX dataset confirms that SurvivalBoost and DQS are faster (taking less than 1 minute on 100k data points), while Han et al., SumoNet, and RSF are slower for larger sample size. They exhibit super-linear time complexity, making them impractical for large datasets; for more than 100k data points they exceed memory limitations (See Appendix I.1).
Table 1 report evaluation metrics, including $S_{Cen-log-simple}$ which is not what SurvivalBoost directly optimizes. Across datasets, SurvivalBoost achieves the best results in terms of IBS and is tied with SumoNet for $S_{Cen-log-simple}$ (also for C-index, Appendix H.1). It is worth noting that SumoNet uses $S_{Cen-log-simple}$ as its training loss. However, this metric is not guaranteed to be a proper scoring rule, meaning it does not necessarily ensure accurate recovery of the true risks. For KKBOX, we only show the results for 1M data points.

Beyond proper scores, we investigate calibration, MAE, MSE, and the AUC adapted for survival analysis (Appendix S5, S7, S8). We assess the calibration using four tests, including distribution calibration (D_c) [Haider et al., 2018] and One-time calibration (ONE_c) [Hosmer et al., 1997]. Kaplan-Meier, SurvivalBoost, and RSF are the most calibrated models (Appendix S9).

DISCUSSION AND CONCLUSION

Code reproducibility and data The code will be made available on GitHub as a library.

Combination of tree-based architecture and loss function makes the difference Our work shares similarities with the equations in Han et al. [2021], which also uses IPCW [introduced by Robins et al., 1994], though for survival and not competing risks. Their learning strategy targets an equilibrium, showing that it recovers the oracle distribution in survival analysis settings. Meanwhile, our optimization uses a loss on all classes to compute the censoring distribution, while the other part optimizes only for the survival distribution. This last part departs from the schema in Han et al. [2021]. Despite similarities, the two approaches behave markedly different our empirical study.
Building upon trees-based model is probably important to this difference and to the success of SurvivalBoost. Yet, comparing to GBS and RSF show that trees in themselves do not suffice. Our loss is crucial for scalability (as it is separable) and to facilitate fitting trees, as it avoids the need for time derivatives. It avoids issues that plague many competing risks methods. The excellent empirical results, superior performance with less computational resources, come from combining the loss function with the tree-based approach results in a very stable algorithm. This double gain is especially valuable as health datasets continue to grow in size.

Acknowledgments JA, JA, and GV acknowledge funding from the ERC grand INTERCEPT-T2D.

Limitations and further work Further work should consider removing the assumption of non-informative censoring 3.1. This assumption is very common in the literature, though some recent work has relaxed it in survival settings Foomani et al. [2023], Zhang et al. [2023].

Conclusion For competing risks, which generalizes survival analysis to classify the type of outcome, we first propose and prove a strictly proper scoring rule. This reweighted log loss can easily be used in machine learning models: it is separable by observation, making it suitable for stochastic solvers, it does not require time derivatives (unlike most survival models) and it can be applied to non-differentiable models. We integrate it into gradient-boosting trees, resulting in an algorithm called SurvivalBoost. By using time as a feature and incorporating a feedback loop to better estimate censoring probabilities, SurvivalBoost outperforms state-of-the-art methods on both synthetic and real-life datasets, for both competing risks (classification on time-censored data) and standard survival analysis (time-to-event regression with right censoring). It also trains faster on large datasets. As a loss function, it allows survival analysis or competing risks modeling to be easily extended to a wide range of models— from scalable linear models to deep learning architectures, including fine-tuning foundation models— replacing clinical standards like Fine and Gray that do not scale.

References

Ahmed M. Aala and Mihaela van der Schaar. Deep Multi-task Gaussian Processes for Survival Analysis with Competing Risks. In *Advances in Neural Information Processing Systems*, 2017.

Odd O. Aalen, Ornulf Borgan, and Haakon K. Gjessing. *Survival and Event History Analysis*. Statistics for Biology and Health. Springer New York, 2008.

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005. ISSN 0277-6715, 1097-0258.

Alexis Bellot and Mihaela Schaar. Tree-based Bayesian Mixture Model for Competing Risks. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, 2018.

Alexis Bellot and Mihaela van der Schaar. Multitask Boosting for Survival Analysis with Competing Risks. In *Advances in Neural Information Processing Systems*, 2018.

Riccardo Benedetti. Scoring Rules for Forecast Verification. *Monthly Weather Review*, 138(1):203–211, 2010.

Paul Blanche, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in medicine*, 32(30):5381–5397, 2013.

Paul Blanche, Michael W Kattan, and Thomas A Gerds. The c-index is not proper for the evaluation of t-year predicted risks. *Biostatistics*, 20(2):347–357, 2019.

Ahmad Chaddad, Christian Desrosiers, and Matthew Toews. Radiomic analysis of multi-contrast brain MRI for the prediction of survival in patients with glioblastoma multiforme. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4035–4038, Orlando, FL, USA, 2016. IEEE.

D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Christina Curtis, Sohrab Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar Rueda, Mark Dunning, Doug Speed, Andy Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, Carlos Caldas, Samuel Aparicio, James Brenton, and Anne-Lise Børresen-Dale. The genomic and transcriptomic architecture of 2,000 breast tumors reveals novel subgroups. *Nature*, 486:–, 04 2012. doi: 10.1038/nature10983,.

Dominic Danks and Christopher Yau. Derivative-Based Neural Modelling of Cumulative Distribution Functions for Survival Analysis. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, 2022.

Jason P. Fine and Robert J. Gray. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999.

Ali Hossein Gharari Foomani, Michael Cooper, Russell Greiner, and Rahul G. Krishnan. Copula-based deep survival models for dependent censoring, 2023.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. 1999.

Jeffrey J. Gaynor, Eric J. Feuer, Claire C. Tan, Danny H. Wu, Claudia R. Little, David J. Straus, Bayard D. Clarkson, and Murray F. Brennan. On the Use of Cause-Specific Failure and Conditional Failure Probabilities: Examples from Clinical Oncology Data. *Journal of the American Statistical Association*, 88(422):400–409, 1993.

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102 (477):359–378, 2007.

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.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on tabular data?, 2022. arXiv:2207.08815.

Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective Ways to Build and Evaluate Individual Survival Distributions, 2018. arXiv:1811.11347.

Xintian Han, Mark Goldstein, Aahlad Puli, Thomas Wies, Adler Perotte, and Rajesh Ranganath. Inverse-weighted survival games. *Advances in neural information processing systems*, 34:2160–2172, 2021.

Jr Harrell, Frank E., Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546, 1982.

D. W. Hosmer, T. Hosmer, S. Le Cessie, and S. Lemeshow. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16(9):965–980, 1997.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3), 2008.

Hemant Ishwaran, Thomas A Gerds, Udaya B Kogalur, Richard D Moore, Stephen J Gange, and Bryan M Lau. Random survival forests for competing risks. *Biostatistics*, 15(4):757–773, 2014.

E. L. Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

Jared L. Katzman, Uri Shaham, Alexander Cloninger, 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(1): 24, 2018.

William Knaus, Frank Harrell, Joanne Lynn, L Goldman, Russell Phillips, Alfred Connors, Jr, Neal Dawson, W Fulkerson, R Califf, N Desbiens, Peter Layde, Robert Oye, P Bellamy, Rabia Hakim, and D Wagner. 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*, 122:191–203, 03 1995.

Ryan J Koene, Anna E Prizment, Anne Blaes, and Suma H Konety. Shared risk factors in cardiovascular disease and cancer. *Circulation*, 133(11):1104–1114, 2016.

Michael T Koller, Heike Raatz, Ewout W Steyerberg, and Marcel Wolbers. Competing risks and the clinical community: irrelevance or ignorance? *Statistics in medicine*, 31(11-12):1089–1097, 2012.

Malgorzata Kretowska. Tree-based models for survival data with competing risks. *Computer Methods and Programs in Biomedicine*, 159:185–198, 2018.

Haavard Kvamme and Ornulf Borgan. The Brier Score under Administrative Censoring: Problems and Solutions, 2019a. arXiv:1912.08581.

Haavard Kvamme and ornulf Borgan. Continuous and Discrete-Time Survival Prediction with Neural Networks, 2019b. arXiv:1910.06724.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

Lucas Maystre and Daniel Russo. Temporally-Consistent Survival Analysis.

Edgar C. Merkle and Mark Steyvers. Choosing a Strictly Proper Scoring Rule. *Decision Analysis*, 10(4):292–304, 2013.

Karla Monterrubio-Gómez, Nathan Constantine-Cooke, and Catalina A. Vallejos. A review on competing risks methods for survival analysis, 2022. arXiv:2212.05157.

Chirag Nagpal, Xinyu Rachel Li, and Artur Dubrawski. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks, 2021.

Wayne Nelson. Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics*, 14(4):945–966, 1972.

Evgeni Y. Ovcharov. Proper scoring rules and Bregman divergence. *Bernoulli*, 24(1), 2018. ISSN 1350-7265.

Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020.

Chava L Ramspek, Lucy Teece, Kym IE Snell, Marie Evans, Richard D Riley, Maarten van Smeden, Nan van Geloven, and Merel van Diepen. Lessons learnt when accounting for competing events in the external validation of time-to-event prognostic models. *International journal of epidemiology*, 51(2):615–625, 2022.

David Rindt, Robert Hu, David Steinsaltz, and Dino Sejdinovic. Survival Regression with Proper Scoring Rules and Monotonic Neural Networks, 2022. arXiv:2103.14755.

Monorom Rith, Jimwell Soliman, Alexis Fillone, Jose Bienvenido M. Biona, and Neil Stephen Lopez. Analysis of Vehicle Survival Rates for Metro-Manila. In *IEEE 10th International Conference on*

Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM), pages 1–4, 2018.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

Rotraut Schoop, Jan Beyersmann, Martin Schumacher, and Harald Binder. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal*, 53(1):88–112, 2011.

Gian Antonio Susto, Andrea Schirru, Simone Pampuri, Sean McLoone, and Alessandro Beghi. Machine Learning for Predictive Maintenance: A Multiple Classifier Approach. *IEEE Transactions on Industrial Informatics*, 11(3):812–820, 2015.

Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117, May 2011.

Vanya Van Belle, Kristiaan Pelckmans, Sabine Van Huffel, and Johan A.K. Suykens. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, 53(2):107–118, 2011.

Ben Van Calster, David J McLernon, Maarten Van Smeden, Laure Wynants, Ewout W Steyerberg, and Topic Group 'Evaluating diagnostic tests prediction models' of the STRATOS initiative. Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):230, 2019.

Carl van Walraven and Finlay A McAlister. Competing risk bias was common in kaplan–meier risk estimates published in prominent medical journals. *Journal of clinical epidemiology*, 69:170–173, 2016.

Zifeng Wang and Jimeng Sun. SurvTRACE: Transformers for Survival Analysis with Competing Events. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–9, 2022.

Marcel Wolbers, Michael T Koller, Jacqueline CM Witteman, and Ewout W Steyerberg. Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology*, 20(4):555–561, 2009.

Hiroki Yanagisawa. Proper scoring rules for survival analysis, 2023.

Weijia Zhang, Chun Kai Ling, and Xuanhui Zhang. Deep copula-based survival analysis for dependent censoring with identifiability guarantees, 2023.

Xinliang Zhu, Jiawen Yao, and Junzhou Huang. Deep convolutional neural network for survival analysis with pathological images. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 544–547. IEEE, 2016.

A Definitions

A.1 Notations

Below, we detail the notations used throughout the main manuscript, as well as in the proofs and derivations. The following conventions apply to all symbols:

- $.^*$: Oracle
- $\hat{.}$: Estimation

The different variables that we use are:

Maths Symbol	Domain	Description
ζ	$\mathbb{R}_+$	Time horizon
K	$\mathbb{N}^*$	number of competing events (events of interest)
$\mathbf{X}$	$\mathcal{X}$	random variable representing an individual
T_k^*	$\mathbb{R}_+$	random variable of the time-to-event for event k
C	$\mathbb{R}_+$	random variable of the time-to-censoring
T^*	$\mathbb{R}_+$	$\min(T_1^*, T_2^*, ..., T_K^*)$
T	$\mathbb{R}_+$	$\min(T, C)$
Δ^*	$[1, K]$	$\underset{k\in[1,K]}{\arg\min}(T_k^*)$
Δ	$[0, K]$	$\arg\min(C, T_1^*, T_2^*, ..., T_K^*)$
S	$\mathcal{S}$	Survival function
F	$\mathcal{F}$	Cumulative Incidence Function
G	$\mathcal{G}$	Censor function
n	$\mathbb{N}^*$	number of individuals in our observation
i	$[1, n]$	one observation
$\mathbf{x}_i$	$\mathcal{X}^n$	individuals observed
t_i	$\mathbb{R}_+^n$	time-to-event/censoring observed
δ_i	$[0, K]$	event observed, 0 indicates censoring

Table S1: Notations used

A.2 Reporting conventions

In the tables, the best results are highlighted in bold, and the second-best results are underlined.

B Theory on our proper scoring rule: proofs and derivations

In this appendix, we give the proofs and derivations concerning the proper scoring rule that we have introduced.

Lemma 4.1. *Accounting for the time horizon ζ, the expectation of the above scoring rule can be written as:* $\forall \zeta, (\mathbf{X}, T, \Delta) \sim \mathcal{D}$,

$$\begin{aligned} &\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathrm{L}_\zeta\left((\hat{F}_1(\zeta|\mathbf{x}), ..., \hat{F}_K(\zeta|\mathbf{x}), \hat{S}(\zeta|\mathbf{x})), (T, \Delta)\right)\right] \\ &= \sum_{k=1}^{K} \log\left(\hat{F}_k(\zeta|\mathbf{x})\right) F_k^*(\zeta|\mathbf{x}) + \log\left(\hat{S}(\zeta|\mathbf{x})\right) S^*(\zeta|\mathbf{x}) \end{aligned} \tag{3}$$

Proof the of Lemma 4.1 on the expectation of the Reweighted NLL.

$\forall \zeta, \forall k \in [\![1, K]\!], (\mathbf{x}, t, \delta) \sim \mathcal{D}$,

$$\mathrm{L}_\zeta\left((\hat{F}_1(\zeta|\mathbf{x}), ..., \hat{F}_K(\zeta|\mathbf{x}), \hat{S}(\zeta|\mathbf{x})), (t, \delta)\right) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{\sum_{k=1}^{K}\frac{\mathbb{1}_{t_i\leq\zeta,\delta_i=k}\ \log\left(\hat{F}_k(\zeta|\mathbf{x}_i)\right)}{G^*(t_i|\mathbf{x}_i)}}_{\stackrel{\text{def}}{=}\Psi_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}),(t,\delta))}\right) + \underbrace{\frac{\mathbb{1}_{t_i>\zeta}\ \log\left(\hat{S}(\zeta|\mathbf{x}_i)\right)}{G^*(\zeta|\mathbf{x}_i)}}_{\stackrel{\text{def}}{=}\Lambda_{k,\zeta}(\hat{S}(\zeta|\mathbf{x}),(t,\delta))} \tag{5}$$

For the next computations, we recall the definition of the different variables.

Computation of the expectation: First:

$$\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\Psi_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}),(T,\Delta))\right] = \mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathbb{1}_{T\leq\zeta}\mathbb{1}_{\Delta=k}\frac{\log\left(\hat{F}_k(\zeta|\mathbf{x})\right)}{G^*(T|\mathbf{x})}\right] \tag{6}$$

$$= \log\left(\hat{F}_k(\zeta|\mathbf{x})\right)\ \mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\frac{\mathbb{1}_{\min(T^*,C)\leq\zeta}\mathbb{1}_{\Delta=k}}{G^*(T|\mathbf{x})}\right] \tag{7}$$

$$= \log\left(\hat{F}_k(\zeta|\mathbf{x})\right)\ \mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\frac{(\mathbb{1}_{T^*\leq\zeta}\mathbb{1}_{T^*\leq C}+\mathbb{1}_{C\leq\zeta}\mathbb{1}_{C\leq T^*})\mathbb{1}_{\Delta=k}}{G^*(T|\mathbf{x})}\right] \tag{8}$$

$$= \log\left(\hat{F}_k(\zeta|\mathbf{x})\right)\ \mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\frac{\mathbb{1}_{T^*\leq\zeta}\mathbb{1}_{T^*\leq C}\mathbb{1}_{\Delta=k}}{G^*(T|\mathbf{x})} + \underbrace{\frac{\mathbb{1}_{C\leq\zeta}\mathbb{1}_{C\leq T^*}\mathbb{1}_{\Delta=k}}{G^*(T|\mathbf{x})}}_{=0 \text{ because } k\neq 0}\right] \tag{9}$$

$$= \log\left(\hat{F}_k(\zeta|\mathbf{x})\right)\ \mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\frac{\mathbb{1}_{T^*\leq\zeta}\mathbb{1}_{T^*\leq C}\mathbb{1}_{\Delta=k}}{G^*(T|\mathbf{x})}\right] \tag{10}$$

$$= \log\left(\hat{F}_k(\zeta|\mathbf{x})\right)\mathbb{P}(T^*\leq\zeta,\Delta=k|\mathbf{X}=\mathbf{x}) \tag{11}$$

The last equality can be expanded as follows:

$$\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\frac{\mathbb{1}_{T^*\leq\zeta}\mathbb{1}_{T^*\leq C}\mathbb{1}_{\Delta=k}}{G^*(T|\mathbf{x})}\right] = \int_0^\infty\int_0^\infty(\mathbb{1}_{\min(t,c)=t} + \underbrace{\mathbb{1}_{\min(t,c)=c}}_{=0 \text{ because } k\neq 0})\frac{\mathbb{1}_{t\leq\zeta}\mathbb{1}_{t\leq c}}{G^*(t|\mathbf{x})}f_{T^*,C,\Delta}(t,c,k|\mathbf{x})dt\,dc \tag{12}$$

T is a composition of T^* and C

$$= \int_0^\infty\int_0^\infty\frac{\mathbb{1}_{t\leq\zeta}\mathbb{1}_{t\leq c}}{G^*(t|\mathbf{x})}f_{T^*,C,\Delta}(t,c,k|\mathbf{x})dt\,dc \tag{13}$$

$$= \int_0^\infty\int_0^\infty\frac{\mathbb{1}_{t\leq\zeta}\mathbb{1}_{t\leq c}}{G^*(t|\mathbf{x})}f_{T^*,\Delta}(t,k|\mathbf{x})f_C(c|\mathbf{x})dt\,dc \tag{14}$$

Because $T^* \perp\!\!\!\perp C|\mathbf{X}$

$$= \int_0^\infty\frac{\mathbb{1}_{t\leq\zeta}}{G^*(t|\mathbf{x})}f_{T^*,\Delta}(t,k|\mathbf{x})\left(\int_0^\infty\mathbb{1}_{t\leq c}f_C(c|\mathbf{x})dc\right)dt \tag{15}$$

$$= \int_0^\infty\frac{\mathbb{1}_{t\leq\zeta}}{G^*(t|\mathbf{x})}f_{T^*,\Delta}(t,k|\mathbf{x})\left(G^*(t|\mathbf{x})\right)dt \tag{16}$$

with the definition of G^*

$$= \int_0^\infty\mathbb{1}_{t\leq\zeta}f_{T^*,\Delta}(t,k|\mathbf{x})dt \tag{17}$$

$$= \mathbb{P}(T^*\leq\zeta,\Delta=k|\mathbf{X}=\mathbf{x}) \tag{18}$$

And:

$$\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left[\Lambda_{k,\zeta}(\hat{S}(\zeta|\mathbf{X}=\mathbf{x}),(T,\Delta))|\mathbf{X}=\mathbf{x}\right] = \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathbb{1}_{T>\zeta}\frac{\log\left(\hat{S}(\zeta|\mathbf{X}=\mathbf{x})\right)}{G^*(\zeta|\mathbf{x})}\right] \tag{19}$$

$$= \log\left(\hat{S}(\zeta|\mathbf{X}=\mathbf{x})\right)\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left[\frac{\mathbb{1}_{\min(T^*,C)>\zeta}}{G^*(\zeta|\mathbf{x})}\right] \tag{20}$$

$$= \log\left(\hat{S}(\zeta|\mathbf{X}=\mathbf{x})\right)\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left[\frac{\mathbb{1}_{T^*>\zeta}\mathbb{1}_{C>\zeta}}{G^*(\zeta|\mathbf{x})}\right] \tag{21}$$

$$= \log\left(\hat{S}(\zeta|\mathbf{X}=\mathbf{x})\right)\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left[\frac{\mathbb{1}_{C>\zeta}}{G^*(\zeta|\mathbf{x})}\right]\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathbb{1}_{T^*>\zeta}\right] \tag{22}$$

Because $T^* \perp\!\!\!\perp C|\mathbf{X}$

$$= \log\left(\hat{S}(\zeta|\mathbf{X}=\mathbf{x})\right)\frac{\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathbb{1}_{C>\zeta}\right]}{G^*(\zeta|\mathbf{x})}\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathbb{1}_{T^*>\zeta}\right] \tag{23}$$

Because $G^*(\zeta|\mathbf{x})$ does not depend of T and Δ

$$= \log\left(\hat{S}(\zeta|\mathbf{X}=\mathbf{x})\right)\mathbb{P}(T^*>\zeta|\mathbf{X}=\mathbf{x}) \tag{24}$$

$$\tag{25}$$

By summing all of the terms, we obtain:

$$\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathrm{L}_\zeta\left((\hat{F}_1(\zeta|\mathbf{x}),...,\hat{F}_K(\zeta|\mathbf{x}),\hat{S}(\zeta|\mathbf{x})),(T,\Delta)\right)\right]$$

$$= \sum_{k=1}^{K}\log\left(\hat{F}_k(\zeta|\mathbf{x})\right)\mathbb{P}(T^*\leq\zeta,\Delta=k|\mathbf{X}=\mathbf{x}) + \log\left(\hat{S}(\zeta|\mathbf{X}=\mathbf{x})\right)\mathbb{P}(T^*>\zeta|\mathbf{X}=\mathbf{x}) \tag{26}$$

$$= \sum_{k=1}^{K}\log\left(\hat{F}_k(\zeta|\mathbf{x})\right)F_k^*(\zeta|\mathrm{x}) + \log\left(\hat{S}(\zeta|\mathbf{x})\right)S^*(\zeta|\mathbf{x}) \tag{27}$$

Finally:

$$\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathrm{L}_\zeta\left((\hat{F}_1(\zeta|\mathbf{x}),...,\hat{F}_K(\zeta|\mathbf{x}),\hat{S}(\zeta|\mathbf{x})),(T,\Delta)\right)\right]$$

$$= \sum_{k=1}^{K}\log\left(\hat{F}_k(\zeta|\mathbf{x})\right)F_k^*(\zeta|\mathrm{x}) + \log\left(\hat{S}(\zeta|\mathbf{x})\right)S^*(\zeta|\mathbf{x}) \tag{28}$$

□

Proof of the Theorem 1.

Theorem 1 (Properness of the scoring rule)**.** *Under the assumption that the weights are appropriately chosen, $L_\zeta : \mathbb{R}^{K+1}\times\mathcal{D}\to\mathbb{R}$ is a strictly proper scoring rule for the global CIF on a fixed time horizon $\zeta\in\mathbb{R}_+$.*

To be more explicit, we can define a new random variable Y:

Definition B.1.

$$\forall\zeta,\ Y_{k,\zeta}\overset{\text{def}}{=} T^*\leq\zeta\cap\Delta=k$$

And:

$$\forall\zeta,\ Y_{0,\zeta}\overset{\text{def}}{=} T^*>\zeta$$

Thus, the previously mentioned quantities of interest can be rewritten as functions of these variables:

$$F_k^*(\zeta|\mathbf{x}) = \mathbb{P}(T^* \leq \zeta, \Delta = k|\mathbf{X} = \mathbf{x}) = \mathbb{P}(Y_{k,\zeta} = 1|\mathbf{X} = \mathbf{x}) \tag{29}$$

$$S^*(\zeta|\mathbf{x}) = \mathbb{P}(T^* > \zeta|\mathbf{X} = \mathbf{x}) = \mathbb{P}(Y_{0,\zeta} = 1|\mathbf{X} = \mathbf{x}) \tag{30}$$

$\hat{F}_k(\zeta|\mathbf{x})$ represents the estimated probability that $Y_{k,\zeta} = 1$, so we rewrite it as $\hat{p}_{k,\zeta} \stackrel{\text{def}}{=} \hat{F}_k(\zeta|\mathbf{x})$.
Therefore:

$$\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathrm{L}_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}), (T, \Delta))\right] = \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}[\mathrm{L}_\zeta(\hat{p}_\zeta, (T, \Delta))] \tag{31}$$

$$= \sum_{k=0}^{K} \log\left(\hat{p}_{k,\zeta}\right) \mathbb{P}(Y_{k,\zeta} = 1|\mathbf{X} = \mathbf{x}) \tag{32}$$

Using Lemma 4.1

Thus, we obtain the following optimization problem:

$$\begin{aligned} \max_{\hat{p}} \quad & \sum_{k=0}^{K} \log\left(\hat{p}_{k,\zeta}\right) \mathbb{P}(Y_{k,\zeta} = 1|\mathbf{X} = \mathbf{x}) \\ \text{s.t.} \quad & \sum_{k=0}^{K} \hat{p}_k = 1 \\ & \hat{p}_k \geq 0 \end{aligned} \tag{33}$$

The problem can be reformulated as a convex optimization problem due to the concavity of the logarithm:

$$\begin{aligned} \min_{\hat{p}} \quad & -\sum_{k=0}^{K} \log\left(\hat{p}_{k,\zeta}\right) \mathbb{P}(Y_{k,\zeta} = 1|\mathbf{X} = \mathbf{x}) \\ \text{s.t.} \quad & \sum_{k=0}^{K} \hat{p}_k = 1 \\ & \hat{p}_k \geq 0 \end{aligned} \tag{34}$$

We apply the Karush-Kuhn-Tucker conditions since the constraints are qualified (as they are linear). These conditions imply that if p is a local minimum of the problem, there exits $\lambda \in \mathbb{R}$ and $\mu \in \mathbb{R}_+^{K+1}$ such that:

$$\nabla\left(-\sum_{k=0}^{K} \log\left(\hat{p}_{k,\zeta}\right) \mathbb{P}(Y_{k,\zeta} = 1|\mathbf{X} = \mathbf{x})\right) - \mu^\top \mathbf{1}_K + \lambda = 0 \tag{35}$$

$$\forall k, \mu_k \hat{p}_{k,\zeta} = 0 \tag{36}$$

If $\exists k, \hat{p}_{k,\zeta} = 0 \implies -\sum_{k=0}^{K} \log\left(\hat{p}_{k,\zeta}\right) \mathbb{P}(Y_{k,\zeta} = 1|\mathbf{X} = \mathbf{x}) = +\infty$.
Hence, equation (36) implies that $\forall k, \mu_k = 0$.

Now,

$$\forall k, \frac{\partial \left(-\sum_{k=0}^{K} \log\left(\hat{p}_{k,\zeta}\right) \mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x}) \right)}{\partial \hat{p}_{k,\zeta}} = -\frac{\mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x})}{\hat{p}_{k,\zeta}} \tag{37}$$

(37) can be rewritten as:

$$\forall k, \quad -\frac{\mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x})}{\hat{p}_{k,\zeta}} + \lambda = 0 \tag{38}$$

$$\implies \forall k, -\mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x}) + \lambda \hat{p}_{k,\zeta} = 0 \tag{39}$$

By summing over k,

$$\implies -\underbrace{\sum_{k=0}^{K} \mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x})}_{=1} + \lambda \underbrace{\sum_{k=0}^{K} \hat{p}_{k,\zeta}}_{=1} = 0 \tag{40}$$

$$\implies \lambda = 1 \tag{41}$$

$$\implies \forall k, \hat{p}_{\zeta,k} = \mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x}) \tag{42}$$

Any local minimum must satisfy the KKT conditions. Therefore, if p is a local minimum, it is a solution to equations (34) and (42). Consequently, as shown above, the only possible solution must be equal to the oracle distribution. Indeed, the loss is strictly proper.

□

C Study of the proper scoring rule used for evaluation

As mentioned earlier, the most commonly used metric in the competing risks setting, the C-index over time, is known to be biased [Blanche et al., 2019, Rindt et al., 2022]. To address this significant issue in evaluation strategies, we propose two alternative evaluation metrics: one based on a reweighted proper scoring rule, which can be applied to any proper binary scoring rule, and another based on accuracy over time, which measures the observed event against the most likely predicted event.

C.1 PSR for evaluation

The PSR introduced in the main paper as the loss function of our algorithm serves as a global loss across all predictions. The following loss is adapted to focus on a specific event k, allowing us to evaluate our estimates for that event. In the paper, we focus on the IBS, though one could alternatively use a logarithmic loss because of its properness.

Proper scoring rule for the k^{th} competing event In our setting, we denote $L_{k,\zeta}$ as a scoring rule for the k^{th} CIF at a time horizon ζ.

Definition C.1 (*PSR for the k^{th} cause-specific event*)**.** The scoring rule $L_{k,\zeta}$ for the k^{th} CIF at time ζ for an observation $(\mathbf{X}, T, \Delta)$ is proper if and only if:

$$\forall \zeta, (\mathbf{X}, T, \Delta) \sim \mathcal{D}, \quad \mathbb{E}_{T^*, C, \Delta | \mathbf{X} = \mathbf{x}}[L_{k,\zeta}(\hat{F}_k(\zeta | \mathbf{x}), (T, \Delta))] \leq \mathbb{E}_{T^*, C, \Delta | \mathbf{X} = \mathbf{x}}[L_{k,\zeta}(F_k^*(\zeta | \mathbf{x}), (T, \Delta))] \tag{43}$$

C.1.1 A proper scoring rule for competing risks

To evaluate our model, we used the following proper scoring rule, which is appropriate for each event. This proper scoring rule allows us to assess the error for each specific event and the global error across all CIFs.

In the following, we prove that any given (strictly) proper scoring rule that can be used in the multiclass setting (*e.g.* the Brier score or negative log-likelihood) leads to a (strictly) proper scoring in competing risks settings by re-weighting the observations.
Indeed, for any (strictly) proper scoring rule $\ell : \mathbb{R} \times \{0, 1\} \to \mathbb{R}$, we can construct a cause-specific scoring rule function $L_{k,\zeta} : \mathbb{R} \times \mathcal{D} \to \mathbb{R}$, which is also a (strictly) proper scoring rule for the k^{th} cause-specific event at the fixed time horizon $\zeta \in \mathbb{R}_+$. It follows that L_ζ is (strictly) proper.

Definition C.2 (*PSR with re-weighting*)**.** We define $L_{k,\zeta}$, considering the observations $(\mathbf{x}, t, \delta)$ for an event k, as the following scoring rule for the k^{th} CIF:

$$\forall \zeta, \forall k \in [\![1, K]\!], \ell : \mathbb{R} \times \{0,1\} \to \mathbb{R}, (\mathbf{x}, t, \delta) \sim \mathcal{D}$$

$$\mathrm{L}_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}), (t,\delta)) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}_{t_i \leq \zeta, \delta_i = k} \ \ell\left(\hat{F}_k(\zeta|\mathbf{x}_i), 1\right)}{G^*(t_i|\mathbf{x}_i)}$$

Probability of remaining at ζ (1 - probability of censoring)

$$+ \frac{\mathbb{1}_{t_i > \zeta} \ \ell\left(\hat{F}_k(\zeta|\mathbf{x}_i), 0\right)}{G^*(\zeta|\mathbf{x}_i)}$$

Probability of remaining at t_i

$$+ \frac{\mathbb{1}_{t_i \leq \zeta, \delta_i \neq 0, \delta_i \neq k} \ \ell\left(\hat{F}_k(\zeta|\mathbf{x}_i), 0\right)}{G^*(t_i|\mathbf{x}_i)} \tag{44}$$

The weights correspond to the Inverse Probability of Censoring Weighting (IPCW), which is used to re-calibrate the observed population to align with the uncensored oracle population Robins et al. [1994]. This PSR is an extension of Graf et al. [1999] and Schoop et al. [2011] when ℓ is the Brier Score.

Lemma C.1. *Considering a proper scoring rule $\ell : \mathbb{R} \times \{0,1\}$, at time horizon ζ and for any cause-specific risk k, the expectation of the scoring rule can be expressed as:*

$$\forall \zeta, \forall k \in [\![1, K]\!], \ell : \mathbb{R} \times \{0,1\} \to \mathbb{R}, (\mathbf{X}, T, \Delta) \sim \mathcal{D},$$
$$\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathrm{L}_{k,\zeta}\left(\hat{F}_k(\zeta|\mathbf{x}), (T,\Delta)\right)\right] = \ell\left(\hat{F}_k(\zeta|\mathbf{x}), 1\right) F_k^*(\zeta|\mathbf{x}) + \ell\left(\hat{F}_k(\zeta|\mathbf{x}), 0\right)(1 - F_k^*(\zeta|\mathbf{x})) \tag{45}$$

Proof. The computations are essentially the same as in the previous section.

$$\forall \zeta, \forall k \in [\![1, K]\!], \ell : \mathbb{R} \times \{0,1\} \to \mathbb{R}, (\mathbf{x}, t, \delta) \sim \mathcal{D}$$

$$\mathrm{L}_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}), (t,\delta)) \stackrel{\text{def}}{=} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}_{t_i \leq \zeta, \delta_i = k} \ \ell\left(\hat{F}_k(\zeta|\mathbf{x}_i), 1\right)}{G^*(t_i|\mathbf{x}_i)}}_{\stackrel{\text{def}}{=}\Psi_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}),(t,\delta))}$$

$$+ \underbrace{\frac{\mathbb{1}_{t_i > \zeta} \ \ell\left(\hat{F}_k(\zeta|\mathbf{x}_i), 0\right)}{G^*(\zeta|\mathbf{x}_i)}}_{\stackrel{\text{def}}{=}\Lambda_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}),(t,\delta))}$$

$$+ \underbrace{\frac{\mathbb{1}_{t_i \leq \zeta, \delta_i \neq 0, \delta_i \neq k} \ \ell\left(\hat{F}_k(\zeta|\mathbf{x}_i), 0\right)}{G^*(t_i|\mathbf{x}_i)}}_{\stackrel{\text{def}}{=}\Phi_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}),(t,\delta))} \tag{46}$$

$$\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\Psi_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}), (T,\Delta))|\mathbf{X}=\mathbf{x}\right] = \mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathbb{1}_{T\leq\zeta}\mathbb{1}_{\Delta=k}\frac{\ell\left(\hat{F}_k(\zeta|\mathbf{x}), 1\right)}{G^*(T|\mathbf{x})}\right] \tag{47}$$

$$= \ell\left(\hat{F}_k(\zeta|\mathbf{x}), 1\right) \ \mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\frac{\mathbb{1}_{T^*\leq\zeta}\mathbb{1}_{T^*\leq C}\mathbb{1}_{\Delta=k}}{G^*(T|\mathbf{x})}\right] \tag{48}$$

$$= \ell\left(\hat{F}_k(\zeta|\mathbf{x}), 1\right) \mathbb{P}(T^* \leq \zeta, \Delta = k|\mathrm{X} = \mathrm{x}) \tag{49}$$

$$\tag{50}$$

$$\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\Phi_{k,\zeta}\left(\hat{F}_k(\zeta|\mathbf{x}),(T,\Delta)\right)\right] = \mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathbb{1}_{T\leq\zeta,\Delta\neq 0,\Delta\neq k}\frac{\ell\left(\hat{F}_k(\zeta|\mathbf{x}),0\right)}{G^*(T|\mathbf{x})}\right] \tag{51}$$

$$= \ell\left(\hat{F}_k(\zeta|\mathbf{x}),0\right)\ \mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\frac{\mathbb{1}_{T^*\leq\zeta}\mathbb{1}_{T^*\leq C}\mathbb{1}_{\Delta\neq k}}{G^*(T|\mathbf{x})}\right] \tag{52}$$

$$= \ell\left(\hat{F}_k(\zeta|\mathbf{x}),0\right)\mathbb{P}(T^*\leq\zeta,\Delta\neq k|\mathrm{X}=\mathrm{x}) \tag{53}$$

$$\tag{54}$$

$$\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\Lambda_{k,\zeta}(\hat{F}_k(\zeta,\mathbf{x}),(T,\Delta))|\mathbf{X}=\mathbf{x}\right] = \mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathbb{1}_{T>\zeta}\frac{\ell\left(1-\hat{F}_k(\zeta|\mathbf{x}),0\right)}{G^*(\zeta|\mathbf{x})}\right] \tag{55}$$

$$= \ell\left(\hat{F}_k(\zeta|\mathbf{x}),0\right)\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[]\frac{\mathbb{1}_{T^*>\zeta}\mathbb{1}_{C>\zeta}}{\mathbb{P}(C>\zeta|\mathbf{x})}\right] \tag{56}$$

$$= \ell\left(\hat{F}_k(\zeta|\mathbf{x}),0\right)\mathbb{P}(T^*>\zeta|\mathbf{X}=\mathbf{x}) \tag{57}$$

$$\tag{58}$$

By summing all of the terms, we obtain:

$$\begin{aligned}\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathrm{L}_{k,\zeta}\left(\hat{F}_k(\zeta|\mathbf{x}),(T,\Delta)\right)\right] &= \ell\left(\hat{F}_k(\zeta|\mathbf{x}),1\right)\mathbb{P}(T^*\leq\zeta,\Delta=k)\\ &+\ell\left(\hat{F}_k(\zeta|\mathbf{x}),0\right)\left(\mathbb{P}(T^*\leq\zeta,\Delta\neq k|\mathrm{X}=\mathrm{x})+\mathbb{P}(T^*>\zeta|\mathbf{X}=\mathbf{x})\right)\end{aligned} \tag{59}$$

Meanwhile,

$$\mathbb{P}(\overline{T^*\leq\zeta\cap\Delta=k}) = \mathbb{P}(T^*>\zeta\cup\Delta\neq k) \tag{60}$$

$$= \mathbb{P}(T^*>\zeta)+\mathbb{P}(\Delta\neq k)-\mathbb{P}(T^*>\zeta\cap\Delta\neq k) \tag{61}$$

$$= \mathbb{P}(T^*>\zeta)+\mathbb{P}(\Delta\neq k\cap T^*>\zeta)+\mathbb{P}(\Delta\neq k\cap T^*\leq\zeta)-\mathbb{P}(T^*>\zeta\cap\Delta\neq k) \tag{62}$$

$$= \mathbb{P}(T^*>\zeta)+\mathbb{P}(\Delta\neq k\cap T^*\leq\zeta) \tag{63}$$

Therefore, we obtain:

$$\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathrm{L}_{k,\zeta}\left(\hat{F}_k(\zeta|\mathbf{x}),(T,\Delta)\right)\right] = \ell\left(\hat{F}_k(\zeta|\mathbf{x}),1\right)F_k^*(\zeta|\mathbf{x})+\ell\left(\hat{F}_k(\zeta|\mathbf{x}),0\right)(1-F_k^*(\zeta|\mathbf{x})) \tag{64}$$

□

Proposition C.1. *If $\ell:\mathbb{R}\times\{0,1\}\to\mathbb{R}$ is a chosen (strictly) proper scoring rule, then $L_{k,\zeta}:\mathbb{R}\times\mathcal{D}\to\mathbb{R}$ is also a (strictly) proper scoring rule for the k^{th} cause-specific event at the fixed time horizon $\zeta\in\mathbb{R}_+$.*

Proof.

$$\begin{aligned}\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathrm{L}_{k,\zeta}\left(\hat{F}_k(\zeta|\mathbf{x}),(T,\Delta)\right)\right] &= \ell\left(\hat{F}_k(\zeta|\mathbf{x}),1\right)\mathbb{P}(T^*\leq\zeta,\Delta=k|\mathbf{X}=\mathbf{x})\\ &+\ell\left(\hat{F}_k(\zeta|\mathbf{x}),0\right)\left(\mathbb{P}(T^*\leq\zeta,\Delta\neq k|\mathbf{X}=\mathbf{x})+\mathbb{P}(T^*>\zeta|\mathbf{X}=\mathbf{x})\right)\end{aligned} \tag{65}$$

To be more explicit, we define a new random variable Y:

Definition C.3.

$$\forall\zeta,\ Y_{k,\zeta}\overset{\text{def}}{=}T^*\leq\zeta\cap\Delta=k$$

$$F_k^*(\zeta|\mathbf{x}) = \mathbb{P}(T^* \leq \zeta, \Delta = k|\mathbf{X} = \mathbf{x}) = \mathbb{P}(Y_{k,\zeta} = 1|\mathbf{X} = \mathbf{x}) \tag{66}$$

$\hat{F}_k(\zeta|\mathbf{x})$ represents the estimated probability that $Y_{k,\zeta} = 1$, allowing us to rewrite it as: $\hat{p}_{k,\zeta} \stackrel{\text{def}}{=} \hat{F}_k(\zeta|\mathbf{x}) \approx \mathbb{P}(Y_{k,\zeta} = 1|\mathbf{X} = \mathbf{x})$ Therefore:

$$\begin{aligned}
\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathrm{L}_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}),(T^*,C,\Delta))\right] &= \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}[\mathrm{L}_{k,\zeta}(\hat{p}_{k,\zeta},(T,\Delta))] && (67)\\
&= \ell\left(\hat{p}_{k,\zeta},0\right)\mathbb{P}(Y_{k,\zeta}=0|\mathbf{X}=\mathbf{x}) + \ell\left(\hat{p}_{k,\zeta},1\right)\mathbb{P}(Y_{k,\zeta}=1|\mathbf{X}=\mathbf{x}) && (68)\\
&= \mathbb{E}_{Y_{k,\zeta}}[\ell(\hat{p}_{k,\zeta},Y_{k,\zeta})|\mathbf{X}=\mathbf{x}] && (69)\\
&\leq \mathbb{E}_{Y_{k,\zeta}}[\ell(p_{k,\zeta},Y_{k,\zeta})|\mathbf{X}=\mathbf{x}] && (70)\\
&\leq \mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}[\mathrm{L}_{k,\zeta}(\mathbb{P}(Y_{k,\zeta}=1),(T,\Delta))] && (71)\\
&\leq \mathbb{E}[\mathrm{L}_{k,\zeta}(F_k^*(\zeta|\mathrm{x}),(T,\Delta))] && (72)
\end{aligned}$$

The last inequality holds because l is a proper scoring rule. Similarly, the same computation leads to a strictly proper scoring rule if l is strictly proper.

Thus, we conclude that $\forall\zeta, \forall k \in [\![1,K]\!]$, $\mathrm{L}_{k,\zeta}(\hat{F}_k(\zeta|\mathrm{x}),(T,\Delta))$ is a proper scoring rule of $F_k^*(\zeta|\mathbf{x})$.

□

Theorem 2. *If $\ell : \mathbb{R}\times\{0,1\}\to\mathbb{R}$, a chosen (strictly) proper scoring rule, then $L_\zeta : \mathbb{R}\times\mathcal{D}\to\mathbb{R}$ is a (strictly) proper scoring rule for the global CIF at a fixed time horizon $\zeta \in \mathbb{R}_+$.*

Proof. This follows straightforwardly from the proposition and the lemma above. □

Corollary: Proper global scoring rule to compare competing risk models The defined scoring rule $\sum_{k=1}^K \mathrm{L}_{k,\zeta}$ is proper on any arbitrarily chosen time horizon ζ. To compare different models, a global measure is necessary, such as summing over time, as introduced by Graf et al. [1999]. Here, we extend the Integrated Brier Score to other (strictly) proper scoring rules l and prove that the Integrated Loss (IL) is also a (strictly) proper scoring rule.
By considering:

$$Z \sim \mathcal{U}(0, t_{max})$$

with t_{max} being the maximum time horizon for prediction.

Definition C.4 (*Integrated global PSR*)**.** With $\ell : \mathbb{R}\times\{0,1\}\to\mathbb{R}$, a chosen scoring rule, the cause-specific scoring rule function $L_{k,\zeta} : \mathbb{R}\times\mathcal{D}\to\mathbb{R}$ defined as above, we define the IL as

$$\begin{aligned}
\mathrm{IL}(\hat{F}_1(.|\mathbf{x}),...,\hat{F}_K(.|\mathbf{x}),(T,\Delta)) &\stackrel{\text{def}}{=} \mathbb{E}_Z\left[\sum_{k=1}^K \mathrm{L}_{k,Z}(\hat{F}_k(Z|\mathbf{x}),(T,\Delta))|\mathbf{X}=\mathbf{x}\right] && (73)\\
&= \sum_{k=1}^K \underbrace{\mathbb{E}_Z\left[\mathrm{L}_{k,Z}(\hat{F}_k(Z|\mathbf{x}),(T,\Delta))|\mathbf{X}=\mathbf{x}\right]}_{\stackrel{\text{def}}{=}\mathrm{IL}_k(\hat{F}_k(.|\mathbf{x}),(T,\Delta))} && (74)
\end{aligned}$$

Corollary C.1. *With $\ell : \mathbb{R}\times\{0,1\}\to\mathbb{R}$, a chosen (strictly) proper scoring rule, the cause-specific loss function $L_{k,\zeta} : \mathbb{R}\times\mathcal{D}\to\mathbb{R}$ defined above* IL *is a (strictly) proper scoring rule.*

Proof. We have already proven that $L_{k,\zeta} : \mathbb{R}\times\mathcal{D}\to\mathbb{R}$ is a (strictly) proper scoring rule. Given the monotonicity and positivity of the expectation, the result follows immediately.

$$\begin{aligned}
\mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x},Z=\zeta}\left[\mathrm{IL}_k(\hat{F}_k(\zeta|\mathbf{x})),(T,\Delta)\right] &= \mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x},Z=\zeta}\left[\mathrm{L}_k(\hat{F}_k(\zeta|\mathbf{x}),(T,\Delta))\right] && (75)\\
&\leq \mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x},Z=\zeta}\left[\mathrm{L}_k(F_k^*(\zeta|\mathbf{x}),(T,\Delta))\right] && (76)\\
&\leq \mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x},Z=\zeta}\left[\mathrm{IL}_k(F_k^*(\zeta|\mathbf{x}),(T,\Delta))\right] && (77)
\end{aligned}$$

And since the expectation is non-decreasing, we have:

$$\mathbb{E}_{T^*,C,\Delta}\left[\mathrm{IL}_k(\hat{F}_k(Z|\mathbf{x}),(T,\Delta))|\mathbf{X}=\mathbf{x}\right] \leq \mathbb{E}_{T^*,C,\Delta}\left[\mathrm{IL}_k(F_k^*(Z|\mathbf{x}),(T,\Delta))|\mathbf{X}=\mathbf{x}\right] \tag{78}$$

This allows us to consider the Integrated Loss (IL) as a global proper scoring rule for comparing different competing risks models. □

D Examples

D.1 Brier Score

When we define $l(y,\hat{y}) \stackrel{\text{def}}{=} (y-\hat{y})^2$, we obtain the censoring-adjusted Brier score for the k^{th} competing event, as defined in equation 14 of Kretowska [2018]:

Definition D.1.

$\forall\zeta, \forall k \in [1,K]$,

$$\mathrm{BS}_k(\hat{F}_k(\zeta,\mathbf{x}),\delta,t,\zeta,\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}_{t_i\leq\zeta,\delta_i=k}\left(1-\hat{F}_k(\zeta|\mathbf{x}_i)\right)^2}{G^*(t_i|\mathbf{x}_i)} + \frac{\mathbb{1}_{t_i>\zeta}\left(\hat{F}_k(\zeta|\mathbf{x}_i)\right)^2}{G^*(\zeta|\mathbf{x}_i)} + \frac{\mathbb{1}_{t_i\leq\zeta,\delta_i\neq 0,\delta_i\neq k}\left(\hat{F}_k(\zeta|\mathbf{x}_i)\right)^2}{G^*(t_i|\mathbf{x}_i)} \tag{79}$$

D.2 Binary cross entropy loss

As explained by Benedetti [2010], the log loss captures uncertainty better than the mean squared error. Therefore, one could also evaluate survival analysis and competing risks models using the following loss.

$\forall k \in [1,K]$,

$$\mathrm{l}_k(\hat{F}_k(\zeta,\mathbf{x}),\delta,t,\zeta) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}_{t_i\leq\zeta,\delta_i=k}\log\left(\hat{F}_k(\zeta|\mathbf{x}_i)\right)}{G^*(t_i|\mathbf{x}_i)} + \frac{\mathbb{1}_{t_i\leq\zeta,\delta_i\neq 0,\delta_i\neq k}\log\left(1-\hat{F}_k(\zeta|\mathbf{x}_i)\right)}{G^*(t_i|\mathbf{x}_i)} + \frac{\mathbb{1}_{t_i>\zeta}\log\left(1-\hat{F}_k(\zeta|\mathbf{x}_i)\right)}{G^*(\zeta|\mathbf{x}_i)} \tag{80}$$

E The Yanagisawa [2023] scoring rule for survival

Yanagisawa [2023] introduce a metric called $S_{Cen-log-simple}$, which is an approximation of the proper scoring metric in Rindt et al. [2022]. The metric in Rindt et al. [2022] requires the hazard function, which is the time derivative of the cumulative incidence function. This derivative can only be computed by differentiable models, implying an implicit assumption on almost-everywhere smooth time dependence. To avoid the need for the hazard function, Yanagisawa [2023] approximate it as piecewise affine. They demonstrate that under the assumption that the "node time points" —the edges of the affine segments— match an actual piecewise-affine breakdown of the CIF, the resulting approximation is proper. They argue that with enough node time points, this metric serves as a good approximation of a proper scoring rule.

$S_{Cen-log-simple}$ is defined as:

$$S_{Cen-log-simple}(\hat{F},(t,\delta);\{\zeta_i\}_{i=0}^{B}) \stackrel{\text{def}}{=} -\delta\sum_{i=0}^{B-1}\mathbb{1}_{\zeta_i<t\leq\zeta_{i+1}}\log(\hat{F}(\zeta_{i+1})-\hat{F}(\zeta_i)) - (1-\delta)\sum_{i=0}^{B-1}\mathbb{1}_{\zeta_i<t\leq\zeta_{i+1}}\log(1-\hat{F}(\zeta_{i+1})) \tag{81}$$

where B is the number of node time points[2], and $\{\zeta_i\}_{i=0}^{B}$ are the node times points, evenly spaced between 0 and t_{max}, dividing the time space into B equal intervals.

F Pseudo-code

Algorithm 2 IPCW Computer

Input: $\mathbf{x}, \delta, t, \hat{G}$
$y \leftarrow \delta \ \mathbb{1}_{t \leq \zeta}$ ▷Computing the target
if $t > \zeta$ **then** ▷The observation is not censored
  $w \leftarrow \frac{1}{\hat{G}(\zeta|\mathbf{x})}$
else if $t \leq \zeta$ and $\delta \neq 0$ **then**
  $w \leftarrow \frac{1}{\hat{G}(t|\mathbf{x})}$
else
  $w \leftarrow 0$
end if
return y, w

Algorithm 3 Censoring Feedback Loop - One Iteration

Input: $\mathbf{x}, \delta, t, \hat{S}$
for $i = 1$ **to** $n_{samples}$ **do**
  $\zeta_i \sim \mathcal{U}(0, t_{max})$
end for
$\zeta \leftarrow (\zeta_i)_{1 \leq i \leq n_{samples}}$
$\tilde{\mathbf{x}} \leftarrow (\mathbf{x}, \zeta)$
$\delta \leftarrow \mathbb{1}_{y=0}$ ▷Changing the target (focusing on the censoring distribution)
$y, w \leftarrow \text{ipcwcomputer}(\mathbf{x}, \delta, t, \hat{S})$ ▷See Alg 2
$L \leftarrow \frac{1}{n} \sum_{i=1}^{n} \left(y_i \ w_i \ \log\left(1 - \hat{G}_k(\zeta_i|\mathbf{x}_i)\right)\right) + (1 - y_i) \ w_i \log\left(\hat{G}(\zeta_i|\mathbf{x}_i)\right)$
$\tilde{h}_m(\tilde{\mathbf{x}}) \leftarrow$ Train one iteration of Gradient Boost with L as the loss ▷$\tilde{h}_m$ is the m^{th} weak learner
$\tilde{H}_m(\zeta|\mathbf{x}) \leftarrow \tilde{h}_m(\zeta|\mathbf{x}) + \nu \tilde{H}_{m-1}(\zeta|\mathbf{x})$ ▷$\tilde{H}_m$ is the m^{th} estimator
$((1 - \hat{G})(\zeta|\mathbf{X} = \mathbf{x}), \hat{G}(\zeta|\mathbf{X} = \mathbf{x})) \leftarrow \tilde{H}_m(\tilde{\mathbf{x}})$
return $\hat{G}(\zeta|\mathbf{X} = \mathbf{x})$

G Additional results for competing risk experiments

G.1 Results in the survival analysis setting

G.1.1 KKBOX

Here, we present the results of the experiments conducted on the KKBOX dataset (Figures S1 and 5). We highlight the trade-offs observed to assess the scalability of the models. Specifically, the models were trained on KKBOX using subsamples of 100k, 1M, and 2M training data points. However, due to computational constraints, it was not possible to run some experiments with 1M or 2M data points.

G.2 Trade-off between training time and performances

Here, we provide the results of our analysis of training time with the performances on the $S_{Cen-log-simple}$ of the different models for the survival analysis.

[2]We use $B = 32$, as in the experiments in Yanagisawa [2023]

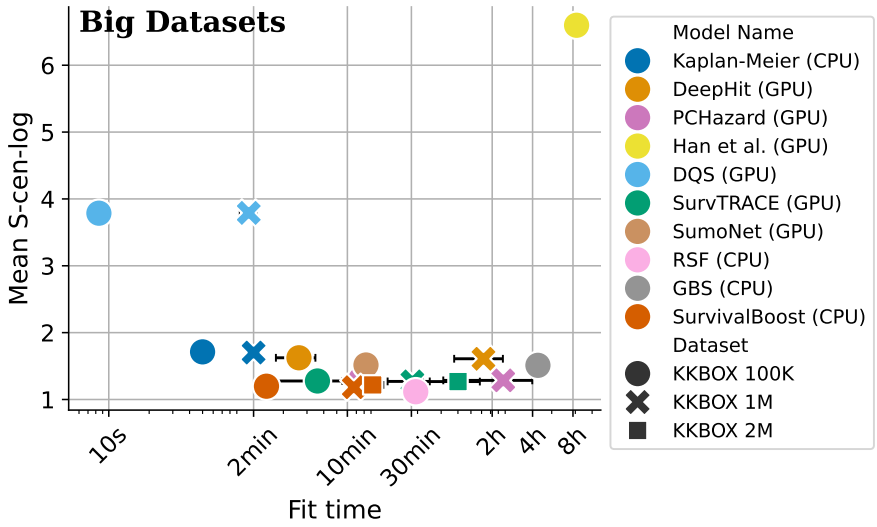

Figure S1: **Trade-off between S_{C-l-s} and fitting time for different sample sizes on the KKBOX dataset**

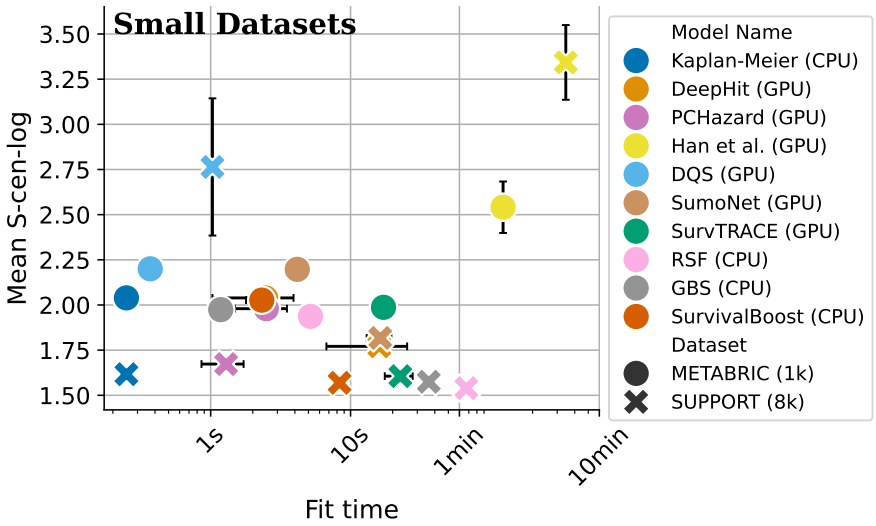

Figure S2: Trade-off between performance and the training time for the $S_{Cen-log-simple}$ metric for the survival model on METABRIC and SUPPORT datasets.

G.3 Results for the SEER Dataset

Learning curves We conducted experiments while varying the number of training points, measuring the KM-adjusted Integrated Brier Score (IBS) for each event. Additionally, we averaged the scores to obtain a global metric. The IBS was computed for each event while training on the full dataset, except for Random Survival Forests, which was trained on 100k data points, and Fine and Gray, which was on 10k data points due to computational limitations. In Table S2, we compare our method with other models, showing that SurvivalBoost outperforms the alternatives. Furthermore, figure 3 illustrates that the models with the best average IBS are also the fastest to train.

C_ζ-index The C-index measures whether the ranking of the risk for different samples aligns with the order of the times when the event of interest occurs[Harrell et al., 1982]. While it was originally developed as a metric for

Table S2: Integrated Brier Score for each cause-specific risk on the SEER Dataset (Lower is better).

Event	1	2	3
Aalen-Johansen	0.1209	0.2832	0.0834
Fine & Gray	0.1055	0.0281	0.0822
Random Survival Forests	**0.0825**	0.0295	0.0803
DeepHit	0.0931	0.0330	0.0831
DSM	0.0875	0.0310	0.0869
DeSurv	0.0975	0.0327	0.0869
SurvTRACE	0.0871	0.0287	0.0800
SurvivalBoost	0.0832	**0.0273**	**0.0757**

survival analysis, it is often adapted to competing risks settings, where it is applied independently to each event [Uno et al., 2011]. However, in such settings, the C-index is biased and does not account for the probabilities of the events. Nonetheless, due to its popularity, we have included it in our experiments.

The tables below present the C_ζ-index over time for the three events S3. At a fixed time horizon ζ, we compute the C_ζ-index for each class, which corresponds to the ROC-AUC, accounting for censored observations. The time horizons ζ are selected based on the any-event distribution, representing quantiles. For instance, at the time corresponding to 0.25, 25% of the events have already occurred. These results differ from those in the SurvTRACE paper [Wang and Sun, 2022] for two main reasons: *1)* The available code online only implements one of their loss functions, *2)* they treated the SEER dataset with two competing risks, classifying any other event as censored, whereas we categorized other events as a third competing risk.

Table S3: C-index for competing risks on the SEER Dataset (Higher is better)

Time-horizon quantile	0.25			0.50			0.75		
Event	1	2	3	1	2	3	1	2	3
Aalen Johansen	.5±.0	.5.±.0	.5.±.0	.5.±.0	.5.±.0	.5.±.0	.5.±.0	.5.±.0	.5.±.0
Fine & Gray	.79.±.01	.67.±.01	.67.±.02	.76.±.01	.66.±.02	.67.±.01	.74.±.01	.66.±.01	.69.±.01
DeepHit	.86.±.01	.72.±.02	.73.±.01	.83.±.0	.70.±.02	.70.±.01	.81.±.01	.68.±.02	.69.±.02
DSM	.87.±.01	**.76.±.01**	.74.±.01	.84.±.01	**.73.±.01**	.72.±.01	.82.±.01	**.72.±.01**	**.72.±.01**
DeSurv	.82.±.01	.70.±.03	.70.±.01	.80.±.01	.69.±.0	.70± .01	.79.±.01	.68.±.01	.71.±.01
SurvTRACE	**.88.±.01**	**.76**.±.01	**.76.±.01**	**.85.±.01**	**.73.±.01**	**.73.±.01**	**.83.±.01**	.71.±.01	**.72.±.01**
SurvivalBoost	.87.±.01	.75.±.01	.74.±.01	.84.±.01	.72.±.01	.72.±.01	.80.±.01	.64.±.01	.62.±.01

H Additional results for survival experiments

H.1 Metrics for the survival analysis

Table S4: **METABRIC**: $S_{Cen-log-simple}$ and C-index

Model Name	S_{C-l-s} (↓)	C-index 0.25 (↑)	C-index 0.5 (↑)	C-index 0.75 (↑)
Kaplan-Meier	2.0393 ± 0.2184	0.5000 ± 0.0000	0.5000 ± 0.0000	0.5000 ± 0.0000
DeepHit	2.0391 ± 0.0005	0.6559 ± 0.0123	0.5918 ± 0.0236	0.6036 ± 0.0226
PCHazard	1.9796 ± 0.0855	0.6633 ± 0.0145	0.6356 ± 0.0112	0.6342 ± 0.0034
Han et al.	2.6648 ± 0.0356	0.6770 ± 0.0341	**0.6537 ± 0.0318**	**0.6407 ± 0.0074**
DQS	2.2002 ± 0.0000	0.6554 ± 0.0126	0.6215 ± 0.0091	0.6275 ± 0.0018
SumoNet	2.1973 ± 0.0000	**0.6872 ± 0.0230**	0.6428 ± 0.0107	0.6292 ± 0.0084
SurvTRACE	1.9871 ± 0.0876	0.6598 ± 0.0094	0.6377 ± 0.0079	0.6357 ± 0.0108
RSF	1.9371 ± 0.2265	0.6736 ± 0.0135	0.6398 ± 0.0101	0.6335 ± 0.0097
GBS	**1.9742 ± 0.4043**	0.6402 ± 0.0131	0.6399 ± 0.0122	0.6388 ± 0.0101
SurvivalBoost	2.0269 ± 0.1592	0.6685 ± 0.0099	0.6374 ± 0.0106	0.6159 ± 0.0082

TABLE S5: **METABRIC:** METRICS.

MODEL NAME	IBS (↓)	MSE (↓)	MAE (↓)	AUC (↑)
KAPLAN-MEIER	0.1854 ± 0.0103	16007.1 ± 2100.4	102.3 ± 2.5	0.5000 ± 0.0000
DEEPHIT	0.1707 ± 0.0086	16229.1 ± 1645.0	98.5 ± 2.3	0.6737 ± 0.0256
PCHAZARD	0.1685 ± 0.011	15374.2 ± 2134.5	93.3 ± 3.2	0.6871 ± 0.0173
HAN ET AL.	0.1959 ± 0.0036	**13714.0 ± 1349.0**	95.5 ± 2.4	0.6752 ± 0.0113
DQS	0.1717 ± 0.018	16833.8 ± 1777.9	97.3 ± 2.5	0.6792 ± 0.0164
SUMONET	0.1698 ± 0.0098	40239.2 ± 1936.9	179.8 ± 3.2	0.5000 ± 0.0000
SURVTRACE	0.1723 ± 0.0064	22733.4 ± 1382.3	109.7 ± 4.7	0.6962 ± 0.0102
RSF	**0.1651 ± 0.0084**	15154.4 ± 1445.0	94.3 ± 1.2	**0.7023 ± 0.0129**
GBS	0.1686 ± 0.0107	14265.3 ± 2025.0	91.6 ± 3.4	0.6896 ± 0.0123
SURVIVALBOOST	0.1679 ± 0.0116	14208.1 ± 1762.8	**91.5 ± 2.7**	0.6993 ± 0.0170

TABLE S6: **SUPPORT**: $S_{Cen-log-simple}$ AND C-INDEX

MODEL NAME	S_{C-l-s} (↓)	C-INDEX 0.25 (↑)	C-INDEX 0.5 (↑)	C-INDEX 0.75 (↑)
KAPLAN-MEIER	1.6169 ± 0.2680	0.5000 ± 0.0000	0.5000 ± 0.0000	0.5000 ± 0.0000
DEEPHIT	2.249±.009	0.5546 ± 0.0158	0.5575 ± 0.0163	0.5600 ± 0.0196
PCHAZARD	1.6730 ± 0.0040	0.6121 ± 0.0052	0.6077 ± 0.0047	0.6054 ± 0.0044
HAN ET AL.	3.2227 ± 0.0054	0.5920 ± 0.0235	0.5740 ± 0.0187	0.5713 ± 0.0143
DQS	2.7641 ± 0.1281	0.5741 ± 0.0043	0.5682 ± 0.0033	0.5645 ± 0.0038
SUMONET	1.8175 ± 0.0000	0.5948 ± 0.0050	0.5952 ± 0.0052	0.5970 ± 0.0050
SURVTRACE	1.6061 ± 0.0026	0.6101 ± 0.0052	0.6099 ± 0.0038	0.6073 ± 0.0030
RSF	1.9421 ± 0.0229	**0.6174 ± 0.0058**	0.6137 ± 0.0045	0.6104 ± 0.0047
GBS	1.5750 ± 0.0002	0.6136 ± 0.0108	0.6140 ± 0.0100	**0.6143 ± 0.0099**
SURVIVALBOOST	1.5692 ± 0.3413	0.6165 ± 0.0052	**0.6159 ± 0.0044**	0.6138 ± 0.0044

TABLE S7: **SUPPORT:** METRICS.

MODEL NAME	IBS (↓)	MSE (↓)	MAE (↓)	AUC (↑)
KAPLAN-MEIER	0.2077 ± 0.004	1503075.2 ± 34398.0	904.4 ± 7.3	0.5000 ± 0.0000
DEEPHIT	0.2061 ± 0.0058	1416882.2 ± 33011.4	898.4 ± 14.0	0.6061 ± 0.0321
PCHAZARD	0.1867 ± 0.0036	1317674.8 ± 27353.9	843.0 ± 8.8	0.6578 ± 0.0074
HAN ET AL.	0.2539 ± 0.0015	1417630.2 ± 40832.6	881.5 ± 23.0	0.5906 ± 0.0139
DQS	0.2025 ± 0.004	1499067.0 ± 44660.7	876.3 ± 11.1	0.5979 ± 0.0029
SUMONET	0.1942 ± 0.0056	1857007.8 ± 36240.4	967.4 ± 8.1	0.5000 ± 0.0000
SURVTRACE	0.1876 ± 0.0037	1294800.3 ± 14983.6	849.8 ± 12.5	0.6555 ± 0.0070
RSF	0.1815 ± 0.0041	1347923.5 ± 53819.8	842.9 ± 15.2	**0.6750 ± 0.0094**
GBS	0.187 ± 0.0041	1292740.7 ± 26527.7	847.8 ± 10.0	0.6617 ± 0.0128
SURVIVALBOOST	**0.1814 ± 0.0049**	**1216995.5 ± 34370.6**	**827.2 ± 12.2**	0.6704 ± 0.0086

TABLE S8: **KKBOX (100K DATA POINTS):** METRICS.

MODEL NAME	IBS (↓)	MSE (↓)	MAE (↓)	AUC (↑)
KAPLAN-MEIER	0.2131 ± 0.0007	177438.3 ± 2250.0	345.3 ± 1.2	0.5000 ± 0.0000
DEEPHIT	0.1523 ± 0.0007	113033.6 ± 593.1	245.7 ± 0.5	0.9397 ± 0.0052
PCHAZARD	0.1095 ± 0.0001	100153.0 ± 1925.2	213.5 ± 3.2	0.9431 ± 0.0046
HAN ET AL. (NLL)	0.1715 ± 0.0036	111820.2 ± 0.0	245.6 ± 0.0	0.8881 ± 0.0086
DQS	0.1301 ± 0.0013	**93820.2 ± 4140.5**	**204.9 ± 2.2**	0.9228 ± 0.0071
SUMONET	0.1078 ± 0.0	224981.4 ± 0.0	360.3 ± 0.0	0.5000 ± 0.0000
SURVTRACE	0.1107 ± 0.0006	133400.5 ± 1353.9	250.0 ± 3.0	0.9379 ± 0.0004
RSF	0.1068 ± 0.0	911586.7 ± 0.0	423.6 ± 0.0	**0.9449 ± 0.0000**
GBS	0.1567 ± 0.0	123348.9 ± 0.0	254.5 ± 0.0	0.8958 ± 0.0000
SURVIVALBOOST	**0.1052 ± 0.0006**	101103.9 ± 9688.4	207.2 ± 4.3	0.9322 ± 0.0006

TABLE S9: **SURVIVAL CALIBRATION METRICS.** RESULTS ARE MARKED WITH ✓ IF THE MODEL IS CALIBRATED AND - OTHERWISE, WITH THE SIGNIFICANCE LEVEL FIXED AT $p_{value} = 0.05$.

DATASET / MODEL	METABRIC				SUPPORT				KKBOX				TOTAL TESTS SUCCESFULL
	KM_c	X_c	D_c	one_c	KM_c	X_c	D_c	one_c	KM_c	X_c	D_c	one_c	
KAPLAN-MEIER	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	-	✓	**10**
DEEPHIT	✓	✓	-	-	✓	-	-	-	✓	✓	-	-	5
PCHAZARD	✓	✓	✓	✓	✓	✓	-	-	✓	✓	-	-	8
HAN ET AL. [2021]	✓	✓	-	✓	✓	✓	-	-	✓	-	-	-	6
DQS	✓	✓	-	-	✓	-	✓	-	✓	-	-	-	5
SUMONET	-	-	-	-	-	-	-	-	-	-	-	-	0
SURVTRACE	-	✓	✓	✓	✓	✓	-	-	✓	✓	-	-	7
RSF	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	-	-	**9**
GBS	✓	✓	✓	✓	✓	✓	-	-	✓	✓	-	-	8
SURVIVALBOOST	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	-	-	**9**

I Implementation Details

I.1 Computing Infrastructure

To conduct our experiments, we used an iternal cluster.
The neural network were trained onto a 4x NVIDIA Tesla V100 32GB GPU with 40 CPUs and 252Gb RAM.
The others methods that do not need GPUs were trained onto a cluster with 48CPUs and 504Gb RAM. We chose to allow only 50Gb RAM for each model.

I.2 Reference of used implementations for baselines

We compare SURVIVALBOOST with several baselines, outlining their main characteristics and the implementation used in Table S10

I.3 GridSearch Parameters

We performed a Randomized Search for these parameters with a budget of 30 iterations. There are no parameters to tune for the Aalen-Johansen and Fine & Gray models.

J Distribution of the competing risks datasets

J.1 SEER Distribution of events

Here, we present the distributions for both competing risks datasets: the SEER Dataset S3 and the synthetic dataset S4. Notably, the censoring distribution is non-uniform over time. Figure S4 illustrates an example of the event distribution with censoring, which is dependent on the covariates. The parameters were selected to represent three distinct behaviors.

Table S10: Characteristics of used baselines.

Name	Competing risks	Proper loss	Implementation	Reference
SurvTRACE	✓		ours	Wang and Sun [2022]
DeepHit	✓		`github.com/havakv/pycox`	Lee et al. [2018]
DSM	✓		`autonlab.github.io/DeepSurvivalMachines`	Nagpal et al. [2021]
DeSurv	✓		`github.com/djdanks/DeSurv`	Danks and Yau [2022]
Random Survival Forests	✓		`scikit-survival.readthedocs.io/` for survival, and `www.randomforestsrc.org/` for competing risks	Ishwaran et al. [2008, 2014]
Fine & Gray	✓		`cran.r-project.org/package=cmprsk`	Fine and Gray [1999]
Aalen-Johansen	✓		ours	Aalen et al. [2008]
Han et al.			`github.com/rajesh-lab/Inverse-Weighted-Survival-Games`	Han et al. [2021]
PCHazard			`github.com/havakv/pycox`	Kvamme and Borgan [2019b]
SumoNet		✓	`github.com/MrHuff/Sumo-Net`	Rindt et al. [2022]
DQS		✓	`ibm.github.io/dqs/`	Yanagisawa [2023]

Table S11: Randomized Search Parameters

Estimator	Parameter	Range
SurvivalBoost	Learning Rate	$loguniform(0.01, 0.5)$
	Nb of iterations	$[\![20, 200]\!]$
	Maximum Depth	$[\![2, 10]\!]$
	Nb of times	$[\![1, 5]\!]$
SurvTRACE	Learning Rate	$loguniform(10^{-5}, 10^{-3})$
	Batch Size	$\{256, 512, 1024\}$
	Hidden parameter	$\{2, 3\}$

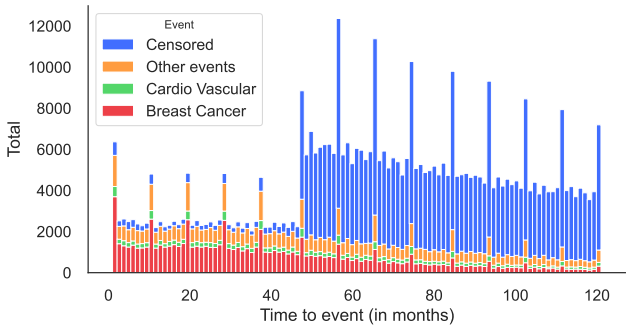

Figure S3: **SEER Dataset Distributions** The censoring rate is approximately 63%. The prevalence of events is 18% for breast cancer, 4.5% for cardiovascular events, and 10% for other events. The shift in the censoring distribution after the 48^{th} month may be challenging for some methods to learn.

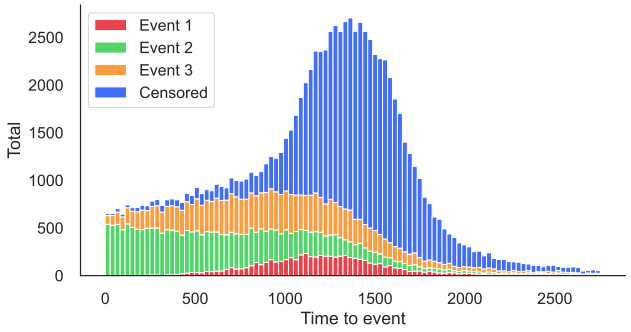

Figure S4: **Synthetic Dataset Distributions** Duration distributions of the synthetic dataset with censoring dependent on X, and a censoring rate of 69%. The events are stacked. To illustrate this distribution, consider truck maintenance. Event 1, occurring throughout the duration, corresponds to drivers' driving skills. Event 2 may represent a design flaw in the trucks, occurring from the start. Event 3 refers to trucks' wear and tear over time.