Teaching Models To Survive: Proper Scoring Rule and Stochastic Optimization with Competing Risks

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

Teaching Models To Survive: Proper Scoring Rule and Stochastic Optimization with Competing Risks

Julie Alberge
SODA Team, Inria Saclay
Palaiseau, France
julie.alberge@inria.fr

Vincent Maladière
:probabl.
Paris, France
vincent@probabl.ai

Olivier Grisel
:probabl.
Paris, France

Judith Abécassis
SODA Team, Inria Saclay
Palaiseau, France

Gaël Varoquaux
SODA Team, Inria Saclay
Palaiseau, France

Abstract

When data are right-censored, *i.e.* some outcomes are missing due to a limited period of observation, survival analysis can compute the "time to event". Multiple classes of outcomes lead to a classification variant: predicting the most likely event, known as competing risks, which has been less studied. To build a loss that estimates outcome probabilities for such settings, we introduce a strictly proper censoring-adjusted separable scoring rule that can be optimized on a subpart of the data because the evaluation is made independently of observations. It enables stochastic optimization for competing risks which we use to train gradient boosting trees. Compared to 11 state-of-the-art models, this model, MultiIncidence, performs best in estimating the probability of outcomes in survival and competing risks. It can predict at any time horizon and is much faster than existing alternatives.

1 Introduction

We all die at some point; some applications call for predicting not whether an event of interest will happen or not, but when it is likely to occur: *time-to-event regression*. In such a setting, 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; valid models use dedicated corrections for censorship: *survival analysis* models. These have long been central to health (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 & Russo). Survival analysis has led to many dedicated models, such as the Kaplan & Meier (1958) estimator or the Cox (1972) proportional hazard model.

Competing risks analysis generalizes survival analysis to account for multiple events, determining which will happen first (Susto et al., 2015; Gaynor et al., 1993). For instance, if a person with breast cancer 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. (National Cancer Institute, 2023). The caregiver may also want to adapt the treatment if the patient is predicted to 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 instance, 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 a single event of interest will be biased if competing risks

Preprint. Under review.

are not included. Hence, adequate models for these risks are critical for decision-making (Ramspek et al., 2022; Koller et al., 2012; van Walraven & 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 & Sun, 2022). Given that the risk, which is the probability of the outcome at a specific time, is crucial for various applications, it can be preferable to use losses that directly control probabilities (proper scoring rules), as 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 typically 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 introduced by 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 to learn a competing risks model with a proper scoring rule. This scoring rule gives a loss easy to plug into any multiclass estimator to create a competing risks model: giving the individual risk of each event at any horizon. We also sum over time for model evaluation, as the resulting Integrated Scoring Rule is also proper.
An interesting property of this new loss is that it can be optimized on a subset of the training data because the evaluation is made independently of observations. Hence, it allows stochastic optimization, enabling computationally efficient learning. With that, we propose an algorithm called MultiIncidence, based on Stochastic Gradient Boosting Trees. We benchmark our algorithm on a synthetic dataset with varying censoring rates, number of features, and number of training samples to show that our method outperforms state-of-the-art methods while exhibiting faster training times. Finally, applying our model to real-life datasets demonstrates that it outperforms other models in both the competing risks context and basic survival analysis.

2 Related work

Survival settings Various survival models have been developed, ranging from approaches like the Kaplan & Meier (1958) estimator, estimating the general survival curve of a whole population, to models that account for covariates. One of them is the Cox (1972) Proportional-Hazards Model, a linear model of *hazard*: the instantaneous probability of an event, *i.e.* the logarithmic derivative of outcome probabilities in 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 & Borgan, 2019b). Although the above do not control risks, more recent neural networks use adequate losses (see below): DQS (Yanagisawa, 2023, though relying on a piecewise constant hazard), SumoNet (Rindt et al., 2022, which requires differentiable models).

Competing risks Competing risks, with multiple outcomes, require new methods (which can naturally adapt to the simpler survival setting). Derived from the Kaplan & Meier (1958) estimator, the Nelson (1972)-Aalen et al. (2008) estimator is an unbiased marginal model for competing risks. The linear Fine & Gray (1999) estimator is inspired by the Cox (1972) estimator in survival analysis and is the most used model in clinical research. Machine-learning models have recently been adapted to the competing risks setting, including tree-based approaches such as the Random Survival Forests (Ishwaran et al., 2008; Kretowska, 2018; Bellot & Schaar, 2018), boosting approaches (Bellot & van der Schaar, 2018), neural networks approaches *e.g.* DeepHit and Gaussian mixtures approaches (Lee et al., 2018; Aala & van der Schaar, 2017; Danks & Yau, 2022a; Nagpal et al., 2021) and transformers approaches with SurvTRACE (Wang & Sun, 2022) using a loss corrected to predict rare competing events but independently forecasts all events without ensuring probabilities sum to one.
For a review of the competing setting, the reader can refer to Monterrubio-Gómez et al. (2022).

Evaluation for such models Prediction evaluation in survival or competing risks settings calls for adapted metrics to account for right-censored points (Harrell et al., 1982), like the C-index which adapts the Area Under the ROC curve in classification. However, the C-index only evaluates the ranking of samples, *i.e.* which samples will undergo the event of interest first, and is dependent on the censoring distribution which may bias the evaluation (Blanche et al., 2019; Rindt et al., 2022). In fact, the score may be higher for distributions other than oracle-censoring distributions. 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 given time horizon ζ. The C-index ranking metric has been extended to competing risks (Uno et al., 2011) but, as in the survival setting, the C-index only evaluates relative risks for pairs of individuals and not the absolute value of the risk for a given individual. Other time-dependent adaptations of the ROC curve have been developed, also assessing a discriminative power rather than risks or probabilities (Blanche et al., 2013). And yet control of the risk is crucial to decision making (Van Calster et al., 2019). Proper scoring rules are alternatives to overcome the limitations of existing metrics because they capture more aspects of the problem. In addition, they can be used for both the training and evaluation of probabilistic predictive models.

Proper Scoring Rules (PSR) Scoring rules are functions of observations and a candidate probability distribution; when *proper* they control for the oracle probability distribution (definition 3.2). This is important in machine learning to create losses that recover probabilities of outcomes. For classification, where discrete events are observed rather than the probability, the Brier score and the log loss give proper scoring rules, with relative merits (Benedetti, 2010; Merkle & Steyvers, 2013).

Graf et al. (1999) adapt the Brier score to survival analysis, with a strong independence assumption on the censoring distribution. However, the assumption can easily be violated (Kvamme & Borgan, 2019a) which leads to bias (Rindt et al., 2022). Rindt et al. (2022) show that the likelihood of the survival function leads to a proper scoring rule but requires obtaining the density function and the survival function, a time-wise derivative of outcome probabilities (definition 3.1). For quantile regression, Yanagisawa (2023) shows that the Pinball loss may lead to a proper scoring rule for survival analysis but requires an oracle parameter. Han et al. (2021) introduces a double optimization problem for which the stationary point is located at the true distributions.

For competing risks, Schoop et al. (2011) extend the Brier score to a proper scoring rule. However, the Brier score does not measure the uncertainty as well 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 calligraphy style $\mathcal{A}$.

3.1 Problem setting

We consider K competing events and for $k \in [\![1, K]\!]$, we denote $T_k^* \in \mathbb{R}_+$ the event time of the event k, depending on the covariate $\mathbf{X} \in \mathcal{X}$. We also denote $T^* \in \mathbb{R}_+, T^* = \min_{k \in [\![1,K]\!]}(T_k^*)$ and $\Delta^* \in [\![1, K]\!], \Delta^* = \arg\min_{k \in [\![1,K]\!]}(T_k^*)$. We observe $(\mathbf{X}, T, \Delta) \sim \mathcal{D}$, with $T = \min(T^*, C)$ where $C \in \mathbb{R}_+$ is the censoring time, which may 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 interested in the distribution of the uncensored data $(\mathbf{X}, T^*, \Delta^*) \sim \mathcal{D}^*$ especially in the joint distribution of $T^*, \Delta^* | \mathbf{X} = \mathbf{x}$ and the marginal distribution of $T^* | \mathbf{X} = \mathbf{x}$.
This paper aims to predict an unbiased estimate of all of the cause-specific Cumulative Incidence functions (CIF) at any time horizon ζ chosen based on the observations $(\mathbf{x}, t, \delta)$:

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

CIF (cumulative incidence function):	$F^*(\zeta \| \mathbf{x}) = \mathbb{P}(T^* \leq \zeta \| \mathbf{X} = \mathbf{x})$
k^{th} CIF:	$F_k^*(\zeta \| \mathbf{x}) = \mathbb{P}(T^* \leq \zeta \cap \Delta^* = k \| \mathbf{X} = \mathbf{x})$
Censoring:	$G^*(\zeta \| \mathbf{x}) = \mathbb{P}(C > \zeta \| \mathbf{X} = \mathbf{x})$
Survival to any event:	$S^*(\zeta \| \mathbf{x}) = \mathbb{P}(T^* > \zeta \| \mathbf{X} = \mathbf{x})$

Assumption 3.1 (*Non informative censoring*)**.** We make the classic assumption of survival analysis that the censoring is noninformative according to the covariates:

$$\forall k, \in [\![1, K]\!], \quad T_k^* \perp\!\!\!\perp C | \mathbf{X}$$

Assumption 3.1 needed for most theoretical results in survival (Rindt et al., 2022; Yanagisawa, 2023; Han et al., 2021). It is key to understanding why single-event survival analysis is invalid in the presence of competing risks: if some observations are censored due to other events sharing 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 and gives a corresponding score $\ell(\mathcal{P}, Y)$. The better the score, the better the model fits the observation. For a proper scoring rule, it corresponds to the degree to which the model can predict the oracle distribution (more on scoring rules in Gneiting & Raftery, 2007; Ovcharov, 2018; Merkle & Steyvers, 2013).

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

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

When equality is reached if and only if $\mathcal{P} = \mathcal{Q}$, the scoring rule is called strictly proper.

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

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

$$\forall \zeta, (\mathbf{X}, T, \Delta) \sim \mathcal{D},$$

Estimated distributions

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

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

Oracle distributions

4 A Proper Scoring Rule for Competing Risks

We prove that the negative log-likelihood re-weighted by the censoring distribution (IPCW) is 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 of all the cumulative incidence functions as well as the survival function.

$$\forall \zeta, (\mathbf{x}, t, \delta) \sim \mathcal{D}, \quad \mathrm{L}_\zeta((\hat{F}_1(\zeta|\mathbf{x}), ..., \hat{F}_K(\zeta|\mathbf{x}), \hat{S}(\zeta|\mathbf{x})), (t, \delta)) \stackrel{\text{def}}{=}$$

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\frac{\mathbb{1}_{t_i\leq\zeta,\delta_i=k}\ \log\left(\hat{F}_k(\zeta|\mathbf{x}_i)\right)}{G^*(t_i|\mathbf{x}_i)}\right) + \frac{\mathbb{1}_{t_i>\zeta}\ \log\left(\hat{S}(\zeta|\mathbf{x}_i)\right)}{G^*(\zeta|\mathbf{x}_i)} \quad (2)$$

Probability of remaining at t_i

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

Eqn.2 can be seen as a standard log-loss (a.k.a cross-entropy), reweighted by appropriate sample weights, the inverse probabilities, IPCW (inverse probabilities of censoring weights). It can thus be easily 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:* $\forall \zeta, (\mathbf{X}, T, \Delta) \sim \mathcal{D}$,

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

Proof sketch. The weights enable moving from the observation distribution T to the distribution of T^*, a key ingredient to show properness. The whole proof can be found in Appendix B. □

Theorem 1 (Properness of the scoring rule)**.** *Under the assumption that the weights are well 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. With the previous result, the properties of the negative log-likelihood, and the Definition 3.3, we obtain that the loss is strictly proper. The whole proof can be found in Appendix B. □

5 MultiIncidence Model: Gradient boosting for competing risks

While eq.2 can be used as a loss in any multiclass machine learning algorithm, we chose Gradient Boosting trees because of their performance on tabular data (Grinsztajn et al., 2022) and their ability to be fit via stochastic optimization. Most survival or competing risk loss cannot be used with such tree-based models as the require time-derivates and thus smoothness.

We introduce a model, MultiIncidence, predicting all of the CIF for each competing event as well as the global survival function. Predicting these jointly easily maintains the stability of the probabilities as outputs of classifications model sum to one and $\mathbb{P}(T^* \leq \zeta | \mathbf{X} = \mathbf{x}) + \mathbb{P}(T^* > \zeta | \mathbf{X} = \mathbf{x}) = 1$ or

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

With loss presented in Eq.3 we can directly predict the CIF instead of predicting the hazards function (the derivative of the CIF) as often done –*e.g.* DeepHit (Lee et al., 2018) or SurvTRACE (Wang & Sun, 2022). This allows us to drop the constant-hazard hypothesis (Yanagisawa, 2023; Kvamme & Borgan, 2019b; Wang & Sun, 2022; Rindt et al., 2022).

Our algorithm uses two classifiers (here gradient-boosted trees), one for the censoring trained on binary censored/non-censored labels (i.e. for time ζ, $\mathbb{P}(C > \zeta | \mathbf{X} = \mathbf{x})$), and a classifier for the multiple events. Both of the censoring and event models are corrected with IPCW weights. To compute these IPCW we iterate the training using a feedback loop (in the like of boosting). We first compute a survival censoring model. Then, with these probabilities, we initiate our MultiIncidence model. After several iterations, we apply a feedback loop to retrain our censoring model.
To model complex time dependence, time is stacked as an additional feature. At each iteration, we sample different times for each sample and stack the different features as well as the targets to provide more information to our algorithm. This approach is made possible by our loss which is separable. An additional benefit is that we can predict the CIF at any time, unlike models that are optimized for a limited number of times (such as SurvTRACE) and need to be interpolated to other times.

As Figure 1 shows an iteration: we compute the weights for each sample, as well as the target according to the sampled time. A censored sample will have a weight equal to 0 (due to the indicator

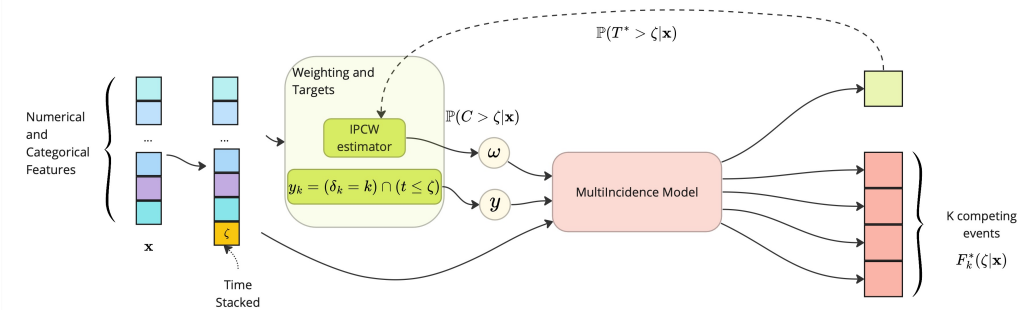


Figure 1: **MultiIncidence Model with its Feedback Loop.** After giving the input to the model, a random time is given and the weights and the target can be computed. After one iteration, the feedback loop trains the censoring probability – $G^\star$ in eq.2.

Algorithm 1 MultiIncidence Algorithm - Training

Input: $\mathbf{x}, \delta, t$
Ensure: $\min(t) > 0$
$\hat{G} \leftarrow$ Train $n_{censoring}$ iterations the censoring algorithm
for $m = 1$ **to** n_{iter} **do** ▷Boosting iterations
  for $i = 1$ **to** $n_{samples}$ **do**
    $\zeta_i \sim \mathcal{U}(0, t_{max})$ ▷Sample a random time horizon
    $y_i \leftarrow \delta_i \, (t_i \leq \zeta_i)$ ▷Computing the target
    $w_i \leftarrow 0$
    if $t_i > \zeta_i$ **then** ▷The observation is not censored
      $w_i \leftarrow \frac{1}{\hat{G}(\zeta_i)}$
    else if $t_i \leq \zeta_i$ and $\delta_i \neq 0$ **then**
      $w_i \leftarrow \frac{1}{\hat{G}(t_i)}$
    end if
    $\tilde{\mathbf{x}}_i \leftarrow (\mathbf{x}_i, \zeta_i)$ ▷Stacking the time to the features
  end for
  $\zeta \leftarrow (\zeta_i)_{1 \leq i \leq n_{samples}}$
  $\hat{H}_m(\tilde{\mathbf{x}}) \leftarrow$ Train one iteration of Gradient Boost with $\hat{G}(\zeta | \mathbf{X} = \mathbf{x})$ ▷Hm is the m-th tree
  $(\hat{S}(\zeta | \mathbf{X} = \mathbf{x}), (\hat{F}_k(\zeta | \mathbf{X} = \mathbf{x})_{1 \leq k \leq K}) \leftarrow \hat{H}_m(\tilde{\mathbf{x}})$
  $\hat{G} \leftarrow$ Train one iteration the Censoring Feedback Loop with $\hat{S}(\zeta | \mathbf{X} = \mathbf{x})$
end for

functions in eq.2). For strictly positive weights, if the target is in $[\![1, K]\!]$, this will represent that the event of interest has happened before ζ. Finally, a target equal to 0 will notify that the sample has survived any event. We give a pseudocode of the algorithm 1.

6 Experimental study: Competing risks

6.1 Evaluation metrics for competing risks models

To evaluate the risks of the different events, we use two metrics[1].

Evaluating the predicted probability We use a proper scoring rule (PSR). To avoid a form of circularity in the evaluation, we do not use the PSR that our model optimizes but rather we extend that used by Graf et al. (1999) and Schoop et al. (2011): we apply it to the Brier Score rather than the log-loss (Appendix C details the formula and the formal proof that it is indeed proper). To evaluate the model at all times, we sum it over time, giving the *integrated Brier score* (IBS).

Prediction accuracy in time For many applications, as in predictive maintenance or medicine, a crucial information is: which is the first event that a subject may encounter. We use a validation metric to check for each sample whether observed events are predicted as the most likely, at given times, chosen as before with quantiles. *E.g.* for an individual that encounters event 2 at t, the probability of surviving before t should be the highest compared to the probabilities of encountering each event. We also want the probability of encountering event 2 after t to be the highest one. To do so, we adapt Multi-Class accuracy to different times:

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

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

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

6.2 Experimental settings

Synthetic Dataset We designed a synthetic dataset with linear relations between features and targets, as well as relations with the censoring distribution of the features (details in Appendix H.2). To create the synthetic dataset, for each sample, we draw $2n_{events}$ parameters from a normal law. Then, we draw the durations from a Weibull distribution for each event from those parameters. To determine the observation, we return the minimum duration with its associated event. Then, the censoring event is computed with the same method.

SEER Dataset This dataset follows more than 470k breast cancer patients for up to ten years with mortality due to various diseases as outcomes. The censoring is around 63% and Figure S9 shows the distribution of the events. Instead of Lee et al. (2018) (DeepHit) or Wang & Sun (2022) (SurvTRACE), which consider only the two most prevalent events and censor the rest, defeating the purpose of competing risks, we consider the SEER data set with 3 competing events, aggregating the other events in a third class. We remove some features following Wang & Sun (2022).

Baselines We compared our approach to 7 other models. Aalen-Johansen's estimator (Aalen et al., 2008), Fine & Gray's linear model (Fine & Gray, 1999), a tree-based approach with the Random Survival Forests (RSF, Ishwaran et al., 2008), and neural networks: DeepHit (Lee et al., 2018), Deep Survival Machines (DSM, Nagpal et al., 2021), DeSurv (Danks & Yau, 2022b) and a transformer model with SurvTRACE (Wang & Sun, 2022). DeepHit is trained with a ranking loss: the C-index summed with a negative log-likelihood, DSM uses a graphical method for feature encodings while DeSurv solves Ordinal Differential Equations for continuous predictions in time. SurvTRACE is trained for three-time horizons (based on quantiles of observed event times) and at time 0, while Aalen-Johansen and Fine & Gray are trained for all observed event times. In contrast, our method is trained on uniformly sampled time horizons, allowing predictions at any time. To compute the Integrated Brier Score over time, other methods require linear interpolation of their trained times. For times exceeding their trained times, we assume the incidence remains constant. To be fair across models, we use the same time budget for hyper-parameter tuning (grid in Appendix S7).

6.3 Results, competing risks

Synthetic dataset Figure 2 shows the trade-off between statistical performance (IBS) and training time for each model compared. With the synthetic dataset, we can compute an oracle IBS. MultiIncidence outperforms the other models over the IBS while being the fastest to train.

We also conduct different experiments on the synthetic dataset varying the number of training points (Figure S1), the censoring rate (Figure S3), and the number of features (Figure S2). More experiments on the synthetic data set can be found in the appendix E.1.

Results on SEER Dataset On the real-life dataset, we keep 30% of the data set to test the models. Figure 3 compares models with the Integrated Brier score (with Kaplan-Meier weights of Graf et al. (1999) due to lack of oracle). MultiIncidence achieves the best score and the shortest fit time. Random Survival Forest is not made to be used with that many samples (100k) and uses more than 50 Gb of RAM. MultiIncidence maintains its marked lead with much fewer training samples (Appendix E.2).

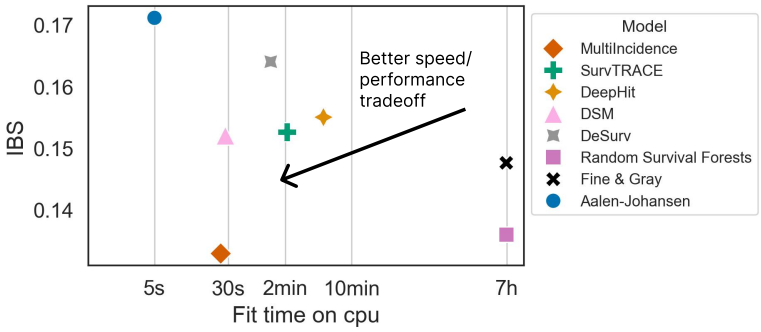


Figure 2: **Trade-off prediction/training time for competing risk on the synthetic dataset** Average IBS compared to the fitting time for each model on 20k training data points, censoring rate around 50%, and a dependant censoring for 6 features.

Figure 3: **Trade-off prediction/training time for competing risk on the SEER dataset** Average IBS compared to the fitting time for each model on the maximum training points (330k) except for Fine & Gray (50k) and RSF (100k). Table S2 gives IBS values for each event.

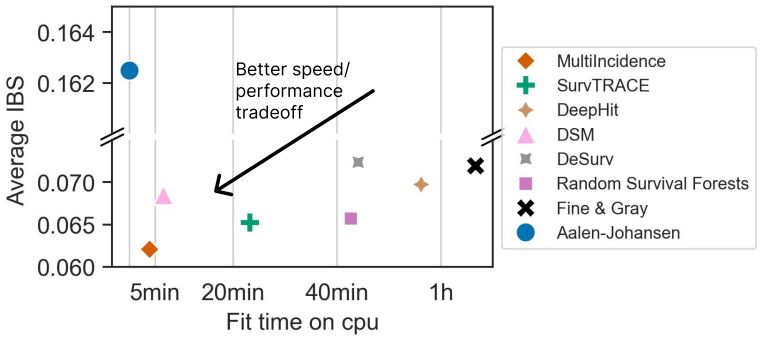
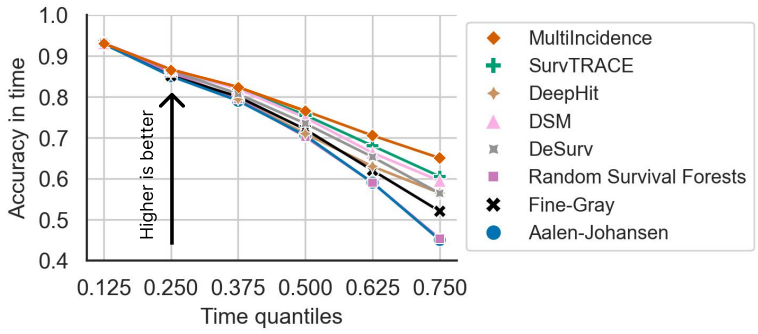


Figure 4: **Prediction accuracy at time** ζ Accuracy of the Argmax of the Cumulative Incidence Functions on different quantiles in time on the SEER Dataset (Higher is Better).



Event and time-specific C-indexes are presented in table S3, but do not capture the models' ability to predict which event is more likely to occur at a given time horizon. This is measured by accuracy in time in Figure 4, and MultiIncidence has the best performance. The benefit grows as time increases, meaning that it better interpolates in times.

7 Usage in Survival Analysis

7.1 Survival experiments

Real-life Datasets As our model can also handle survival analysis, we perform survival analysis on two real-life survival datasets: SUPPORT and METABRIC, both available in the Pycox library.

METABRIC The Molecular Taxonomy of Breast Cancer International Consortium is a dataset on gene expression with around 2k data points

SUPPORT Study to Understand Prognoses Preferences Outcomes and Risks of Treatment is a dataset on the survival time of hospital patients with more than 8k datapoints.

Evaluation We use different metrics to evaluate our models. As above we use the Integrated Brier Score (detailed in Appendix C), but we also add another metric from Yanagisawa (2023), called $S_{Cen-log-simple}$ (detailed in Appendix D). This last metric approximates the proper scoring metric in Rindt et al. (2022) –and is not exactly proper, see Appendix D. It is useful because it can be used on any model as it does not require the density of the Cumulative Incidence Function.

Baselines We compare our model with SOTA competing risks models, including SurvTRACE (Wang & Sun, 2022), DeepHit (Lee et al., 2018) and Random Survival Forests (Ishwaran et al., 2008). We also benchmark some SOTA survival ones: neural networks *e.g.* (PCHazard Kvamme & Borgan, 2019b), survival game (Han et al., 2021) and neural networks trained with a proper survival-analysis scoring rule, *e.g.* SumoNet (Rindt et al., 2022), and DQS (Yanagisawa, 2023).

Table 1: **Survival dataset**: Integrated Brier Score and $S_{Cen-log-simple}$ (Lower is Better)

Dataset	SUPPORT		METABRIC	
Model	IBS	$S_{Cen-log-simple}$	IBS	$S_{Cen-log-simple}$
Random Survival Forest	0.225±0.004	1.942±0.023	0.197±0.025	2.442±0.044
DeepHit	0.217±0.004	2.249±0.009	0.180±0.014	2.271±0.019
Han et al. (2021)	0.260±0.012	3.483±0.307	0.191±0.003	2.420±0.150
PCHazard	0.210±0.007	2.192±0.024	0.176±0.014	2.246±0.046
Han	0.260±0.012	3.483±0.307	0.191±0.003	2.420±0.150
DQS	0.202±0.007	1.987±0.069	0.180±0.034	2.205±0.044
SuMo net	0.194±0.010	**1.721±0.016**	0.169±0.009	**2.302±0.059**
SurvTRACE	0.194±0.005	1.870±0.018	**0.168±0.011**	2.270±0.034
MultiIncidence	**0.191±0.006**	1.740±0.020	**0.168±0.019**	**2.169±0.056**

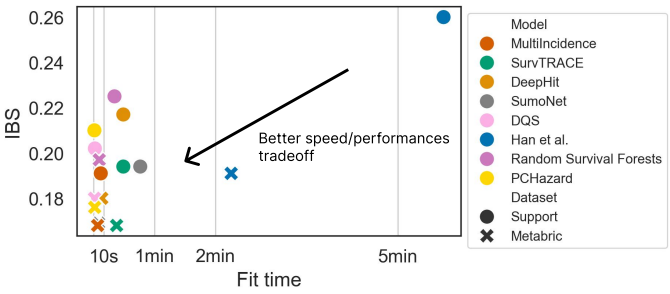


Figure 5: **Trade-off prediction/training time in survival usage** Performances (measured by IBS, integrated Brier score) function of fitting time for each model.

7.2 Results in survival usage

Prediction performance For both datasets, MultiIncidence achieves the best results on IBS and tied with Sumo Net for $S_{Cen-log-simple}$ (Table 1 and Appendix F.1 for the C-index). Sumo Net uses $S_{Cen-log-simple}$ as a training loss; note however that this metric is not guaranteed to be a proper scoring rule thus it does not ensure recovering the actual risks.

Computational time Figure 5 shows the trade-off between training time and performance in IBS, a trade-off that MultiIncidence excels at, being the best model for statistical performance and also one of the fastest. Appendix F.2 gives the same figure for the $S_{Cen-log-simple}$ metric, and MultiIncidence reaches a great trade-off rivaled only by SumoNet, which has competing performance on the $S_{Cen-log-simple}$ loss. Varying sample size from 1k to 100k on a synthetic dataset confirms that MultiIncidence and DQS are faster (less than 1min on 100k data points), Han et al., SumoNet, and Random Survival Forests slower for large sample size, with a super-linear time complexity for SumoNet and Random Survival Forests that makes them untractable for large data (Appendix E.1).

Discussion and Conclusion

Code reproducibility and data The code is available on GitHub as a library called hazardous.

Social impact Our contribution is not directly applied and has no immediate social impact, but we hope that it will improve medical applications where survival analysis is central.

Limitations and further work Further work should consider removing the assumption of noninformative censoring. This assumption is very common in the literature, though some recent works have relaxed it in survival settings (Foomani et al., 2023; Zhang et al., 2023).

Conclusion For competing risks, which is a generalization of survival analysis to classify the type of outcome, we first propose and prove a (strictly) proper scoring rule. It is a reweighted log loss that can easily be used as a loss for machine learning: it is separable in the observations and thus suited to stochastic solvers; it does not require time-wise derivative (unlike most survival models) and can be used in non-differentiable models. We plug it into gradient-boosting trees, in an algorithm called

MultiIncidence. Thanks to time used as a feature and its feedback loop to better estimate censoring probabilities, MultiIncidence outperforms state-of-the-art methods on a synthetic dataset as well as real-life datasets both for competing risk (classification on time-censored data) and standard survival (time-to-event regression with right censoring). It is also faster to train over many samples. As a loss, it easily brings survival or competing risks to many models: scalable linear models to replace clinical standard Fine & Gray that do not scale, or deep learning, including fine-tuning foundation models.

References

Aala, A. M. and van der Schaar, M. Deep Multi-task Gaussian Processes for Survival Analysis with Competing Risks. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/861dc9bd7f4e7dd3cccd534d0ae2a2e9-Abstract.html.

Aalen, O. O., Borgan, O., and Gjessing, H. K. *Survival and Event History Analysis*. Statistics for Biology and Health. Springer New York, New York, NY, 2008. ISBN 978-0-387-20287-7 978-0-387-68560-1. doi: 10.1007/978-0-387-68560-1. URL http://link.springer.com/10.1007/978-0-387-68560-1.

Antolini, L., Boracchi, P., and Biganzoli, E. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, December 2005. ISSN 0277-6715, 1097-0258. doi: 10.1002/sim.2427. URL https://onlinelibrary.wiley.com/doi/10.1002/sim.2427.

Bellot, A. and Schaar, M. Tree-based Bayesian Mixture Model for Competing Risks. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pp. 910–918. PMLR, March 2018. URL https://proceedings.mlr.press/v84/bellot18a.html. ISSN: 2640-3498.

Bellot, A. and van der Schaar, M. Multitask Boosting for Survival Analysis with Competing Risks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/2afe4567e1bf64d32a5527244d104cea-Abstract.html.

Benedetti, R. Scoring Rules for Forecast Verification. *Monthly Weather Review*, 138(1):203–211, January 2010. ISSN 1520-0493, 0027-0644. doi: 10.1175/2009MWR2945.1. URL http://journals.ametsoc.org/doi/10.1175/2009MWR2945.1.

Blanche, P., Dartigues, J.-F., and Jacqmin-Gadda, H. 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.

Blanche, P., Kattan, M. W., and Gerds, T. A. The c-index is not proper for the evaluation of t-year predicted risks. *Biostatistics*, 20(2):347–357, April 2019. ISSN 1465-4644, 1468-4357. doi: 10.1093/biostatistics/kxy006. URL https://academic.oup.com/biostatistics/article/20/2/347/4864363.

Chaddad, A., Desrosiers, C., and Toews, M. 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)*, pp. 4035–4038, Orlando, FL, USA, August 2016. IEEE. ISBN 978-1-4577-0220-4. doi: 10.1109/EMBC.2016.7591612. URL http://ieeexplore.ieee.org/document/7591612/.

Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, January 1972. ISSN 0035-9246, 2517-6161. doi: 10.1111/j.2517-6161.1972.tb00899.x. URL https://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1972.tb00899.x.

Danks, D. and Yau, C. Derivative-Based Neural Modelling of Cumulative Distribution Functions for Survival Analysis. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 7240–7256. PMLR, March 2022a. URL https://proceedings.mlr.press/v151/danks22a.html.

Danks, D. and Yau, C. Derivative-based neural modelling of cumulative distribution functions for survival analysis, 28–30 Mar 2022b. URL https://proceedings.mlr.press/v151/danks22a.html.

Fine, J. P. and Gray, R. J. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, 94(446):496–509, June 1999. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1999.10474144. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10474144.

Foomani, A. H. G., Cooper, M., Greiner, R., and Krishnan, R. G. Copula-based deep survival models for dependent censoring, 2023.

Gaynor, J. J., Feuer, E. J., Tan, C. C., Wu, D. H., Little, C. R., Straus, D. J., Clarkson, B. D., and Brennan, M. F. 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, June 1993. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1993.10476289. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476289.

Gneiting, T. and Raftery, A. E. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214506000001437. URL http://www.tandfonline.com/doi/abs/10.1198/016214506000001437.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, September 1999. ISSN 0277-6715, 1097-0258. doi: 10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5. URL https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. Why do tree-based models still outperform deep learning on tabular data?, July 2022. URL http://arxiv.org/abs/2207.08815. arXiv:2207.08815 [cs, stat].

Han, X., Goldstein, M., Puli, A., Wies, T., Perotte, A., and Ranganath, R. Inverse-weighted survival games. *Advances in neural information processing systems*, 34:2160–2172, 2021.

Harrell, Frank E., J., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546, 05 1982. ISSN 0098-7484. doi: 10.1001/jama.1982.03320430047030. URL https://doi.org/10.1001/jama.1982.03320430047030.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. Random survival forests. *The Annals of Applied Statistics*, 2(3), September 2008. ISSN 1932-6157. doi: 10.1214/08-AOAS169. URL http://arxiv.org/abs/0811.1645. arXiv:0811.1645 [stat].

Ishwaran, H., Gerds, T. A., Kogalur, U. B., Moore, R. D., Gange, S. J., and Lau, B. M. Random survival forests for competing risks. *Biostatistics*, 15(4):757–773, 2014.

Kaplan, E. L. and Meier, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, June 1958. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1958.10501452. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501452.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, December 2018. ISSN 1471-2288. doi: 10.1186/s12874-018-0482-1. URL https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0482-1.

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

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

Kretowska, M. Tree-based models for survival data with competing risks. *Computer Methods and Programs in Biomedicine*, 159:185–198, June 2018. ISSN 01692607. doi: 10.1016/j.cmpb.2018.03.017. URL https://linkinghub.elsevier.com/retrieve/pii/S0169260717314347.

Kvamme, H. and Borgan, O. The Brier Score under Administrative Censoring: Problems and Solutions, December 2019a. URL http://arxiv.org/abs/1912.08581. arXiv:1912.08581 [cs, stat].

Kvamme, H. and Borgan, o. Continuous and Discrete-Time Survival Prediction with Neural Networks, October 2019b. URL http://arxiv.org/abs/1910.06724. arXiv:1910.06724 [cs, stat].

Lee, C., Zame, W., Yoon, J., and Van Der Schaar, M. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v32i1.11842. URL https://ojs.aaai.org/index.php/AAAI/article/view/11842.

Maystre, L. and Russo, D. Temporally-Consistent Survival Analysis.

Merkle, E. C. and Steyvers, M. Choosing a Strictly Proper Scoring Rule. *Decision Analysis*, 10(4):292–304, December 2013. ISSN 1545-8490, 1545-8504. doi: 10.1287/deca.2013.0280. URL https://pubsonline.informs.org/doi/10.1287/deca.2013.0280.

Monterrubio-Gómez, K., Constantine-Cooke, N., and Vallejos, C. A. A review on competing risks methods for survival analysis, December 2022. URL http://arxiv.org/abs/2212.05157. arXiv:2212.05157 [stat].

Nagpal, C., Li, X. R., and Dubrawski, A. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks, 2021.

National Cancer Institute, DCCPS, S. R. P. Surveillance, epidemiology, and end results (seer) program (www.seer.cancer.gov) seer*stat database: Incidence - seer research data, 8 registries, nov 2021 sub (1975-2020) - linked to county attributes - time dependent (1990-2020) income/rurality, 1969-2020 counties, national cancer institute, dccps, surveillance research program, , based on the november 2022 submission, 2023.

Nelson, W. Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics*, 14(4):945–966, November 1972. ISSN 0040-1706, 1537-2723. doi: 10.1080/00401706.1972.10488991. URL http://www.tandfonline.com/doi/abs/10.1080/00401706.1972.10488991.

Ovcharov, E. Y. Proper scoring rules and Bregman divergence. *Bernoulli*, 24(1), February 2018. ISSN 1350-7265. doi: 10.3150/16-BEJ857. URL https://projecteuclid.org/journals/bernoulli/volume-24/issue-1/Proper-scoring-rules-and-Bregman-divergence/10.3150/16-BEJ857.full.

Ramspek, C. L., Teece, L., Snell, K. I., Evans, M., Riley, R. D., van Smeden, M., van Geloven, N., and van Diepen, M. 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.

Rindt, D., Hu, R., Steinsaltz, D., and Sejdinovic, D. Survival Regression with Proper Scoring Rules and Monotonic Neural Networks, February 2022. URL http://arxiv.org/abs/2103.14755. arXiv:2103.14755 [cs, stat].

Rith, M., Soliman, J., Fillone, A., Biona, J. B. M., and Lopez, N. S. Analysis of Vehicle Survival Rates for Metro-Manila. In *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM)*, pp. 1–4, Baguio City, Philippines, November 2018. IEEE. ISBN 978-1-5386-7767-4. doi: 10.1109/HNICEM.2018.8666408. URL https://ieeexplore.ieee.org/document/8666408/.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866, September 1994. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1994.10476818. URL https://www.tandfonline.com/doi/full/10.1080/01621459.1994.10476818.

Schoop, R., Beyersmann, J., Schumacher, M., and Binder, H. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal*, 53(1):88–112, February 2011. ISSN 03233847. doi: 10.1002/bimj.201000073. URL https://onlinelibrary.wiley.com/doi/10.1002/bimj.201000073.

Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., and Beghi, A. Machine Learning for Predictive Maintenance: A Multiple Classifier Approach. *IEEE Transactions on Industrial Informatics*, 11 (3):812–820, June 2015. ISSN 1551-3203, 1941-0050. doi: 10.1109/TII.2014.2349359. URL http://ieeexplore.ieee.org/document/6879441/.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117, May 2011. ISSN 0277-6715, 1097-0258. doi: 10.1002/sim.4154. URL https://onlinelibrary.wiley.com/doi/10.1002/sim.4154.

Van Belle, V., Pelckmans, K., Van Huffel, S., and Suykens, J. A. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, 53(2):107–118, October 2011. ISSN 09333657. doi: 10.1016/j.artmed.2011.06.006. URL https://linkinghub.elsevier.com/retrieve/pii/S0933365711000765.

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., Steyerberg, E. W., 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.

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

Wang, Z. and Sun, J. SurvTRACE: Transformers for Survival Analysis with Competing Events. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 1–9, August 2022. doi: 10.1145/3535508.3545521. URL http://arxiv.org/abs/2110.00855. arXiv:2110.00855 [cs, stat].

Wolbers, M., Koller, M. T., Witteman, J. C., and Steyerberg, E. W. Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology*, 20(4):555–561, 2009.

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

Zhang, W., Ling, C. K., and Zhang, X. Deep copula-based survival analysis for dependent censoring with identifiability guarantees, 2023.

Zhu, X., Yao, J., and Huang, J. Deep convolutional neural network for survival analysis with pathological images. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 544–547, Shenzhen, China, December 2016. IEEE. ISBN 978-1-5090-1611-2. doi: 10.1109/BIBM.2016.7822579. URL http://ieeexplore.ieee.org/document/7822579/.

A Definitions

A.1 Notations

Here we detail the notations used in the main manuscript as well as in the proofs and derivations below.

For all symbols, we use the following conventions:

- $.^*$: 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 when the event k will occur
C	$\mathbb{R}_+$	random variable when the censoring will occur
T^*	$\mathbb{R}_+$	$\min(T_1^*, T_2^*, ..., T_K^*)$
T	$\mathbb{R}_+$	$\min(T, C)$
Δ^*	$[1, K]$	$\underset{i\in[1,K]}{\arg\min}(T_i^*)$
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 means censoring

Table S1: Notations used

A.2 Reporting conventions

In tables, the best results are reported in bold characters, and the second best is 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.

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

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

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

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

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

Derivation of the expectation :

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

And:

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

By summing all of the terms, we obtain:

$$\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}} \left[\mathrm{L}_\zeta \left(\hat{F}(\zeta|\mathbf{x}), (T, \Delta) \right) \right] = \sum_{k=1}^{K} \log\left(\hat{F}_k(\zeta|\mathbf{x}), 1\right) \mathbb{P}(T^* \leq \zeta, \Delta = k) + \log\left(\hat{S}(\zeta|\mathbf{X} = \mathbf{x})\right) (\mathbb{P}(T^* > \zeta | \mathbf{X} = \mathbf{x})) \tag{6}$$

$$= \sum_{k=1}^{K} \log\left(\hat{F}_k(\zeta|\mathbf{x})\right) F_k^*(\zeta|\mathrm{x}) + \log\left(\hat{S}(\zeta|\mathbf{x})\right) S^*(\zeta|\mathbf{x}) \tag{7}$$

Finally:

$$\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}} \left[\mathrm{L}_\zeta \left(\hat{F}(\zeta|\mathbf{x}), (T, \Delta) \right) \right] = \sum_{k=1}^{K} \log\left(\hat{F}_k(\zeta|\mathbf{x})\right) F_k^*(\zeta|\mathrm{x}) + \log\left(\hat{S}(\zeta|\mathbf{x})\right) S^*(\zeta|\mathbf{x}) \tag{8}$$

□

Proof of the Theorem 1.

Theorem 1 (Properness of the scoring rule)**.** *Under the assumption that the weights are well 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.

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

And:

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

$$F_k^*(\zeta|\mathbf{x}) = \mathbb{P}(T^* \leq \zeta, \Delta = k|\mathbf{X} = \mathbf{x}) = \mathbb{P}(Y_{k,\zeta} = 1|\mathbf{X} = \mathbf{x}) \tag{9}$$

$$S^*(\zeta|\mathbf{x}) = \mathbb{P}(T^* > \zeta|\mathbf{X} = \mathbf{x}) = \mathbb{P}(Y_{0,\zeta} = 1|\mathbf{X} = \mathbf{x}) \tag{10}$$

$\hat{F}_k(\zeta|\mathbf{x})$ represents the estimated probability for $Y_{k,\zeta} = 1$, so we can rewrite: $\hat{p}_{k,\zeta} \overset{\text{def}}{=} \hat{F}_k(\zeta|\mathbf{x}) \approx \mathbb{P}(Y_{k,\zeta} = 1|\mathbf{X} = \mathbf{x})$ Therefore:

$$\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathrm{L}_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}), (T, \Delta))\right] = \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}[\mathrm{L}_\zeta(\hat{p}_\zeta, (T, \Delta))] \tag{11}$$

$$= \sum_{k=1}^{K} \log\left(\hat{p}_{k,\zeta}\right) \mathbb{P}(Y_{k,\zeta} = 1|\mathbf{X} = \mathbf{x}) + \log\left(\hat{p}_{0,\zeta}\right) \mathbb{P}(Y_{0,\zeta} = 0|\mathbf{X} = \mathbf{x}) \tag{12}$$

Thus:

$$\begin{aligned} \min_{\hat{p}} \quad & \sum_{k=1}^{K} \log\left(\hat{p}_{k,\zeta}\right) \mathbb{P}(Y_{k,\zeta} = 1|\mathbf{X} = \mathbf{x}) + \log\left(\hat{p}_{0,\zeta}\right) \mathbb{P}(Y_{0,\zeta} = 1|\mathbf{X} = \mathbf{x}) \\ \text{s.t.} \quad & \sum_{k=0}^{K} \hat{p}_k = 1 \\ & \hat{p}_k \geq 0 \end{aligned} \tag{13}$$

is obtained for $\hat{p} = p^*$.

$$\begin{aligned} \frac{p_k}{\hat{p}_k} + \mu &= 0 \\ \sum_{k=0}^{K} \hat{p}_k &= 1 \\ \sum_{k=0}^{K} \frac{p_k}{-\mu} &= 1 \\ \mu &= -1 \\ \implies \hat{p}_k &= a_k \end{aligned}$$

□

C Study of the proper scoring rule used for evaluation

As mentioned above, the metric most used in the competing risks setting, the C-index in time, is biased (Blanche et al., 2019; Rindt et al., 2022). To overcome this issue, which is major for any evaluation strategy, we propose here two evaluation metrics: one re-weighting proper scoring rule, that can be effective with any proper binary scoring rule. The second is the accuracy in time that measures the observed event versus the most likely predicted event.

C.1 PSR for evaluation

The PSR introduced in the main paper to be the loss of our model is a global loss over all of our predictions. The following loss is adapted to focus on a special event k to evaluate our estimations on a specific event. In the paper, we chose to focus on the IBS, but one could use a logarithmic loss because of its properness.

Proper scoring rule for the k^{th} competing event In our setting, we will denote $L_{k,\zeta}$, 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:

$$\forall \zeta, (\mathbf{X}, T, \Delta) \sim \mathcal{D}, \quad \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}[L_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}), (T, \Delta))] \leq \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}[L_{k,\zeta}(F_k^*(\zeta|\mathbf{x}), (T, \Delta))] \tag{14}$$

C.1.1 A Proper Scoring Rule for Competing Risks

To evaluate our model, we used the following proper scoring rule is adequate for each event. Thanks to this proper scoring rule, we can understand the error for each event and the global error of all of the CIF.

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, the negative log-likelihood) leads to a (strictly) proper scoring in competing risks settings thanks to the re-weighting of the observations.
Indeed, for any (strictly) proper scoring rule $\ell : \mathbb{R} \times \{0, 1\} \to \mathbb{R}$, one can build a cause-specific scoring rule function $L_{k,\zeta} : \mathbb{R} \times \mathcal{D} \to \mathbb{R}$ that is also a (strictly) proper scoring rule for the cause-specific event k^{th} in 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 $(\mathbf{x}, t, \delta)$ and for an event k, the following scoring rule of the k^{th} CIF:

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

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

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

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

$$+ \frac{\mathbb{1}_{t_i \leq \zeta, \delta_i \neq 0, \delta_i \neq k} \ \ell\left(\hat{F}_k(\zeta|\mathbf{x}_i), 0\right)}{G^*(t_i|\mathbf{x}_i)} \tag{15}$$

Probability of remaining at t_i

The weights correspond to the Inverse Probability of Censoring Weighting (IPCW) 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 former scoring rule can be written as:*

$$\forall \zeta, \forall k \in [\![1, K]\!], \ell : \mathbb{R} \times \{0, 1\} \to \mathbb{R}, (\mathbf{X}, T, \Delta) \sim \mathcal{D},$$
$$\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathrm{L}_{k,\zeta}\left(\hat{F}_k(\zeta|\mathbf{x}), (T, \Delta)\right)\right] = \ell\left(\hat{F}_k(\zeta|\mathbf{x}), 1\right) F_k^*(\zeta|\mathbf{x}) + \ell\left(\hat{F}_k(\zeta|\mathbf{x}), 0\right)(1 - F_k^*(\zeta|\mathbf{x})) \tag{16}$$

Proof.

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

$$\begin{aligned}
\mathrm{L}_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}), (t,\delta)) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} & \underbrace{\frac{\mathbb{1}_{t_i \leq \zeta, \delta_i = k} \ \ell\left(\hat{F}_k(\zeta|\mathbf{x}_i), 1\right)}{G^*(t_i|\mathbf{x}_i)}}_{\stackrel{\text{def}}{=}\Psi_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}),(t,\delta))} \\
& + \underbrace{\frac{\mathbb{1}_{t_i > \zeta} \ \ell\left(\hat{F}_k(\zeta|\mathbf{x}_i), 0\right)}{G^*(\zeta|\mathbf{x}_i)}}_{\stackrel{\text{def}}{=}\Lambda_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}),(t,\delta))} \\
& + \underbrace{\frac{\mathbb{1}_{t_i \leq \zeta, \delta_i \neq 0, \delta_i \neq k} \ \ell\left(\hat{F}_k(\zeta|\mathbf{x}_i), 0\right)}{G^*(t_i|\mathbf{x}_i)}}_{\stackrel{\text{def}}{=}\Phi_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}),(t,\delta))} \qquad (17)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left[\Phi_{k,\zeta}\left(\hat{F}_k(\zeta|\mathbf{x}),(T,\Delta)\right)\right] &= \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathbb{1}_{T\leq\zeta,\Delta\neq0,\Delta\neq k}\frac{\ell\left(\hat{F}_k(\zeta|\mathbf{x}),0\right)}{G^*(T|\mathbf{x})}\right] \\
&= \ell\left(\hat{F}_k(\zeta|\mathbf{x}),0\right) \ \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left[\frac{\mathbb{1}_{T^*\leq\zeta}\mathbb{1}_{T^*\leq C^*}\mathbb{1}_{\Delta\neq k}}{G^*(T|\mathbf{x})}\right] \\
&= \ell\left(\hat{F}_k(\zeta|\mathbf{x}),0\right)\mathbb{P}(T^*\leq\zeta,\Delta\neq k|\mathrm{X}=\mathrm{x})
\end{aligned}$$

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

By summing all of the terms, we obtain:

$$\begin{aligned}
\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathrm{L}_{k,\zeta}\left(\hat{F}_k(\zeta|\mathbf{x}),(T,\Delta)\right)\right] &= \ell\left(\hat{F}_k(\zeta|\mathbf{x}),1\right)\mathbb{P}(T^*\leq\zeta,\Delta=k) \\
&+ \ell\left(\hat{F}_k(\zeta|\mathbf{x}),0\right)\left(\mathbb{P}(T^*\leq\zeta,\Delta\neq k|\mathrm{X}=\mathrm{x})+\mathbb{P}(T^*>\zeta|\mathbf{X}=\mathbf{x})\right)
\end{aligned} \qquad (18)$$

Meanwhile,

$$\mathbb{P}(\overline{T^* \leq \zeta \cap \Delta = k}) = \mathbb{P}(T^* > \zeta \cup \Delta \neq k) \tag{19}$$
$$= \mathbb{P}(T^* > \zeta) + \mathbb{P}(\Delta \neq k) - \mathbb{P}(T^* > \zeta \cap \Delta \neq k) \tag{20}$$
$$= \mathbb{P}(T^* > \zeta) + \mathbb{P}(\Delta \neq k \cap T^* > \zeta) + \mathbb{P}(\Delta \neq k \cap T^* \leq \zeta) - \mathbb{P}(T^* > \zeta \cap \Delta \neq k) \tag{21}$$
$$= \mathbb{P}(T^* > \zeta) + \mathbb{P}(\Delta \neq k \cap T^* \leq \zeta) \tag{22}$$

So, we obtain:

$$\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathrm{L}_{k,\zeta}\left(\hat{F}_k(\zeta|\mathbf{x}),(T,\Delta)\right)\right] = \ell\left(\hat{F}_k(\zeta|\mathbf{x}),1\right)F_k^*(\zeta|\mathbf{x}) + \ell\left(\hat{F}_k(\zeta|\mathbf{x}),0\right)(1-F_k^*(\zeta|\mathbf{x})) \tag{23}$$

□

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

Proof.

$$\begin{aligned}\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathrm{L}_{k,\zeta}\left(\hat{F}_k(\zeta|\mathbf{x}),(T,\Delta)\right)\right] &= \ell\left(\hat{F}_k(\zeta|\mathbf{x}),1\right)\mathbb{P}(T^* \leq \zeta, \Delta = k|\mathbf{X}=\mathbf{x}) \\ &+ \ell\left(\hat{F}_k(\zeta|\mathbf{x}),0\right)(\mathbb{P}(T^* \leq \zeta, \Delta \neq k|\mathbf{X}=\mathbf{x}) + \mathbb{P}(T^* > \zeta|\mathbf{X}=\mathbf{x}))\end{aligned} \tag{24}$$

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

Definition C.3.

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

$$F_k^*(\zeta|\mathbf{x}) = \mathbb{P}(T^* \leq \zeta, \Delta = k|\mathbf{X}=\mathbf{x}) = \mathbb{P}(Y_{k,\zeta} = 1|\mathbf{X}=\mathbf{x}) \tag{25}$$

$\hat{F}_k(\zeta|\mathbf{x})$ represents the estimated probability for $Y_{k,\zeta} = 1$, so we can rewrite: $\hat{p}_{k,\zeta} \stackrel{\text{def}}{=} \hat{F}_k(\zeta|\mathbf{x}) \approx \mathbb{P}(Y_{k,\zeta} = 1|\mathbf{X}=\mathbf{x})$ Therefore:

$$\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left[\mathrm{L}_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}),(T,\Delta))\right] = \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}[\mathrm{L}_{k,\zeta}(\hat{p}_{k,\zeta},(T,\Delta))] \tag{26}$$
$$= \ell\left(\hat{p}_{k,\zeta},0\right)\mathbb{P}(Y_{k,\zeta}=0|\mathbf{X}=\mathbf{x}) + \ell\left(\hat{p}_{k,\zeta},1\right)\mathbb{P}(Y_{k,\zeta}=1|\mathbf{X}=\mathbf{x}) \tag{27}$$
$$= \mathbb{E}_{Y_{k,\zeta}}[\ell(\hat{p}_{k,\zeta},Y_{k,\zeta})|\mathbf{X}=\mathbf{x}] \tag{28}$$
$$\leq \mathbb{E}_{Y_{k,\zeta}}[\ell(p_{k,\zeta},Y_{k,\zeta})|\mathbf{X}=\mathbf{x}] \tag{29}$$
$$\leq \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}[\mathrm{L}_{k,\zeta}(\mathbb{P}(Y_{k,\zeta}=1),(T,\Delta))] \tag{30}$$
$$\leq \mathbb{E}[\mathrm{L}_{k,\zeta}(F_k^*(\zeta|\mathrm{x}),(T,\Delta))] \tag{31}$$

The last inequality is valid because l is a proper scoring rule. The same computation leads to a strictly proper scoring rule if l is a strictly proper scoring rule.

So, we obtain 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, thus $L_\zeta : \mathbb{R}\times\mathcal{D} \to \mathbb{R}$ is a (strictly) proper scoring rule for the global CIF at a fixed time horizon $\zeta \in \mathbb{R}_+$.*

Proof. Straight forward thanks to 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 the time horizon ζ chosen arbitrarily. To be able to compare different models, a global measure is necessary, *eg* by summing over time, as introduced in Graf et al. (1999). Here, we extend the Integrated Brier Score to other (strictly) proper scoring rules l and we 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} 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

$$\mathrm{IL}(\hat{F}_1(.|\mathbf{x}), ..., \hat{F}_K(.|\mathbf{x}), (T, \Delta)) \stackrel{\text{def}}{=} \mathbb{E}_Z \left[\sum_{k=1}^{K} \mathrm{L}_{k,Z}(\hat{F}_k(Z|\mathbf{x}), (T, \Delta)) | \mathbf{X} = \mathbf{x} \right] \tag{32}$$

$$= \sum_{k=1}^{K} \underbrace{\mathbb{E}_Z \left[\mathrm{L}_{k,Z}(\hat{F}_k(Z|\mathbf{x}), (T, \Delta)) | \mathbf{X} = \mathbf{x} \right]}_{\stackrel{\text{def}}{=} \mathrm{IL}_k(\hat{F}_k(.|\mathbf{x}), (T, \Delta))} \tag{33}$$

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. Using the monotonicity /positivity of the expectation, the result is immediate.

$$\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x},Z=\zeta} \left[\mathrm{IL}_k(\hat{F}_k(\zeta|\mathbf{x})), (T, \Delta) \right] = \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x},Z=\zeta} \left[\mathrm{L}_k(\hat{F}_k(\zeta|\mathbf{x}), (T, \Delta)) \right] \tag{34}$$

$$\leq \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x},Z=\zeta} \left[\mathrm{L}_k(F_k^*(\zeta|\mathbf{x}), (T, \Delta)) \right] \tag{35}$$

$$\leq \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x},Z=\zeta} \left[\mathrm{IL}_k(F_k^*(\zeta|\mathbf{x}), (T, \Delta)) \right] \tag{36}$$

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

$$\mathbb{E}_{(\Delta,T)} \left[\mathrm{IL}_k(\hat{F}_k(Z|\mathbf{x}), (T, \Delta)) | \mathbf{X} = \mathbf{x} \right] \leq \mathbb{E}_{(\Delta,T)} \left[\mathrm{IL}_k(F_k^*(Z|\mathbf{x}), (T, \Delta)) | \mathbf{X} = \mathbf{x} \right] \tag{37}$$

This allows us to consider the IL as a global proper scoring rule to compare different competing risks models. □

D The Yanagisawa (2023) scoring rule for survival

Yanagisawa (2023) introduce a metric, called $S_{Cen-log-simple}$, is an approximation of the proper scoring metric in Rindt et al. (2022). Indeed, the metric in Rindt et al. (2022) requires the hazard function, the time derivative of the cumulative incidence function, which is exposed only by differentiable models –and hence with an implicit assumption on almost-everywhere smooth time dependence. To avoid requesting this hazard function, Yanagisawa (2023) approximate it as piecewise affine. They show that under the assumption that the "node time points", edges of the affine, parts match an actual piecewise-affine breakdown of the CIF, the resulting approximation is proper. They argue that with enough node time points, the metric is a good approximation of a proper scoring rule.

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

$$S_{Cen-log-simple}(\hat{F}, (t, \delta); \{\zeta_i\}_{i=0}^{B}) \stackrel{\text{def}}{=}$$

$$- \delta \sum_{i=0}^{B-1} \mathbb{1}_{\zeta_i < t \leq \zeta_{i+1}} \log(\hat{F}(\zeta_{i+1}) - \hat{F}(\zeta_i))$$

$$- (1 - \delta) \sum_{i=0}^{B-1} \mathbb{1}_{\zeta_i < t \leq \zeta_{i+1}} \log(1 - \hat{F}(\zeta_{i+1})) \tag{38}$$

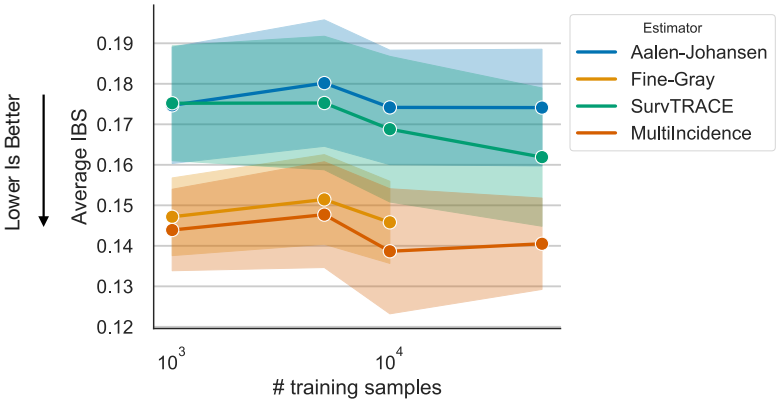


Figure S1: **Integrated Brier Score (IBS) vs Training Samples on Synthetic Dataset** Integrated Brier Score for the synthetic dataset with linear relation over the features when we vary the number of samples. The test set was made into five different seeds.

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

E Additional results for competing risk experiments

E.1 Results on synthetic dataset

Varying the number of training points shows a slow improvement of SurvTrace, but at $n = 5 \cdot 10^4$ MultiIncidence still has the best IBS (Figure S1). MultiIncidence also maintains its benefit with an increased censoring rate (Figure S3). In terms of computation time, MultiIncidence is the fastest, but the dependence on the number of features is similar across MultiIncidence, Fine & Gray, and SurvTRACE (Figure S2).

Integrated Brier Score with a varying number of points By varying the number of training points in our synthetic dataset, while the Oracle Integrated Brier Score is decreasing, we see in Figure S1 that our method obtains better results than the transformer (SurvTRACE) in particular for a smaller number of training points. The number of training points may be a huge bottleneck for medical studies, as the number of patients may be low. We also see that, as the number of training points increases, SurvTRACE improves. With too many points, here 20,000, the Fine & Gray model was too long to run. We also see that the Fine & Gray model achieves approximately the same performance as our model, as expected because we model linear relations between the targets and the features.

Computational cost vs performances To emphasize this phenomenon, we measured the time to fit each model, while varying the number of samples and the number of features in Figure S5. We show that for a limited number of samples, all of the methods take approximately the same amount of time to fit while having the worst results for SurvTRACE. With a higher number of samples, our method was faster to train than the other ones while achieving the same performance. We did not obtain the results for the Fine & Gray model because the time to fit was higher than the given budget.

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

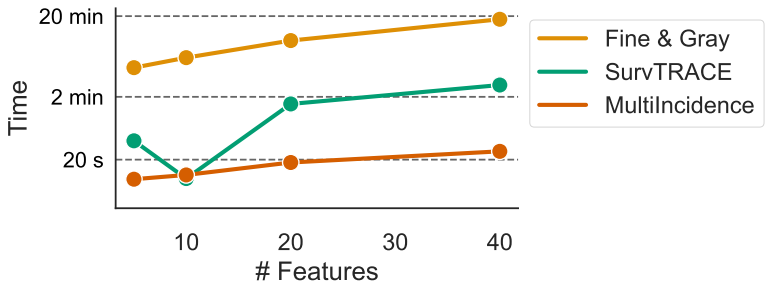


Figure S2: **Fitting time vs number of features** Time to fit 10,000 samples depending on the number of features.

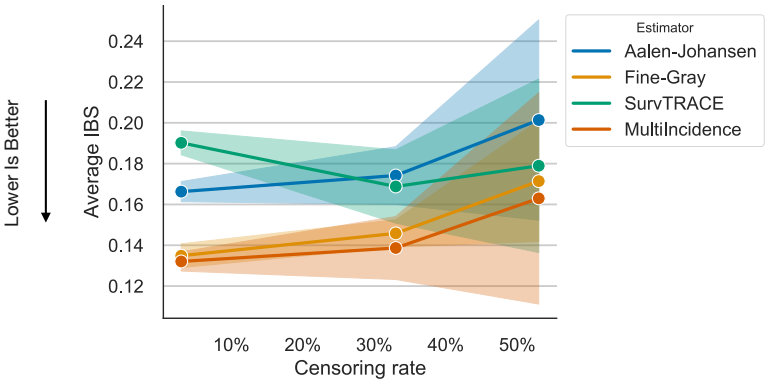


Figure S3: **Integrated Brier Score vs Censoring Rate** Integrated Brier Score for the synthetic dataset with 10,000 training points when we vary the censoring rate. Shaded areas represent the standard deviation across the different seeds. We used the Oracle censoring distribution to compute the weights

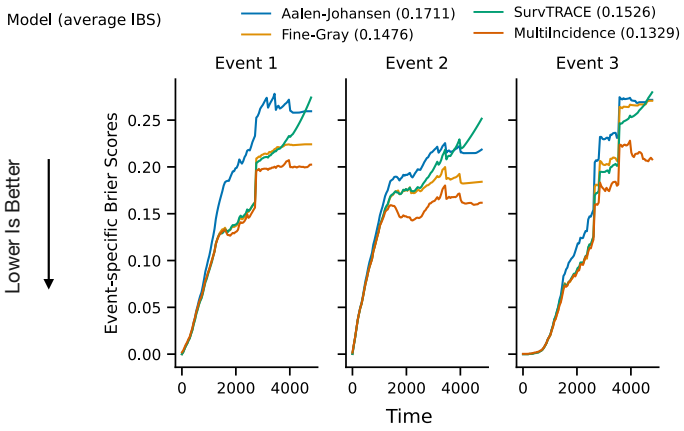


Figure S4: **Brier Score in time** Evolution of the Brier Score for the synthetic dataset H.2 with 20,000 training points with 50% of censoring. The weights are computed with the Oracle censoring distribution.

We show the dependence of time to fit with the number of features in Figure S2. In this figure, we highlight that our method takes less time to fit; the increase in time to fit with the number of features is similar among all methods. Another study of the impact of the features and the number of samples to fit the models can be found in Appendix S8.

Censoring Scale We studied the impact of censoring on the different models. To do so, we vary the censoring distribution to understand the effect of the learning scheme. In Figure S3, we see that our method outperforms SurvTRACE at different censoring rates. As expected, all models get worse as the censoring rate increases.

Brier Score in time We compared the Brier Score over time for each model, as shown in Figure S4. The Brier Score increases over time for all models, which is expected due to the smaller number of individuals toward the end. Additionally, the associated weights contribute significantly to errors at later times. In this context, MultiIncidence consistently outperforms every other model for each event.

Impact of the number of features and the training samples on fit time of competing risks

E.2 Results for the SEER Dataset

Learning curves We ran the experiments while varying the number of training points. In doing so, we measured the KM-adjusted Integrated Brier Score for each event. We also average it to have one global metric. We see in Figure S7 that our model of the global evaluation metric is quite stable and lower than the average Integrated Brier Score on SurvTRACE for any number of training points. We expanded the Integrated Brier Score for each event while training on the whole dataset except for the Random Survival Forests we trained with 100k data points and Fine and Gray with 10k data points because the last two methods could not handle such an amount of data. In Table S2, we compare our

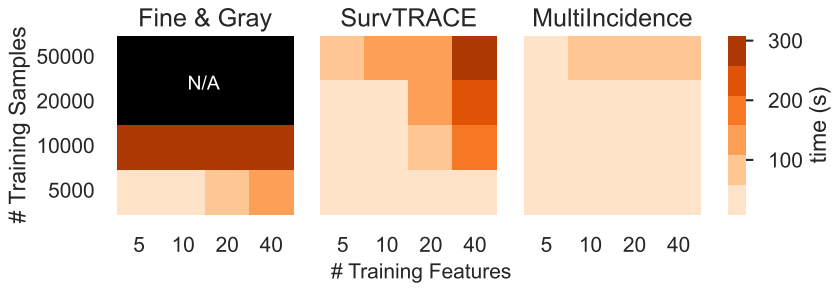


Figure S5: **Fit time for competing risks models**. We have measured the time to fit for each of them depending on the number of training points and the number of features.

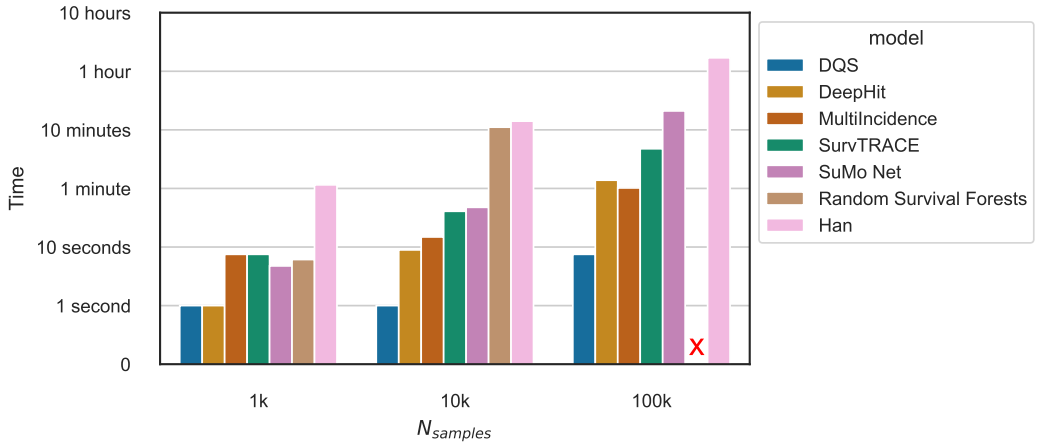


Figure S6: **Synthetic Dataset, training time for survival** Time to fit each survival method while varying the number of samples generated.

method with the other models. We see that our model MultiIncidence outperforms the other methods. Furthermore, figure 3 shows 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 of the different samples is in agreement with the order of the times in which the event of interest happens(Harrell et al., 1982). It is originally a metric for survival settings but is often adapted to competing risks settings where it is applied independently to each event (Uno et al., 2011). In such settings, it is biased and does not

Figure S7: **Integrated Brier Score vs Number of Training Samples: SEER** Integrated Brier Score (Lower is Better) on the SEER dataset varying the number of samples: 50,000 samples, 100,000, and the full Training Dataset, aside for the Fine&Gray model, which was tractable only for 10,000 samples.

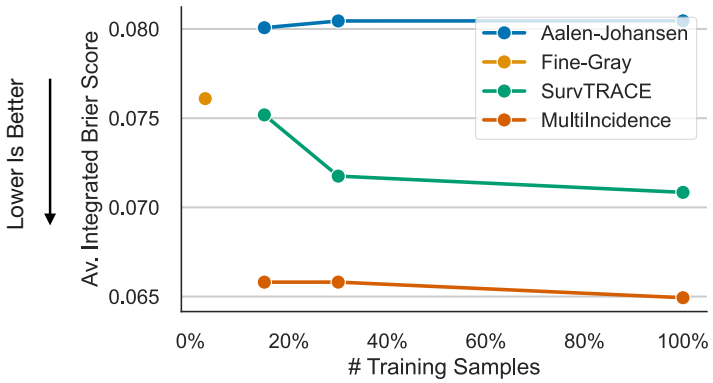


Table S2: Integrated Brier Score for each cause-specific risk on the SEER Dataset (Lower is Better).

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
MULTIINCIDENCE	0.0832	**0.0273**	**0.0757**

control for the probabilities of the events. However, as it is a popular metric, we have included it in our experiments.

We give tables below for the C_ζ-index toward time for the three events S3. At a fixed time horizon ζ, we compute the C_ζ-index for each class (corresponding to the ROC-AUC where we handled censored observations). The time horizons ζ are selected based on the any-event distribution, representing quantiles, indicating that at the time corresponding to 0.25, 25% of events have already occurred. These results differ from those in the SurvTRACE paper (Wang & Sun, 2022) for two reasons: *1)* The available code online only implements one of their losses, *2)* they treated the SEER dataset with two competing risks, and any other event was classified as censored, instead of collapsing them in a third competing event.

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	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
FINE & GRAY	0.80	0.67	0.67	0.77	0.67	0.69	0.76	0.68	0.71
RANDOM SURVIVAL FORESTS	**0.89**	0.79	0.79	**0.87**	**0.78**	**0.77**	**0.85**	**0.77**	**0.77**
DEEPHIT	0.83	**0.86**	**0.85**	0.75	0.75	0.75	0.73	0.75	0.75
DSM	0.88	0.85	0.84	0.77	0.74	0.75	0.76	0.75	0.75
DESURV	0.83	0.82	0.81	0.72	0.70	0.71	0.74	0.73	0.73
SURVTRACE	0.88	0.78	0.77	0.86	0.76	0.75	0.84	0.76	0.75
MULTIINCIDENCE	0.88	0.79	0.77	0.85	0.72	0.71	0.81	0.66	0.62

F Additional results for survival experiments

F.1 Metrics for the survival analysis

Table S4: **METABRIC**: Integrated Brier Score, $S_{Cen-log-simple}$ and c-index at 50%

MODEL	C-INDEX 0.25	C-INDEX. 0.5	C-INDEX 0.75	IBS	$S_{Cen-log-simple}$
RANDOM SURVIVAL FORESTS	0.502±0.009	0.483±0.027	0.502±0.021	0.197±0.025	2.442±0.044
DEEPHIT	0.525±0.041	**0.639±0.024**	0.613±0.016	0.180±0.014	2.271±0.019
PCHAZARD	0.595±0.088	**0.639±0.019**	**0.639±0.014**	0.176±0.014	2.246±0.046
HAN	0.626±0.035	0.622±0.007	0.628±0.006	0.191±0.003	2.420±0.150
DQS	0.601±0.019	0.630±0.032	0.633±0.014	0.180±0.034	2.205±0.044
SUMO NET	**0.660±0.022**	0.634±0.017	0.589±0.015	0.169±0.009	2.302±0.059
SURVTRACE	0.589±0.082	0.627±0.015	0.629±0.007	**0.168±0.011**	2.270±0.034
MULTIINCIDENCE	0.627±0.016	0.636±0.015	0.635±0.011	**0.168±0.019**	**2.169±0.056**

Table S5: **SUPPORT**: Integrated Brier Score and $S_{Cen-log-simple}$ (Lower is Better)

MODEL	C-INDEX 0.25	C-INDEX 0.50	C-INDEX 0.75	IBS	$S_{Cen-log-simple}$
RANDOM SURVIVAL FORESTS	0.481±0.024	0.527±0.019	0.531±0.020	0.225±0.004	1.942±0.023
DEEPHIT	0.449±0.041	0.609±0.004	0.599±0.003	0.217±0.005	2.251±0.021
PCHAZARD	0.585±0.014	0.584±0.014	0.584±0.016	0.210±0.007	2.192±0.024
HAN	0.576±0.016	0.574±0.007	0.587±0.011	0.260±0.012	3.483±0.307
DQS	**0.601±0.019**	0.598±0.012	0.592±0.009	0.201±0.007	1.987±0.069
SUMO NET	0.590±0.016	0.589±0.016	0.589±0.015	0.194±0.010	**1.721±0.016**
SURVTRACE	0.578±0.008	0.609±0.005	0.610±0.006	0.194±0.005	1.870±0.018
MULTIINCIDENCE	0.572±0.019	**0.618±0.007**	**0.615±0.007**	**0.191±0.006**	1.740±0.020

F.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.

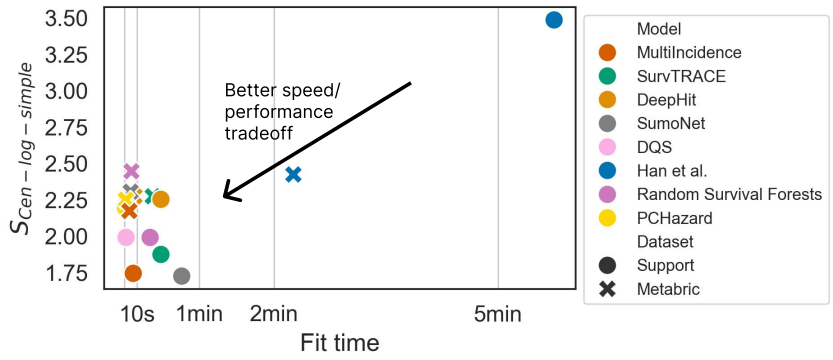


Figure S8: Trade-off between the performances and the training time for the $S_{Cen-log-simple}$ for the survival model over METABRIC and SUPPORT

G Implementation Details

G.1 Reference of used implementations for baselines

We compare MultiIncidence with several baselines and describe their main characteristics and the implementation used in Table S6

G.2 GridSearch Parameters

We ran a Randomized Search for those parameters with a budget of 30. There are no parameters to tune for Aalen-Johansen and Fine & Gray.

H Distribution of the competing risks datasets

H.1 SEER Distribution of events

Here, we present the distributions of the event of the SEER Dataset. We can highlight that the censoring distribution is non-uniform in time. The change in the censoring distribution from the 48^{th} month may be difficult to learn for some methods.

Table S6: Characteristics of used baselines.

Name	Competing risks	Proper loss	Implementation	Reference
SurvTRACE	✓		ours	Wang & 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 & Yau (2022a)
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 & 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 & Borgan (2019b)
SumoNet		✓	github.com/MrHuff/Sumo-Net	Rindt et al. (2022)
DQS		✓	ibm.github.io/dqs/	Yanagisawa (2023)

Table S7: Randomized Search Parameters

Estimator	Parameter	Range
MultiIncidence	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	$\{2, 3\}$

H.2 Example of distribution of one synthetic dataset

Figure S10 shows an example of the distribution of the events with the censoring (dependent on the covariates). The parameters are chosen to fit three different behaviors possible. To illustrate this distribution, we can think of truck maintenance. Event 1, happening during the whole period duration, corresponds to the driver's driving skills. Event 2 may correspond to a misconception of the truck, happening from the beginning. Event 3 will refer to the truck's wear and tear.

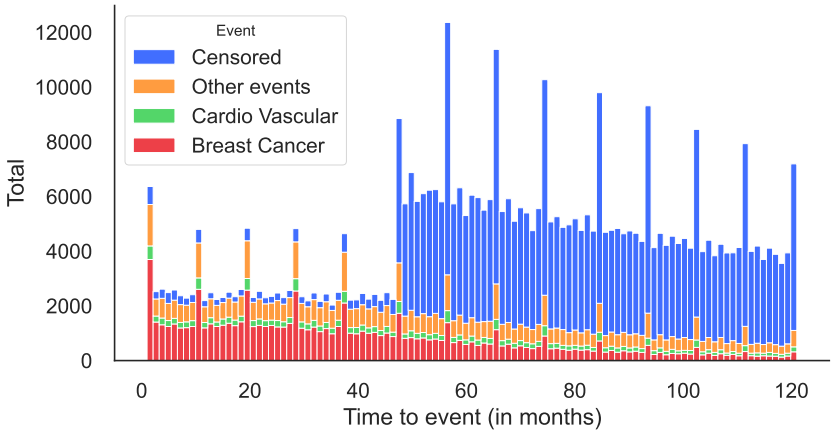


Figure S9: **SEER Dataset Distributions** The censoring rate is around 63%. The prevalence of the different events is 18% for Breast Cancer, 4.5% for Cardio Vascular events, and 10% for other events.

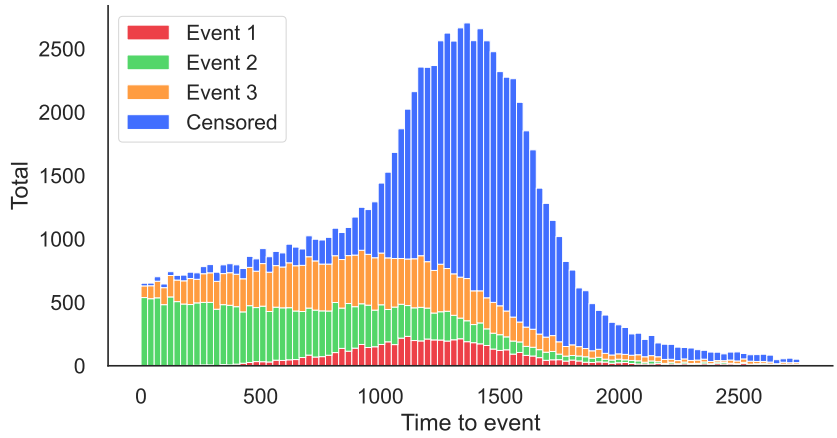


Figure S10: **Synthetic Dataset Distributions** Duration distributions of the synthetic dataset when censoring is dependent on X, censoring rate 69%. The events are stacked.

I Corollary: Bregman divergence

Here, we propose another proof with a scoring rule in the form of a Bregman Divergence. A Bregman divergence is a form of distance, and because of that, we want to minimize the Bregman divergence.

Definition I.1. Considering $U : \mathbb{R}^d \to \mathbb{R}$ strictly convex and differentiable,

Bregman divergence
$$D_U(p, q) = U(p) - U(q) - \langle \nabla U(q), p - q \rangle. \geq 0 \tag{39}$$

The specific choice of l as D_U does not change any computations of the expectation, so we obtain:

$$\begin{aligned}
\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left(\mathrm{L}_{k,\zeta}\left(\hat{F}_k(\zeta|\mathbf{x}), (T, \Delta)\right)\right) &= D_U\left(0, \hat{F}_k(\zeta|\mathbf{x})\right)(1 - F_k^*(\zeta|\mathbf{x})) + D_U\left(1, \hat{F}_k(\zeta|\mathbf{x})\right) F_k^*(\zeta|\mathbf{x}) \\
&= (U(0) - U(\hat{F}_k(\zeta|\mathbf{x})) + \langle \nabla U(\hat{F}_k(\zeta|\mathbf{x})), \hat{F}_k(\zeta|\mathbf{x})\rangle)(1 - F_k^*(\zeta|\mathbf{x})) \\
&\quad + (U(1) - U(\hat{F}_k(\zeta|\mathbf{x})) - \langle \nabla U(\hat{F}_k(\zeta|\mathbf{x})), 1 - \hat{F}_k(\zeta|\mathbf{x})\rangle) F_k^*(\zeta|\mathbf{x}) \\
&= U(1)F_k^*(\zeta|\mathbf{x}) + U(0)(1 - F_k^*(\zeta|\mathbf{x})) - U(\hat{F}_k(\zeta|\mathbf{x})) \\
&\quad + \langle \nabla U(\hat{F}_k(\zeta|\mathbf{x})), \hat{F}_k(\zeta|\mathbf{x}) - F_k^*(\zeta|\mathbf{x})\rangle
\end{aligned}$$

Meanwhile, because U is strictly convex and differentiable:

$$\forall p, \hat{p}, \quad U(p) > U(\hat{p}) + \langle \nabla U(\hat{p}), p - \hat{p} \rangle \tag{40}$$

$$-U(\hat{p}) + \langle \nabla U(\hat{p}), \hat{p} - p \rangle > -U(p) \tag{41}$$

This implies:

$$\begin{aligned}\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left(\mathrm{L}_{k,\zeta}\left(\hat{F}_k(\zeta|\mathbf{x}),(T,\Delta)\right)\right) &= D_U\left(0,\hat{F}_k(\zeta|\mathbf{x})\right)\left(1-F_k^*(\zeta|\mathbf{x})\right)+D_U\left(1,\hat{F}_k(\zeta|\mathbf{x})\right)F_k^*(\zeta|\mathbf{x}) \\ &> U(1)F_k^*(\zeta|\mathbf{x})+U(0)(1-F_k^*(\zeta|\mathbf{x}))-U(F_k^*(\zeta|\mathbf{x})) \\ &> D_U\left(0,F_k^*(\zeta|\mathbf{x})\right)\left(1-F_k^*(\zeta|\mathbf{x})\right)+D_U\left(1,F_k^*(\zeta|\mathbf{x})\right)F_k^*(\zeta|\mathbf{x}) \\ &> \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}}\left(\mathrm{L}_{k,\zeta}\left(F_k^*(\zeta|\mathbf{x}),(T,\Delta)\right)\right)\end{aligned}$$

We obtain that, a negative Bregman Divergence leads to a strictly proper scoring rule.

J Examples

J.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 define in Eq. 14 of Kretowska (2018):

Definition J.1.

$$\begin{aligned}&\forall\zeta,\forall k\in[1,K], \\ &\mathrm{BS}_k(\hat{F}_k(\zeta,\mathbf{x}),\delta,t,\zeta,\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}_{t_i\leq\zeta,\delta_i=k}\left(1-\hat{F}_k(\zeta|\mathbf{x}_i)\right)^2}{G^*(t_i|\mathbf{x}_i)}+\frac{\mathbb{1}_{t_i>\zeta}\left(\hat{F}_k(\zeta|\mathbf{x}_i)\right)^2}{G^*(\zeta|\mathbf{x}_i)} \\ &\qquad\qquad +\frac{\mathbb{1}_{t_i\leq\zeta,\delta_i\neq0,\delta_i\neq k}\left(\hat{F}_k(\zeta|\mathbf{x}_i)\right)^2}{G^*(t_i|\mathbf{x}_i)} \qquad (42)\end{aligned}$$

J.2 Binary cross entropy loss

As it is explained in Benedetti (2010), the log loss captures better the uncertainty than the mean squared error. So, one could also evaluate survival and competing risks models with the following loss.

$$\begin{aligned}&\forall k\in[1,K], \\ &\mathrm{l}_k(\hat{F}_k(\zeta,\mathbf{x}),\delta,t,\zeta) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}_{t_i\leq\zeta,\delta_i=k}\log\left(\hat{F}_k(\zeta|\mathbf{x}_i)\right)}{G^*(t_i|\mathbf{x}_i)}+\frac{\mathbb{1}_{t_i\leq\zeta,\delta_i\neq0,\delta_i\neq k}\log\left(1-\hat{F}_k(\zeta|\mathbf{x}_i)\right)}{G^*(t_i|\mathbf{x}_i)} \\ &\qquad\qquad +\frac{\mathbb{1}_{t_i>\zeta}\log\left(1-\hat{F}_k(\zeta|\mathbf{x}_i)\right)}{G^*(\zeta|\mathbf{x}_i)} \qquad (43)\end{aligned}$$