



**HAL**  
open science

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

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

► **To cite this version:**

Julie Alberge, Vincent Maladière, Olivier Grisel, Judith Abécassis, Gaël Varoquaux. Teaching Models To Survive: Proper Scoring Rule and Stochastic Optimization with Competing Risks. 2024. hal-04617672v2

**HAL Id: hal-04617672**

**<https://hal.science/hal-04617672v2>**

Preprint submitted on 12 Sep 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

---

# 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**  
:proabl.  
Paris, France  
vincent@proabl.ai

**Olivier Grisel**  
:proabl.  
Paris, France

**Judith Abécassis**  
SODA Team, Inria Saclay  
Palaiseau, France

**Gaël Varoquaux**  
SODA Team, Inria Saclay  
Palaiseau, France

## Abstract

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

## 12 **1 Introduction**

13       We all die at some point; some applications call for predicting not whether an event of interest will  
14       happen or not, but when it is likely to occur: *time-to-event regression*. In such a setting, samples often  
15       have unobserved outcomes, *e.g.* individuals that have not been followed long enough for the event  
16       of interest to occur. Limiting the analysis to fully observed samples creates a censoring bias; valid  
17       models use dedicated corrections for censorship: *survival analysis* models. These have long been  
18       central to health (Zhu et al., 2016; Chaddad et al., 2016; Gaynor et al., 1993). Nowadays, survival  
19       analysis is also used in diverse fields, such as predictive maintenance (Rith et al., 2018; Susto et al.,  
20       2015), or user-engagement studies (Maystre & Russo). Survival analysis has led to many dedicated  
21       models, such as the Kaplan & Meier (1958) estimator or the Cox (1972) proportional hazard model.

22       Competing risks analysis generalizes survival analysis to account for multiple events, determining  
23       which will happen first (Susto et al., 2015; Gaynor et al., 1993). For instance, if a person with  
24       breast cancer dies from a different cause, it is impossible to determine when they would have  
25       succumbed to cancer, regardless of the length of the observation period. (National Cancer Institute,  
26       2023). The caregiver may also want to adapt the treatment if the patient is predicted to die of a  
27       competing event such as a heart attack sooner than from cancer. As the risks of the various events  
28       are seldom independent—for instance, cancer and cardiovascular disease share inflammation or age  
29       risk factors (Koene et al., 2016)—competing risks cannot be solved by running a survival model for  
30       each event (Wolbers et al., 2009). The estimated risk of a single event of interest will be biased if

31 the competing risks are not included. Hence adequate models for those risks are critical for decision  
32 making (Ramspek et al., 2022; Koller et al., 2012; van Walraven & McAlister, 2016).

33 Survival models have traditionally been developed with ad hoc adjustments for censoring. The most  
34 common approach is to design a likelihood using the probability of censoring per unit time—*i.e.* the  
35 time-derivative of the risk—which either comes with strong parametric assumptions (Cox, 1972) or ad  
36 hoc corrections (Wang & Sun, 2022). Given that the risk, which is the probability of the outcome at a  
37 specific time, is crucial for various applications, it can be preferable to use losses that directly control  
38 probabilities (proper scoring rules), as developed by Graf et al. (1999); Rindt et al. (2022). However,  
39 no metric (or loss) has been shown to control probabilities in the competing risks setting.

40 In application domains typical of survival analysis and competing risks—health, predictive mainte-  
41 nance, insurance, marketing—the data are typically tabular with categorical variables, where tree-based  
42 models shine (Grinsztajn et al., 2022). Existing survival and competing risks models do not fit well  
43 with these requirements. In particular, the proper scoring rule introduced by Rindt et al. (2022)  
44 requires a time derivative of the risk, typically via an auto-diff operator in a neural architecture. This  
45 approach is challenging to adapt to tree-based algorithms. In addition, the ever-growing volume of  
46 data calls for computationally efficient algorithms.

47 **Contributions** Here, we provide a general theoretical framework to learn a competing risks model  
48 with a proper scoring rule. This scoring rule gives a loss easy to plug into any multiclass estimator to  
49 create a competing risks model: giving the individual risk of each event at any horizon. We also sum  
50 over time for model evaluation, as the resulting Integrated Scoring Rule is also proper.

51 An interesting property of this new loss is that it can be optimized on a subset of the training  
52 data because the evaluation is made independently of observations. Hence, it allows stochastic  
53 optimization, enabling computationally efficient learning. With that, we propose an algorithm called  
54 MultiIncidence, based on Stochastic Gradient Boosting Trees. We benchmark our algorithm on a  
55 synthetic dataset with varying censoring rates, number of features, and number of training samples  
56 to show that our method outperforms state-of-the-art methods while exhibiting faster training times.  
57 Finally, applying our model to real-life datasets demonstrates that it outperforms other models in both  
58 the competing risks context and basic survival analysis.

## 59 2 Related work

60 **Survival settings** Various survival models have been developed, ranging from approaches like the  
61 Kaplan & Meier (1958) estimator, estimating the general survival curve of a whole population, to  
62 models that account for covariates. One of them is the Cox (1972) Proportional-Hazards Model,  
63 a linear model of *hazard*: the instantaneous probability of an event, *i.e.* the logarithmic derivative  
64 of outcome probabilities in time. More complex models have been adapted to the survival setting:  
65 Support Vector Machines (Van Belle et al., 2011), survival games (Han et al., 2021) and Neural  
66 networks with DeepSurv (Katzman et al., 2018) or PCHazard (Kvamme & Borgan, 2019b). While  
67 the above do not control risks, more recent neural networks use adequate losses (see below): DQS  
68 (Yanagisawa, 2023, though relying on a piecewise constant hazard), SumoNet (Rindt et al., 2022,  
69 which requires differentiable models).

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

82 **Evaluation for such models** Prediction evaluation in survival or competing risks settings calls for  
 83 adapted metrics to account for right-censored points (Harrell et al., 1982), like the C-index which  
 84 adapts the Area Under the ROC curve in classification. However, the C-index only evaluates the  
 85 ranking of samples, *i.e.* which samples will undergo the event of interest first, and is dependent on  
 86 the censoring distribution which may bias the evaluation (Blanche et al., 2019; Rindt et al., 2022). In  
 87 fact, the score may be higher for distributions other than oracle-censoring distributions. Alternative  
 88 methods have been proposed such as the *time-dependent* C-index,  $C_\zeta$  (Antolini et al., 2005), which  
 89 is the same metric but computed at a given time horizon  $\zeta$ . The C-index ranking metric has been  
 90 extended to competing risks (Uno et al., 2011) but, as in the survival setting, the C-index only  
 91 evaluates relative risks for pairs of individuals and not the absolute value of the risk for a given  
 92 individual. Other time-dependent adaptations of the ROC curve have been developed, also assessing  
 93 a discriminative power rather than risks or probabilities (Blanche et al., 2013). And yet control of the  
 94 risk is crucial to decision making (Van Calster et al., 2019). Proper scoring rules are alternatives to  
 95 overcome the limitations of existing metrics because they capture more aspects of the problem. In  
 96 addition, they can be used for both the training and evaluation of probabilistic predictive models.

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

102 Graf et al. (1999) adapt the Brier score to survival analysis, with a strong independence assumption  
 103 on the censoring distribution. However, the assumption can easily be violated (Kvamme & Borgan,  
 104 2019a) which leads to bias (Rindt et al., 2022). Rindt et al. (2022) show that the likelihood of  
 105 the survival function leads to a proper scoring rule but requires obtaining the density function and  
 106 the survival function, a time-wise derivative of outcome probabilities (definition 3.1). For quantile  
 107 regression, Yanagisawa (2023) shows that the Pinball loss may lead to a proper scoring rule for  
 108 survival analysis but requires an oracle parameter. Han et al. (2021) introduces a double optimization  
 109 problem for which the stationary point is located at the true distributions.

110 For competing risks, Schoop et al. (2011) extend the Brier score to a proper scoring rule. However,  
 111 the Brier score does not measure the uncertainty as well as the log loss (Benedetti, 2010).

### 112 3 Problem Formulation

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

#### 115 3.1 Problem setting

116 We consider  $K$  competing events and for  $k \in \llbracket 1, K \rrbracket$ , we denote  $T_k^* \in \mathbb{R}_+$  the event time of the  
 117 event  $k$ , depending on the covariate  $\mathbf{X} \in \mathcal{X}$ . We also denote  $T^* \in \mathbb{R}_+$ ,  $T^* = \min_{k \in \llbracket 1, K \rrbracket} (T_k^*)$  and

118  $\Delta^* \in \llbracket 1, K \rrbracket$ ,  $\Delta^* = \arg \min_{k \in \llbracket 1, K \rrbracket} (T_k^*)$ . We observe  $(\mathbf{X}, T, \Delta) \sim \mathcal{D}$ , with  $T = \min(T^*, C)$  where

119  $C \in \mathbb{R}_+$  is the censoring time, which may depend on  $\mathbf{X}$  and  $\Delta \in \llbracket 0, K \rrbracket$ ,  $\Delta = \arg \min_{k \in \llbracket 0, K \rrbracket} (T_k^*)$  where

120 0 denotes a censored observation. However, we are interested in the distribution of the uncensored  
 121 data  $(\mathbf{X}, T^*, \Delta^*) \sim \mathcal{D}^*$  especially in the joint distribution of  $T^*, \Delta^* | \mathbf{X} = \mathbf{x}$  and the marginal  
 122 distribution of  $T^* | \mathbf{X} = \mathbf{x}$ .

123 This paper aims to predict an unbiased estimate of all of the cause-specific Cumulative Incidence  
 124 functions (CIF) at any time horizon  $\zeta$  chosen based on the observations  $(\mathbf{x}, t, \delta)$ :

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

$$\begin{aligned} \text{Survival to any event:} & S^*(\zeta | \mathbf{x}) = \mathbb{P}(T^* > \zeta | \mathbf{X} = \mathbf{x}) \\ \text{CIF (cumulative incidence function) of any event:} & F^*(\zeta | \mathbf{x}) = \mathbb{P}(T^* \leq \zeta | \mathbf{X} = \mathbf{x}) = 1 - S^*(\zeta | \mathbf{X} = \mathbf{x}) \\ \text{CIF of the } k^{\text{th}} \text{ event:} & F_k^*(\zeta | \mathbf{x}) = \mathbb{P}(T^* \leq \zeta \cap \Delta^* = k | \mathbf{X} = \mathbf{x}) \\ \text{Censoring:} & G^*(\zeta | \mathbf{x}) = \mathbb{P}(C > \zeta | \mathbf{X} = \mathbf{x}) \end{aligned}$$

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

$$\forall k, \in \llbracket 1, K \rrbracket, T_k^* \perp\!\!\!\perp C | \mathbf{X}$$

125 Assumption 3.1 needed for most theoretical results in survival (Rindt et al., 2022; Yanagisawa,  
126 2023; Han et al., 2021). It is key to understanding why single-event survival analysis is invalid  
127 in the presence of competing risks: if some observations are censored due to other events sharing  
128 unobserved risk factors with the event of interest, this assumption is violated.

### 129 3.2 CIF scoring rule

130 **Proper Scoring Rule** A scoring rule  $\ell$  evaluates a distribution  $\mathcal{P}$  on an observation  $Y$  and gives a  
131 corresponding score  $\ell(\mathcal{P}, Y)$ . The higher the score, the better the model fits the observation. For a  
132 proper scoring rule, it corresponds to the degree to which the model can predict the oracle distribution  
133 (more on scoring rules in Gneiting & Raftery, 2007; Ovcharov, 2018; Merkle & Steyvers, 2013).

**Definition 3.2** (*Proper Scoring Rule*). A scoring rule  $\ell$  is proper if

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

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

135 **Proper scoring rule for the Global CIF** We will denote  $L_\zeta$ , a scoring rule for the global CIF at a  
136 time horizon  $\zeta$ .

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

$$\begin{aligned} \forall \zeta, (\mathbf{X}, T, \Delta) \sim \mathcal{D}, \\ \mathbb{E}_{T^*, C, \Delta | \mathbf{X}=\mathbf{x}}[L_\zeta((\hat{F}_1(\zeta|\mathbf{x}), \dots, \hat{F}_K(\zeta|\mathbf{x}), \hat{S}(\zeta|\mathbf{x})), (T, \Delta))] \leq \\ \mathbb{E}_{T^*, C, \Delta | \mathbf{X}=\mathbf{x}}[L_\zeta((F_1^*(\zeta|\mathbf{x}), \dots, F_K^*(\zeta|\mathbf{x}), S^*(\zeta|\mathbf{x})), (T, \Delta))] \quad (1) \end{aligned}$$

← Estimated distributions

← Oracle distributions

## 140 4 A Proper Scoring Rule for Competing Risks

141 We prove that the negative log-likelihood re-weighted by the censoring distribution (IPCW) is proper.

142 **Definition 4.1** (*Competitive Weights Negative LogLoss*). We introduce the multiclass negative  
143 log-likelihood re-weighted with the censoring distribution. The different classes represent the loss of  
144 all the cumulative incidence functions as well as the survival function.

$$\begin{aligned} \forall \zeta, (\mathbf{x}, t, \delta) \sim \mathcal{D}, \quad L_\zeta((\hat{F}_1(\zeta|\mathbf{x}), \dots, \hat{F}_K(\zeta|\mathbf{x}), \hat{S}(\zeta|\mathbf{x})), (t, \delta)) \stackrel{\text{def}}{=} \\ \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \left( \frac{\mathbb{1}_{t_i \leq \zeta, \delta_i = k} \log(\hat{F}_k(\zeta|\mathbf{x}_i))}{G^*(t_i|\mathbf{x}_i)} \right) + \frac{\mathbb{1}_{t_i > \zeta} \log(\hat{S}(\zeta|\mathbf{x}_i))}{G^*(\zeta|\mathbf{x}_i)} \quad (2) \end{aligned}$$

← Probability of remaining at  $t_i$

← Probability of remaining at  $\zeta$   
(1 - probability of censoring)

146 Eqn.2 can be seen as a standard log-loss (a.k.a cross-entropy), reweighted by appropriate sample  
147 weights, the inverse probabilities, IPCW (inverse probabilities of censoring weights). It can thus be  
148 easily added to most multiclass estimators.

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

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

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

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

156 *Proof sketch.* With the previous result, the properties of the negative log-likelihood, and the Definition  
 157 3.3, we obtain that the loss is strictly proper. The whole proof can be found in Appendix B.  $\square$

## 158 5 MultiIncidence Model: Gradient boosting for competing risks

159 While eq.2 can be used as a loss in any multiclass machine learning algorithm, we chose Gradient  
 160 Boosting trees because of their performance on tabular data (Grinsztajn et al., 2022) and their ability  
 161 to be fit via stochastic optimization. Gradient boosting methods are designed to approximate complex  
 162 functions through a combination of weak learners (or base learners). At each iteration, the algorithm  
 163 focuses on the residuals of the loss and constructs a base learner  $h_m$  that minimizes the residuals. For  
 164 gradient boosting trees, the final estimator typically takes the form  $H_m(x) = H_{m-1}(x) + \nu h_m(x)$   
 165 where  $\nu$  represents a chosen learning rate. More explanations on gradient boosting methods are  
 166 provided in Friedman (1999).

167 Most survival or competing risk loss cannot be used with such tree-based models as the require  
 168 time-derivates and thus smoothness. So, we introduce a model, MultiIncidence, that predicts all  
 169 CIFs for each competing event, as well as the global survival function. Predicting these jointly  
 170 easily maintains the stability of the probabilities as outputs of classifications model sum to one and  
 171  $\mathbb{P}(T^* \leq \zeta | \mathbf{X} = \mathbf{x}) + \mathbb{P}(T^* > \zeta | \mathbf{X} = \mathbf{x}) = 1$  or

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

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

176 Our algorithm uses two classifiers (here gradient-boosted trees), one for the censoring trained on  
 177 binary censored/non-censored labels (i.e. for time  $\zeta$ ,  $\mathbb{P}(C > \zeta | \mathbf{X} = \mathbf{x})$ ), and one for the multiple  
 178 events. Both of the censoring and event models are corrected with IPCW weights. To compute these  
 179 IPCW we iterate the training using a feedback loop (in the like of boosting). We first compute a  
 180 survival censoring model. Then, with these probabilities, we initiate our MultiIncidence model. After  
 181 several iterations, we apply a feedback loop to retrain our censoring model.  
 182

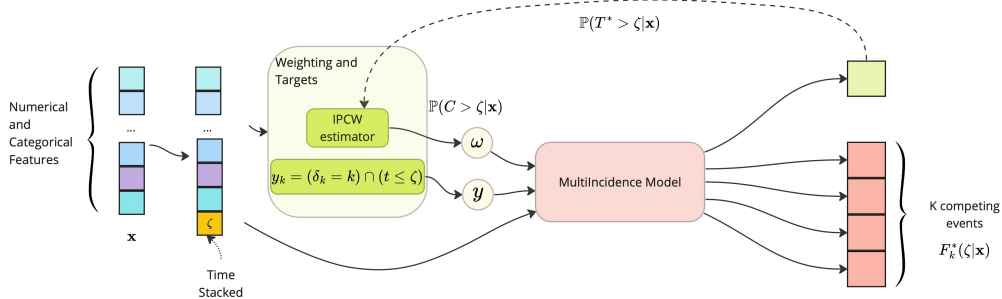


Figure 1: **MultiIncidence Model with its Feedback Loop.** After giving the input to the model, a random time is given and the weights and the target can be computed. After one iteration, the feedback loop trains the censoring probability –  $G^*$  in eq.2.

---

**Algorithm 1** MultiIncidence 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 random time horizon  
 $\tilde{\mathbf{x}} \leftarrow (\mathbf{x}, \zeta)$  ▷Stacking the time to the features  
 $y, w \leftarrow \text{ipcwComputer}(\mathbf{x}, \delta, t, \hat{G})$  ▷Pseudo-code is written in Algo 2  
 $L \leftarrow \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \left( \mathbb{1}_{y_i=k} y_i w_i \log \left( \hat{F}_k(\zeta_i | \mathbf{x}_i) \right) \right) + \mathbb{1}_{y_i=0} y_i w_i \log \left( \hat{S}(\zeta_i | \mathbf{x}_i) \right)$   
 $h_m(\tilde{\mathbf{x}}) \leftarrow \text{Train one iteration of Gradient Boost with } L \text{ as the loss}$  ▷ $h_m$  is the  $m^{th}$  weak learner  
 $H_m(\zeta | \mathbf{x}) \leftarrow h_m(\zeta | \mathbf{x}) + \nu H_{m-1}(\zeta | \mathbf{x})$  ▷ $H_m$  is the  $m^{th}$  estimator  
 $(\hat{S}(\zeta | \mathbf{X} = \mathbf{x}), (\hat{F}_k(\zeta | \mathbf{X} = \mathbf{x})_{1 \leq k \leq K})) \leftarrow \hat{H}_m(\tilde{\mathbf{x}})$   
 $\hat{G} \leftarrow \text{Train one iteration the Censoring Feedback Loop with } \hat{S}(\zeta | \mathbf{X} = \mathbf{x})$  ▷Pseudo-code is written in Algo 3

---

183 To incorporate more complex temporal dependencies into the model, we uniformly sample a time  
 184 point for each observation and include it as an additional feature. Multiple time points can be  
 185 sampled per iteration for each observation. This approach generates a richer dataset, where the  
 186 targets may vary based on the specific times sampled, thus providing the model with a broader  
 187 range of temporal information. This approach is made possible by our loss which is separable. An  
 188 additional benefit is that we can predict the CIF at any time, unlike models that are optimized for a  
 189 limited number of times and need to be interpolated to other times.  
 190

191 As Figure 1 shows an iteration: we compute the weights  $w_i$  and targets  $y_i$  according to the sampled  
 192 times for each individual. Specifically, for censored samples, the corresponding weight is set to 0, as  
 193 determined by the indicator function in eq.2. A target  $y_i \in \llbracket 1, K \rrbracket$  indicates that the event of interest  
 194 has occurred before  $\zeta$ . However, when  $y_i = 0$ , the individual has survived any event. We give a  
 195 pseudocode of the Algorithm 1.

## 196 6 Experimental study: Competing risks

### 197 6.1 Evaluation metrics for competing risks models

198 To evaluate the risks of the different events, we use two metrics<sup>1</sup>.

199 **Evaluating the predicted probability** We extend the method proposed by Graf et al. (1999) and  
 200 Schoop et al. (2011). The detailed formula and a formal proof of the propperness of the loss can be  
 201 found in Appendix ???. To avoid potential circularity with the loss function that our model optimizes,  
 202 we apply this evaluation metric to the Brier Score rather than the log-loss. To evaluate the model at  
 203 all times, we sum it over time, giving the *Integrated Brier Score* (IBS).

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

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

---

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

214 uncensored at  $\zeta$ .

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

215 **6.2 Experimental settings**

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

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

228 **Baselines** We compared our approach to 7 other models. Aalen-Johansen’s estimator (Aalen et al.,  
 229 2008), Fine & Gray’s linear model (Fine & Gray, 1999), a tree-based approach with the Random  
 230 Survival Forests (RSF, Ishwaran et al., 2008), and neural networks: DeepHit (Lee et al., 2018), Deep  
 231 Survival Machines (DSM, Nagpal et al., 2021), DeSurv (Danks & Yau, 2022b) and a transformer  
 232 model with SurvTRACE (Wang & Sun, 2022). DeepHit is trained with a ranking loss: the C-index  
 233 summed with a negative log-likelihood, DSM uses a graphical method for feature encodings while  
 234 DeSurv solves Ordinal Differential Equations for continuous predictions in time. SurvTRACE is  
 235 trained for three-time horizons (based on quantiles of observed event times) and at time 0, while  
 236 Aalen-Johansen and Fine & Gray are trained for all observed event times. In contrast, our method  
 237 is trained on uniformly sampled time horizons, allowing predictions at any time. To compute the  
 238 Integrated Brier Score over time, other methods require linear interpolation of their trained times.  
 239 For times exceeding their trained times, we assume the incidence remains constant. To be fair across  
 240 models, we use the same time budget for hyper-parameter tuning (grid in Appendix S7).

241 **6.3 Results, competing risks**

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

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

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.

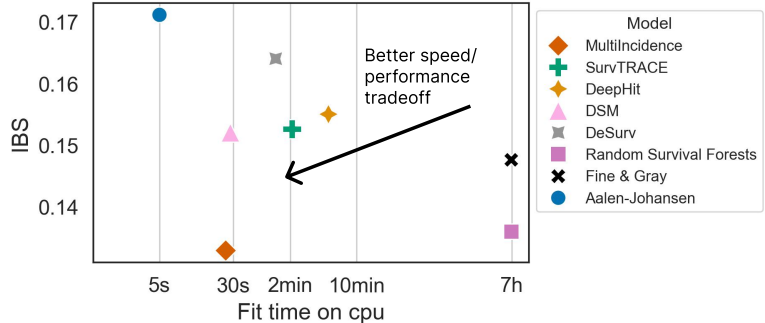




Figure 3: **Trade-off prediction/training time for competing risk on the SEER dataset** Average IBS compared to the fitting time for each model on the maximum training points (330k) except for Fine & Gray (50k) and RSF (100k). Table S2 gives IBS values for each event.

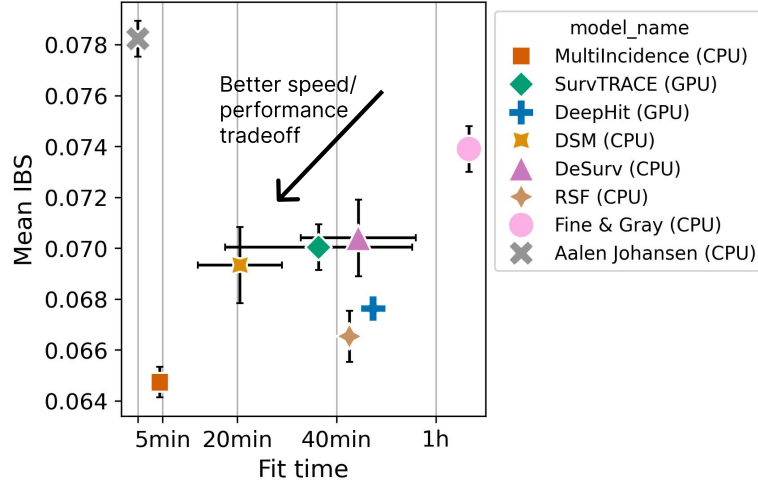
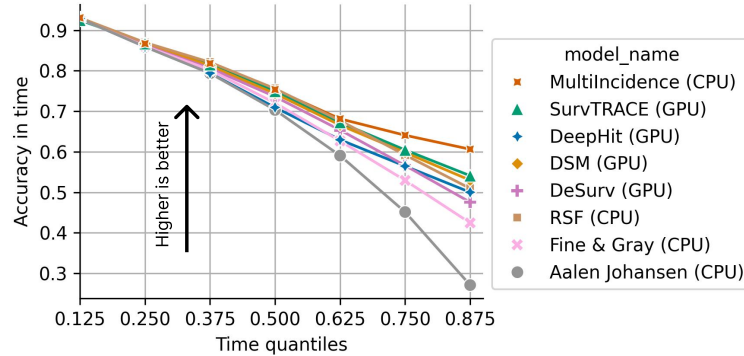


Figure 4: **Prediction accuracy at time  $\zeta$**  Accuracy of the Argmax of the Cumulative Incidence Functions on different quantiles in time on the SEER Dataset (Higher is Better).



248 **Results on SEER Dataset** On the real-life dataset, we keep 30% of the data set to test the models.  
 249 Figure 3 compares models with the Integrated Brier score (with Kaplan-Meier weights of Graf et al.  
 250 (1999) due to lack of oracle). MultiIncidence achieves the best score and the shortest fit time. Random  
 251 Survival Forest is not made to be used with that many samples (100k) and uses more than 50 Gb of  
 252 RAM. MultiIncidence maintains its marked lead with much fewer training samples (Appendix F.2).

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

## 257 7 Usage in Survival Analysis

### 258 7.1 Survival experiments

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

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

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

265 **Evaluation** We use different metrics to evaluate our models. As above we use the Integrated Brier  
 266 Score (detailed in Appendix C), but we also add another metric from Yanagisawa (2023), called  
 267  $S_{Cen-log-simple}$  (detailed in Appendix D). This last metric approximates the proper scoring metric

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          |
| DQS                    | 0.202±0.007        | 1.987±0.069          | 0.180±0.034        | 2.205±0.044          |
| SUMO NET               | 0.194±0.010        | <b>1.721±0.016</b>   | 0.169±0.009        | 2.302±0.059          |
| SURVTRACE              | 0.194±0.005        | 1.870±0.018          | <b>0.168±0.011</b> | 2.270±0.034          |
| MULTIINCIDENCE         | <b>0.191±0.006</b> | 1.740±0.020          | <b>0.168±0.019</b> | <b>2.169±0.056</b>   |

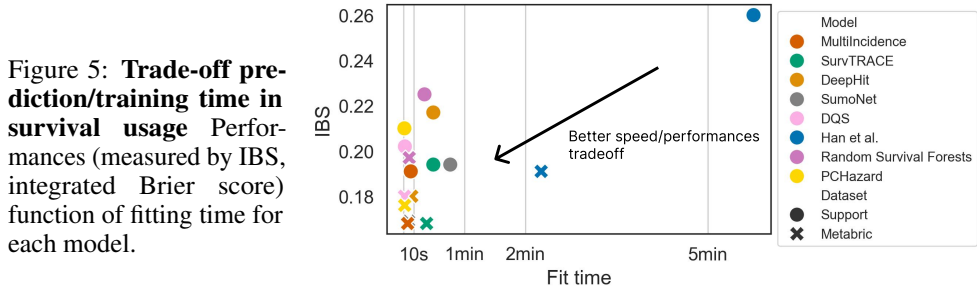


Figure 5: **Trade-off prediction/training time in survival usage** Performances (measured by IBS, integrated Brier score) function of fitting time for each model.

268 in Rindt et al. (2022) –and is not exactly proper, see Appendix D. It is useful because it can be used  
 269 on any model as it does not require the density of the Cumulative Incidence Function.

270 **Baselines** We compare our model with SOTA competing risks models, including SurvTRACE  
 271 (Wang & Sun, 2022), DeepHit (Lee et al., 2018) and Random Survival Forests (Ishwaran et al., 2008).  
 272 We also benchmark some SOTA survival ones: neural networks e.g. (PCHazard Kvamme & Borgan,  
 273 2019b), survival game (Han et al., 2021) and neural networks trained with a proper survival-analysis  
 274 scoring rule, e.g. SumoNet (Rindt et al., 2022), and DQS (Yanagisawa, 2023).

## 275 7.2 Results in survival usage

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

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

## 288 Discussion and Conclusion

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

290 **Acknowledgement** JA, JA, and GV acknowledge funding from the ERC grand INTERCEPT-T2D.

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

293 **Limitations and further work** Further work should consider removing the assumption of non-  
294 informative censoring. This assumption is very common in the literature, though some recent work  
295 has relaxed it in survival settings (Foomani et al., 2023; Zhang et al., 2023).

296 **Conclusion** For competing risks, which is a generalization of survival analysis to classify the type  
297 of outcome, we first propose and prove a (strictly) proper scoring rule. It is a reweighted log loss that  
298 can easily be used as a loss for machine learning: it is separable in the observations and thus suited  
299 to stochastic solvers; it does not require time-wise derivative (unlike most survival models) and can  
300 be used in non-differentiable models. We plug it into gradient-boosting trees, in an algorithm called  
301 MultiIncidence. Thanks to time used as a feature and its feedback loop to better estimate censoring  
302 probabilities, MultiIncidence outperforms state-of-the-art methods on a synthetic dataset as well as  
303 real-life datasets both for competing risk (classification on time-censored data) and standard survival  
304 (time-to-event regression with right censoring). It is also faster to train over many samples. As a loss,  
305 it easily brings survival or competing risks to many models: scalable linear models to replace clinical  
306 standard Fine & Gray that do not scale, or deep learning, including fine-tuning foundation models.

## 307 References

- 308 Aala, A. M. and van der Schaar, M. Deep Multi-task Gaussian Processes for Survival Analysis with  
309 Competing Risks. In *Advances in Neural Information Processing Systems*, volume 30. Curran  
310 Associates, Inc., 2017. URL [https://papers.nips.cc/paper\\_files/paper/2017/hash/8](https://papers.nips.cc/paper_files/paper/2017/hash/861dc9bd7f4e7dd3cccd534d0ae2a2e9-Abstract.html)  
311 [61dc9bd7f4e7dd3cccd534d0ae2a2e9-Abstract.html](https://papers.nips.cc/paper_files/paper/2017/hash/861dc9bd7f4e7dd3cccd534d0ae2a2e9-Abstract.html).
- 312 Aalen, O. O., Borgan, O., and Gjessing, H. K. *Survival and Event History Analysis*. Statistics  
313 for Biology and Health. Springer New York, New York, NY, 2008. ISBN 978-0-387-20287-7  
314 978-0-387-68560-1. doi: 10.1007/978-0-387-68560-1. URL [http://link.springer.com/10](http://link.springer.com/10.1007/978-0-387-68560-1)  
315 [.1007/978-0-387-68560-1](http://link.springer.com/10.1007/978-0-387-68560-1).
- 316 Antolini, L., Boracchi, P., and Biganzoli, E. A time-dependent discrimination index for survival data.  
317 *Statistics in Medicine*, 24(24):3927–3944, December 2005. ISSN 0277-6715, 1097-0258. doi:  
318 10.1002/sim.2427. URL <https://onlinelibrary.wiley.com/doi/10.1002/sim.2427>.
- 319 Bellot, A. and Schaar, M. Tree-based Bayesian Mixture Model for Competing Risks. In *Proceedings*  
320 *of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pp. 910–918.  
321 PMLR, March 2018. URL <https://proceedings.mlr.press/v84/bellot18a.html>. ISSN:  
322 2640-3498.
- 323 Bellot, A. and van der Schaar, M. Multitask Boosting for Survival Analysis with Competing Risks.  
324 In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.  
325 URL [https://proceedings.neurips.cc/paper\\_files/paper/2018/hash/2afe4567e](https://proceedings.neurips.cc/paper_files/paper/2018/hash/2afe4567e1bf64d32a5527244d104cea-Abstract.html)  
326 [1bf64d32a5527244d104cea-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2018/hash/2afe4567e1bf64d32a5527244d104cea-Abstract.html).
- 327 Benedetti, R. Scoring Rules for Forecast Verification. *Monthly Weather Review*, 138(1):203–  
328 211, January 2010. ISSN 1520-0493, 0027-0644. doi: 10.1175/2009MWR2945.1. URL  
329 <http://journals.ametsoc.org/doi/10.1175/2009MWR2945.1>.
- 330 Blanche, P., Dartigues, J.-F., and Jacqmin-Gadda, H. Estimating and comparing time-dependent  
331 areas under receiver operating characteristic curves for censored event times with competing risks.  
332 *Statistics in medicine*, 32(30):5381–5397, 2013.
- 333 Blanche, P., Kattan, M. W., and Gerds, T. A. The c-index is not proper for the evaluation of  $t$ -year  
334 predicted risks. *Biostatistics*, 20(2):347–357, April 2019. ISSN 1465-4644, 1468-4357. doi:  
335 10.1093/biostatistics/kxy006. URL [https://academic.oup.com/biostatistics/article](https://academic.oup.com/biostatistics/article/20/2/347/4864363)  
336 [/20/2/347/4864363](https://academic.oup.com/biostatistics/article/20/2/347/4864363).
- 337 Chaddad, A., Desrosiers, C., and Toews, M. Radiomic analysis of multi-contrast brain MRI for the  
338 prediction of survival in patients with glioblastoma multiforme. In *2016 38th Annual International*  
339 *Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4035–4038,  
340 Orlando, FL, USA, August 2016. IEEE. ISBN 978-1-4577-0220-4. doi: 10.1109/EMBC.2016.75  
341 91612. URL <http://ieeexplore.ieee.org/document/7591612/>.

- 342 Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B*  
343 *(Methodological)*, 34(2):187–202, January 1972. ISSN 0035-9246, 2517-6161. doi: 10.1111/j.25  
344 17-6161.1972.tb00899.x. URL [https://rss.onlinelibrary.wiley.com/doi/10.1111/j.  
345 2517-6161.1972.tb00899.x](https://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1972.tb00899.x).
- 346 Danks, D. and Yau, C. Derivative-Based Neural Modelling of Cumulative Distribution Functions  
347 for Survival Analysis. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings*  
348 *of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of  
349 *Proceedings of Machine Learning Research*, pp. 7240–7256. PMLR, March 2022a. URL [https:  
350 //proceedings.mlr.press/v151/danks22a.html](https://proceedings.mlr.press/v151/danks22a.html).
- 351 Danks, D. and Yau, C. Derivative-based neural modelling of cumulative distribution functions for  
352 survival analysis, 28–30 Mar 2022b. URL [https://proceedings.mlr.press/v151/danks  
353 22a.html](https://proceedings.mlr.press/v151/danks22a.html).
- 354 Fine, J. P. and Gray, R. J. A Proportional Hazards Model for the Subdistribution of a Competing Risk.  
355 *Journal of the American Statistical Association*, 94(446):496–509, June 1999. ISSN 0162-1459,  
356 1537-274X. doi: 10.1080/01621459.1999.10474144. URL [http://www.tandfonline.com/do  
357 i/abs/10.1080/01621459.1999.10474144](http://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10474144).
- 358 Foomani, A. H. G., Cooper, M., Greiner, R., and Krishnan, R. G. Copula-based deep survival models  
359 for dependent censoring, 2023.
- 360 Friedman, J. H. Greedy function approximation: A gradient boosting machine. 1999.
- 361 Gaynor, J. J., Feuer, E. J., Tan, C. C., Wu, D. H., Little, C. R., Straus, D. J., Clarkson, B. D.,  
362 and Brennan, M. F. On the Use of Cause-Specific Failure and Conditional Failure Probabilities:  
363 Examples from Clinical Oncology Data. *Journal of the American Statistical Association*, 88(422):  
364 400–409, June 1993. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1993.10476289. URL  
365 <http://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476289>.
- 366 Gneiting, T. and Raftery, A. E. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of*  
367 *the American Statistical Association*, 102(477):359–378, March 2007. ISSN 0162-1459, 1537-  
368 274X. doi: 10.1198/016214506000001437. URL [http://www.tandfonline.com/doi/abs/  
369 10.1198/016214506000001437](http://www.tandfonline.com/doi/abs/10.1198/016214506000001437).
- 370 Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. Assessment and comparison of prognostic  
371 classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, September  
372 1999. ISSN 0277-6715, 1097-0258. doi: 10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529:  
373 AID-SIM274>3.0.CO;2-5. URL [https://onlinelibrary.wiley.com/doi/10.1002/\(SI  
374 CI\)1097-0258\(19990915/30\)18:17/18<2529::AID-SIM274>3.0.CO;2-5](https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5).
- 375 Grinsztajn, L., Oyallon, E., and Varoquaux, G. Why do tree-based models still outperform deep learn-  
376 ing on tabular data?, July 2022. URL <http://arxiv.org/abs/2207.08815>. arXiv:2207.08815  
377 [cs, stat].
- 378 Han, X., Goldstein, M., Puli, A., Wies, T., Perotte, A., and Ranganath, R. Inverse-weighted survival  
379 games. *Advances in neural information processing systems*, 34:2160–2172, 2021.
- 380 Harrell, Frank E., J., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. Evaluating the Yield of  
381 Medical Tests. *JAMA*, 247(18):2543–2546, 05 1982. ISSN 0098-7484. doi: 10.1001/jama.1982.  
382 03320430047030. URL <https://doi.org/10.1001/jama.1982.03320430047030>.
- 383 Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. Random survival forests. *The*  
384 *Annals of Applied Statistics*, 2(3), September 2008. ISSN 1932-6157. doi: 10.1214/08-AOAS169.  
385 URL <http://arxiv.org/abs/0811.1645>. arXiv:0811.1645 [stat].
- 386 Ishwaran, H., Gerds, T. A., Kogalur, U. B., Moore, R. D., Gange, S. J., and Lau, B. M. Random  
387 survival forests for competing risks. *Biostatistics*, 15(4):757–773, 2014.
- 388 Kaplan, E. L. and Meier, P. Nonparametric Estimation from Incomplete Observations. *Journal of the*  
389 *American Statistical Association*, 53(282):457–481, June 1958. ISSN 0162-1459, 1537-274X. doi:  
390 10.1080/01621459.1958.10501452. URL [http://www.tandfonline.com/doi/abs/10.108  
391 0/01621459.1958.10501452](http://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501452).

- 392 Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. DeepSurv: personalized  
393 treatment recommender system using a Cox proportional hazards deep neural network. *BMC*  
394 *Medical Research Methodology*, 18(1):24, December 2018. ISSN 1471-2288. doi: 10.1186/s128  
395 74-018-0482-1. URL [https://bmcmedresmethodol.biomedcentral.com/articles/10.](https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0482-1)  
396 [1186/s12874-018-0482-1](https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0482-1).
- 397 Koene, R. J., Prizment, A. E., Blaes, A., and Konety, S. H. Shared risk factors in cardiovascular  
398 disease and cancer. *Circulation*, 133(11):1104–1114, 2016.
- 399 Koller, M. T., Raatz, H., Steyerberg, E. W., and Wolbers, M. Competing risks and the clinical  
400 community: irrelevance or ignorance? *Statistics in medicine*, 31(11-12):1089–1097, 2012.
- 401 Kretowska, M. Tree-based models for survival data with competing risks. *Computer Methods and*  
402 *Programs in Biomedicine*, 159:185–198, June 2018. ISSN 01692607. doi: 10.1016/j.cmpb.2018.  
403 03.017. URL <https://linkinghub.elsevier.com/retrieve/pii/S0169260717314347>.
- 404 Kvamme, H. and Borgan, O. The Brier Score under Administrative Censoring: Problems and  
405 Solutions, December 2019a. URL <http://arxiv.org/abs/1912.08581>. arXiv:1912.08581  
406 [cs, stat].
- 407 Kvamme, H. and Borgan, o. Continuous and Discrete-Time Survival Prediction with Neural Networks,  
408 October 2019b. URL <http://arxiv.org/abs/1910.06724>. arXiv:1910.06724 [cs, stat].
- 409 Lee, C., Zame, W., Yoon, J., and Van Der Schaar, M. DeepHit: A Deep Learning Approach to  
410 Survival Analysis With Competing Risks. *Proceedings of the AAAI Conference on Artificial*  
411 *Intelligence*, 32(1), April 2018. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v32i1.11842.  
412 URL <https://ojs.aaai.org/index.php/AAAI/article/view/11842>.
- 413 Maystre, L. and Russo, D. Temporally-Consistent Survival Analysis.
- 414 Merkle, E. C. and Steyvers, M. Choosing a Strictly Proper Scoring Rule. *Decision Analysis*, 10(4):  
415 292–304, December 2013. ISSN 1545-8490, 1545-8504. doi: 10.1287/deca.2013.0280. URL  
416 <https://pubsonline.informs.org/doi/10.1287/deca.2013.0280>.
- 417 Monterrubio-Gómez, K., Constantine-Cooke, N., and Vallejos, C. A. A review on competing risks  
418 methods for survival analysis, December 2022. URL <http://arxiv.org/abs/2212.05157>.  
419 arXiv:2212.05157 [stat].
- 420 Nagpal, C., Li, X. R., and Dubrawski, A. Deep survival machines: Fully parametric survival  
421 regression and representation learning for censored data with competing risks, 2021.
- 422 National Cancer Institute, DCCPS, S. R. P. Surveillance, epidemiology, and end results (seer) program  
423 (www.seer.cancer.gov) seer\*stat database: Incidence - seer research data, 8 registries, nov 2021 sub  
424 (1975-2020) - linked to county attributes - time dependent (1990-2020) income/rurality, 1969-2020  
425 counties, national cancer institute, dccps, surveillance research program, , based on the november  
426 2022 submission, 2023.
- 427 Nelson, W. Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics*, 14  
428 (4):945–966, November 1972. ISSN 0040-1706, 1537-2723. doi: 10.1080/00401706.1972.104889  
429 91. URL <http://www.tandfonline.com/doi/abs/10.1080/00401706.1972.10488991>.
- 430 Ovcharov, E. Y. Proper scoring rules and Bregman divergence. *Bernoulli*, 24(1), February 2018.  
431 ISSN 1350-7265. doi: 10.3150/16-BEJ857. URL [https://projecteuclid.org/journals/b](https://projecteuclid.org/journals/bernoulli/volume-24/issue-1/Proper-scoring-rules-and-Bregman-divergence/10.3150/16-BEJ857.full)  
432 [ernoulli/volume-24/issue-1/Proper-scoring-rules-and-Bregman-divergence/10](https://projecteuclid.org/journals/bernoulli/volume-24/issue-1/Proper-scoring-rules-and-Bregman-divergence/10.3150/16-BEJ857.full)  
433 [.3150/16-BEJ857.full](https://projecteuclid.org/journals/bernoulli/volume-24/issue-1/Proper-scoring-rules-and-Bregman-divergence/10.3150/16-BEJ857.full).
- 434 Ramspek, C. L., Teece, L., Snell, K. I., Evans, M., Riley, R. D., van Smeden, M., van Geloven, N.,  
435 and van Diepen, M. Lessons learnt when accounting for competing events in the external validation  
436 of time-to-event prognostic models. *International journal of epidemiology*, 51(2):615–625, 2022.
- 437 Rindt, D., Hu, R., Steinsaltz, D., and Sejdinovic, D. Survival Regression with Proper Scoring Rules  
438 and Monotonic Neural Networks, February 2022. URL <http://arxiv.org/abs/2103.14755>.  
439 arXiv:2103.14755 [cs, stat].

- 440 Rith, M., Soliman, J., Fillone, A., Biona, J. B. M., and Lopez, N. S. Analysis of Vehicle Survival Rates  
441 for Metro-Manila. In *2018 IEEE 10th International Conference on Humanoid, Nanotechnology,  
442 Information Technology, Communication and Control, Environment and Management (HNICEM)*,  
443 pp. 1–4, Baguio City, Philippines, November 2018. IEEE. ISBN 978-1-5386-7767-4. doi: 10.110  
444 9/HNICEM.2018.8666408. URL <https://ieeexplore.ieee.org/document/8666408/>.
- 445 Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of Regression Coefficients When Some  
446 Regressors are not Always Observed. *Journal of the American Statistical Association*, 89(427):  
447 846–866, September 1994. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1994.10476818.  
448 URL <https://www.tandfonline.com/doi/full/10.1080/01621459.1994.10476818>.
- 449 Schoop, R., Beyersmann, J., Schumacher, M., and Binder, H. Quantifying the predictive accuracy  
450 of time-to-event models in the presence of competing risks. *Biometrical Journal*, 53(1):88–112,  
451 February 2011. ISSN 03233847. doi: 10.1002/bimj.201000073. URL [https://onlinelibrar  
452 y.wiley.com/doi/10.1002/bimj.201000073](https://onlinelibrary.wiley.com/doi/10.1002/bimj.201000073).
- 453 Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., and Beghi, A. Machine Learning for Predictive  
454 Maintenance: A Multiple Classifier Approach. *IEEE Transactions on Industrial Informatics*, 11  
455 (3):812–820, June 2015. ISSN 1551-3203, 1941-0050. doi: 10.1109/TII.2014.2349359. URL  
456 <http://ieeexplore.ieee.org/document/6879441/>.
- 457 Uno, H., Cai, T., Pencina, M. J., D’Agostino, R. B., and Wei, L. J. On the C-statistics for evaluating  
458 overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*,  
459 30(10):1105–1117, May 2011. ISSN 0277-6715, 1097-0258. doi: 10.1002/sim.4154. URL  
460 <https://onlinelibrary.wiley.com/doi/10.1002/sim.4154>.
- 461 Van Belle, V., Pelckmans, K., Van Huffel, S., and Suykens, J. A. Support vector methods for survival  
462 analysis: a comparison between ranking and regression approaches. *Artificial Intelligence in  
463 Medicine*, 53(2):107–118, October 2011. ISSN 09333657. doi: 10.1016/j.artmed.2011.06.006.  
464 URL <https://linkinghub.elsevier.com/retrieve/pii/S0933365711000765>.
- 465 Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., Steyerberg, E. W., and Topic Group  
466 ‘Evaluating diagnostic tests prediction models’ of the STRATOS initiative. Calibration: the achilles  
467 heel of predictive analytics. *BMC medicine*, 17(1):230, 2019.
- 468 van Walraven, C. and McAlister, F. A. Competing risk bias was common in kaplan–meier risk  
469 estimates published in prominent medical journals. *Journal of clinical epidemiology*, 69:170–173,  
470 2016.
- 471 Wang, Z. and Sun, J. SurvTRACE: Transformers for Survival Analysis with Competing Events.  
472 In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational  
473 Biology and Health Informatics*, pp. 1–9, August 2022. doi: 10.1145/3535508.3545521. URL  
474 <http://arxiv.org/abs/2110.00855>. arXiv:2110.00855 [cs, stat].
- 475 Wolbers, M., Koller, M. T., Wittman, J. C., and Steyerberg, E. W. Prognostic models with competing  
476 risks: methods and application to coronary risk prediction. *Epidemiology*, 20(4):555–561, 2009.
- 477 Yanagisawa, H. Proper scoring rules for survival analysis, 2023.
- 478 Zhang, W., Ling, C. K., and Zhang, X. Deep copula-based survival analysis for dependent censoring  
479 with identifiability guarantees, 2023.
- 480 Zhu, X., Yao, J., and Huang, J. Deep convolutional neural network for survival analysis with  
481 pathological images. In *2016 IEEE International Conference on Bioinformatics and Biomedicine  
482 (BIBM)*, pp. 544–547, Shenzhen, China, December 2016. IEEE. ISBN 978-1-5090-1611-2. doi:  
483 10.1109/BIBM.2016.7822579. URL <http://ieeexplore.ieee.org/document/7822579/>.

484 **A Definitions**

485 **A.1 Notations**

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

488 For all symbols, we use the following conventions:

- 489 •  $\cdot^*$ : Oracle
- 490 •  $\hat{\cdot}$ : Estimation

The different variables that we use are:

| Maths Symbol   | Domain           | Description  |
|----------------|------------------|--|
| $\zeta$        | $\mathbb{R}_+$   | 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           |
| $T^*$          | $\mathbb{R}_+$   | $\min(T_1^*, T_2^*, \dots, T_K^*)$                 |
| $T$            | $\mathbb{R}_+$   | $\min(T, C)$                                       |
| $\Delta^*$     | $[1, K]$         | $\arg \min_{k \in [1, K]} (T_k^*)$                 |
| $\Delta$       | $[0, K]$         | $\arg \min(C, T_1^*, T_2^*, \dots, T_K^*)$         |
| $S$            | $\mathcal{S}$    | 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                   |
| $\delta_i$     | $[0, K]$         | event observed, 0 indicates censoring              |

Table S1: Notations used

491

492 **A.2 Reporting conventions**

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

494 **B Theory on our proper scoring rule: proofs and derivations**

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

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

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

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

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

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

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

500 **Computation of the expectation:** First:

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

501 The last equality can be detailed as in:

$$\mathbb{E}_{T^*, C, \Delta | \mathbf{X} = \mathbf{x}} \left[ \frac{\mathbb{1}_{T^* \leq \zeta} \mathbb{1}_{T^* \leq C} \mathbb{1}_{\Delta = k}}{G^*(T|\mathbf{x})} \right] = \int_0^\infty \int_0^\infty (\mathbb{1}_{\min(t, c) = t} + \underbrace{\mathbb{1}_{\min(t, c) = c}}_{=0 \text{ because } k \neq 0}) \frac{\mathbb{1}_{t \leq \zeta} \mathbb{1}_{t \leq c}}{G^*(t|\mathbf{x})} f_{T^*, C, \Delta}(t, c, k | \mathbf{x}) dt dc \quad (6)$$

$$\mathbf{T} \text{ is a composition of } T^* \text{ and } C \quad (7)$$

$$= \int_0^\infty \int_0^\infty \frac{\mathbb{1}_{t \leq \zeta} \mathbb{1}_{t \leq c}}{G^*(t|\mathbf{x})} f_{T^*, C, \Delta}(t, c, k | \mathbf{x}) dt dc \quad (8)$$

$$= \int_0^\infty \int_0^\infty \frac{\mathbb{1}_{t \leq \zeta} \mathbb{1}_{t \leq c}}{G^*(t|\mathbf{x})} f_{T^*, \Delta}(t, k | \mathbf{x}) f_C(c | \mathbf{x}) dt dc \quad (9)$$

$$\text{Because } T^* \perp C | \mathbf{X} \quad (10)$$

$$= \int_0^\infty \frac{\mathbb{1}_{t \leq \zeta}}{G^*(t|\mathbf{x})} f_{T^*, \Delta}(t, k | \mathbf{x}) \left( \int_0^\infty \mathbb{1}_{t \leq c} f_C(c | \mathbf{x}) dc \right) dt \quad (11)$$

$$= \int_0^\infty \frac{\mathbb{1}_{t \leq \zeta}}{G^*(t|\mathbf{x})} f_{T^*, \Delta}(t, k | \mathbf{x}) (G^*(t|\mathbf{x})) dt \quad (12)$$

$$\text{with the definition of } G^* \quad (13)$$

$$= \int_0^\infty \mathbb{1}_{t \leq \zeta} f_{T^*, \Delta}(t, k | \mathbf{x}) dt \quad (14)$$

$$= \mathbb{P}(T^* \leq \zeta, \Delta = k | \mathbf{X} = \mathbf{x}) \quad (15)$$



502 And:

$$\begin{aligned}
\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}} \left[ \Lambda_{k,\zeta}(\hat{S}(\zeta|\mathbf{X}=\mathbf{x}), (T, \Delta)) | \mathbf{X}=\mathbf{x} \right] &= \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}} \left[ \mathbb{1}_{T>\zeta} \frac{\log \left( \hat{S}(\zeta|\mathbf{X}=\mathbf{x}) \right)}{G^*(\zeta|\mathbf{x})} \right] \\
&= \log \left( \hat{S}(\zeta|\mathbf{X}=\mathbf{x}) \right) \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}} \left[ \frac{\mathbb{1}_{\min(T^*, C) > \zeta}}{G^*(\zeta|\mathbf{x})} \right] \\
&= \log \left( \hat{S}(\zeta|\mathbf{X}=\mathbf{x}) \right) \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}} \left[ \frac{\mathbb{1}_{T^* > \zeta} \mathbb{1}_{C > \zeta}}{G^*(\zeta|\mathbf{x})} \right] \\
&= \log \left( \hat{S}(\zeta|\mathbf{X}=\mathbf{x}) \right) \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}} \left[ \frac{\mathbb{1}_{C > \zeta}}{G^*(\zeta|\mathbf{x})} \right] \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}} [\mathbb{1}_{T^* > \zeta}] \\
&\text{Because } T^* \perp\!\!\!\perp C | \mathbf{X} \\
&= \log \left( \hat{S}(\zeta|\mathbf{X}=\mathbf{x}) \right) \frac{\mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}} [\mathbb{1}_{C > \zeta}]}{G^*(\zeta|\mathbf{x})} \mathbb{E}_{T,\Delta|\mathbf{X}=\mathbf{x}} [\mathbb{1}_{T^* > \zeta}] \\
&\text{Because } G^*(\zeta|\mathbf{x}) \text{ does not depend of } T \text{ and } \Delta \\
&= \log \left( \hat{S}(\zeta|\mathbf{X}=\mathbf{x}) \right) \mathbb{P}(T^* > \zeta | \mathbf{X}=\mathbf{x})
\end{aligned}$$

503 By summing all of the terms, we obtain:

$$\begin{aligned}
\mathbb{E}_{T^*, C, \Delta | \mathbf{X}=\mathbf{x}} \left[ L_\zeta \left( (\hat{F}_1(\zeta|\mathbf{x}), \dots, \hat{F}_K(\zeta|\mathbf{x}), \hat{S}(\zeta|\mathbf{x})), (T, \Delta) \right) \right] \\
= \sum_{k=1}^K \log \left( \hat{F}_k(\zeta|\mathbf{x}), 1 \right) \mathbb{P}(T^* \leq \zeta, \Delta = k | \mathbf{X}=\mathbf{x}) \\
+ \log \left( \hat{S}(\zeta|\mathbf{X}=\mathbf{x}) \right) \mathbb{P}(T^* > \zeta | \mathbf{X}=\mathbf{x}) \quad (16)
\end{aligned}$$

504

$$= \sum_{k=1}^K \log \left( \hat{F}_k(\zeta|\mathbf{x}) \right) F_k^*(\zeta|\mathbf{x}) + \log \left( \hat{S}(\zeta|\mathbf{x}) \right) S^*(\zeta|\mathbf{x}) \quad (17)$$

505 Finally:

$$\begin{aligned}
\mathbb{E}_{T^*, C, \Delta | \mathbf{X}=\mathbf{x}} \left[ L_\zeta \left( (\hat{F}_1(\zeta|\mathbf{x}), \dots, \hat{F}_K(\zeta|\mathbf{x}), \hat{S}(\zeta|\mathbf{x})), (T, \Delta) \right) \right] \\
= \sum_{k=1}^K \log \left( \hat{F}_k(\zeta|\mathbf{x}) \right) F_k^*(\zeta|\mathbf{x}) + \log \left( \hat{S}(\zeta|\mathbf{x}) \right) S^*(\zeta|\mathbf{x}) \quad (18)
\end{aligned}$$

506

□

507 *Proof of the Theorem 1.*

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

511 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:

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

512 So the previous quantities of interest can be rewritten as functions of those variables:

$$F_k^*(\zeta|\mathbf{x}) = \mathbb{P}(T^* \leq \zeta, \Delta = k | \mathbf{X} = \mathbf{x}) = \mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x}) \quad (19)$$

513

$$S^*(\zeta|\mathbf{x}) = \mathbb{P}(T^* > \zeta | \mathbf{X} = \mathbf{x}) = \mathbb{P}(Y_{0,\zeta} = 1 | \mathbf{X} = \mathbf{x}) \quad (20)$$

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

$$\mathbb{E}_{T^*, C, \Delta | \mathbf{X} = \mathbf{x}} \left[ \mathbb{L}_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}), (T, \Delta)) \right] = \mathbb{E}_{T, \Delta | \mathbf{X} = \mathbf{x}} [\mathbb{L}_\zeta(\hat{p}_\zeta, (T, \Delta))] \quad (21)$$

$$= \sum_{k=0}^K \log(\hat{p}_{k,\zeta}) \mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x}) \quad (22)$$

516 Thus, we obtain the following optimization problem:

$$\begin{aligned} \max_{\hat{p}} \quad & \sum_{k=0}^K \log(\hat{p}_{k,\zeta}) \mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x}) \\ \text{s.t.} \quad & \sum_{k=0}^K \hat{p}_k = 1 \\ & \hat{p}_k \geq 0 \end{aligned} \quad (23)$$

517 The problem can be rewritten as a convex optimization problem because of the concavity of the  
518 logarithm:

$$\begin{aligned} \min_{\hat{p}} \quad & - \sum_{k=0}^K \log(\hat{p}_{k,\zeta}) \mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x}) \\ \text{s.t.} \quad & \sum_{k=0}^K \hat{p}_k = 1 \\ & \hat{p}_k \geq 0 \end{aligned} \quad (24)$$

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

$$\nabla \left( - \sum_{k=1}^K \log(\hat{p}_{k,\zeta}) \mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x}) - \log(\hat{p}_{0,\zeta}) \mathbb{P}(Y_{0,\zeta} = 1 | \mathbf{X} = \mathbf{x}) \right) + \lambda - \mu \mathbf{1} = 0 \quad (25)$$

$$\forall k, \mu_k p_k = 0 \quad (26)$$

522 If  $\exists k, p_k = 0$ ,  $-\sum_{k=1}^K \log(\hat{p}_{k,\zeta}) \mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x}) - \log(\hat{p}_{0,\zeta}) \mathbb{P}(Y_{0,\zeta} = 1 | \mathbf{X} = \mathbf{x}) = \infty$ .  
523 Hence, (24) implies that  $\forall k, \mu_k = 0$ .

524

525 Now,

$$\forall k, \frac{\partial \left( -\sum_{k=0}^K \log(\hat{p}_{k,\zeta}) \mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x}) \right)}{\partial \hat{p}_{k,\zeta}} = -\frac{\mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x})}{\hat{p}_{k,\zeta}} \quad (27)$$

(24) can be rewritten as: (28)

$$\forall k, -\frac{\mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x})}{\hat{p}_{k,\zeta}} + \lambda = 0 \quad (29)$$

$$\implies \forall k, -\mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x}) + \lambda \hat{p}_{k,\zeta} = 0 \quad (30)$$

By summing over  $k$ , (31)

$$\implies -\underbrace{\sum_{k=0}^K \mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x})}_{=1} + \lambda \underbrace{\sum_{k=0}^K \hat{p}_{k,\zeta}}_{=1} = 0 \quad (32)$$

$$\implies \lambda = 1 \quad (33)$$

$$\implies \forall k, \hat{p}_{\zeta,k} = \mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x}) \quad (34)$$

526 Any local minima must fulfill the KKT theorem. Thus if  $p$  is a local minima, then the local minima  
 527 is a solution to (24) and (25). Consequently the above applies, we do obtain that the only possible  
 528 solution must be equal to the oracle distribution. Indeed, the loss is strictly proper.

529

□

## 530 C Study of the proper scoring rule used for evaluation

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

### 536 C.1 PSR for evaluation

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

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

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

$$\forall \zeta, (\mathbf{X}, T, \Delta) \sim \mathcal{D}, \quad \mathbb{E}_{T^*, C, \Delta | \mathbf{X}=\mathbf{x}}[L_{k,\zeta}(\hat{F}_k(\zeta | \mathbf{x}), (T, \Delta))] \leq \mathbb{E}_{T, \Delta | \mathbf{X}=\mathbf{x}}[L_{k,\zeta}(F_k^*(\zeta | \mathbf{x}), (T, \Delta))] \quad (35)$$

#### 545 C.1.1 A Proper Scoring Rule for Competing Risks

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

549 In the following, we prove that any given (strictly) proper scoring rule that can be used in the  
 550 multiclass setting (e.g. the Brier score, the negative log-likelihood) leads to a (strictly) proper scoring

551 in competing risks settings thanks to the re-weighting of the observations.  
 552 Indeed, for any (strictly) proper scoring rule  $\ell : \mathbb{R} \times \{0, 1\} \rightarrow \mathbb{R}$ , one can build a cause-specific scoring  
 553 rule function  $L_{k,\zeta} : \mathbb{R} \times \mathcal{D} \rightarrow \mathbb{R}$  that is also a (strictly) proper scoring rule for the cause-specific  
 554 event  $k^{\text{th}}$  in the fixed time horizon  $\zeta \in \mathbb{R}_+$ . It follows that  $L_\zeta$  is (strictly) proper.

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

$$\begin{aligned} & \forall \zeta, \forall k \in \llbracket 1, K \rrbracket, \ell : \mathbb{R} \times \{0, 1\} \rightarrow \mathbb{R}, (\mathbf{x}, t, \delta) \sim \mathcal{D} \\ & L_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}), (t, \delta)) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}_{t_i \leq \zeta, \delta_i = k} \ell(\hat{F}_k(\zeta|\mathbf{x}_i), 1)}{G^*(t_i|\mathbf{x}_i)} \\ & \quad + \frac{\mathbb{1}_{t_i > \zeta} \ell(\hat{F}_k(\zeta|\mathbf{x}_i), 0)}{G^*(\zeta|\mathbf{x}_i)} \\ & \quad + \frac{\mathbb{1}_{t_i \leq \zeta, \delta_i \neq 0, \delta_i \neq k} \ell(\hat{F}_k(\zeta|\mathbf{x}_i), 0)}{G^*(t_i|\mathbf{x}_i)} \end{aligned} \quad (36)$$

Probability of remaining at  $\zeta$  (1 - probability of censoring) Probability of remaining at  $t_i$

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

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

$$\begin{aligned} & \forall \zeta, \forall k \in \llbracket 1, K \rrbracket, \ell : \mathbb{R} \times \{0, 1\} \rightarrow \mathbb{R}, (\mathbf{X}, T, \Delta) \sim \mathcal{D}, \\ & \mathbb{E}_{T^*, C, \Delta | \mathbf{X} = \mathbf{x}} \left[ L_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}), (T, \Delta)) \right] = \ell(\hat{F}_k(\zeta|\mathbf{x}), 1) F_k^*(\zeta|\mathbf{x}) + \ell(\hat{F}_k(\zeta|\mathbf{x}), 0) (1 - F_k^*(\zeta|\mathbf{x})) \end{aligned} \quad (37)$$

*Proof.*

$$\forall \zeta, \forall k \in \llbracket 1, K \rrbracket, \ell : \mathbb{R} \times \{0, 1\} \rightarrow \mathbb{R}, (\mathbf{x}, t, \delta) \sim \mathcal{D}$$

$$\begin{aligned} L_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}), (t, \delta)) & \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}_{t_i \leq \zeta, \delta_i = k} \ell(\hat{F}_k(\zeta|\mathbf{x}_i), 1)}{G^*(t_i|\mathbf{x}_i)} \\ & \quad + \frac{\mathbb{1}_{t_i > \zeta} \ell(\hat{F}_k(\zeta|\mathbf{x}_i), 0)}{G^*(\zeta|\mathbf{x}_i)} \\ & \quad + \frac{\mathbb{1}_{t_i \leq \zeta, \delta_i \neq 0, \delta_i \neq k} \ell(\hat{F}_k(\zeta|\mathbf{x}_i), 0)}{G^*(t_i|\mathbf{x}_i)} \end{aligned} \quad (38)$$

$\stackrel{\text{def}}{=} \Psi_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}), (t, \delta))$   
 $\stackrel{\text{def}}{=} \Lambda_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}), (t, \delta))$   
 $\stackrel{\text{def}}{=} \Phi_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}), (t, \delta))$

562

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

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

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

563 By summing all of the terms, we obtain:

$$\begin{aligned}
\mathbb{E}_{T^*, C, \Delta | \mathbf{X} = \mathbf{x}} \left[ \mathbf{L}_{k, \zeta} \left( \hat{F}_k(\zeta | \mathbf{x}), (T, \Delta) \right) \right] &= \ell \left( \hat{F}_k(\zeta | \mathbf{x}), 1 \right) \mathbb{P}(T^* \leq \zeta, \Delta = k) \\
&\quad + \ell \left( \hat{F}_k(\zeta | \mathbf{x}), 0 \right) \left( \mathbb{P}(T^* \leq \zeta, \Delta \neq k | \mathbf{X} = \mathbf{x}) + \mathbb{P}(T^* > \zeta | \mathbf{X} = \mathbf{x}) \right)
\end{aligned} \tag{39}$$

564 Meanwhile,

$$\mathbb{P}(\overline{T^* \leq \zeta \cap \Delta = k}) = \mathbb{P}(T^* > \zeta \cup \Delta \neq k) \tag{40}$$

$$= \mathbb{P}(T^* > \zeta) + \mathbb{P}(\Delta \neq k) - \mathbb{P}(T^* > \zeta \cap \Delta \neq k) \tag{41}$$

$$= \mathbb{P}(T^* > \zeta) + \mathbb{P}(\Delta \neq k \cap T^* > \zeta) + \mathbb{P}(\Delta \neq k \cap T^* \leq \zeta) - \mathbb{P}(T^* > \zeta \cap \Delta \neq k) \tag{42}$$

$$= \mathbb{P}(T^* > \zeta) + \mathbb{P}(\Delta \neq k \cap T^* \leq \zeta) \tag{43}$$

565 So, we obtain:

$$\mathbb{E}_{T^*, C, \Delta | \mathbf{X} = \mathbf{x}} \left[ \mathbf{L}_{k, \zeta} \left( \hat{F}_k(\zeta | \mathbf{x}), (T, \Delta) \right) \right] = \ell \left( \hat{F}_k(\zeta | \mathbf{x}), 1 \right) F_k^*(\zeta | \mathbf{x}) + \ell \left( \hat{F}_k(\zeta | \mathbf{x}), 0 \right) (1 - F_k^*(\zeta | \mathbf{x})) \tag{44}$$

566  $\square$

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

*Proof.*

$$\begin{aligned}
\mathbb{E}_{T^*, C, \Delta | \mathbf{X} = \mathbf{x}} \left[ \mathbf{L}_{k, \zeta} \left( \hat{F}_k(\zeta | \mathbf{x}), (T, \Delta) \right) \right] &= \ell \left( \hat{F}_k(\zeta | \mathbf{x}), 1 \right) \mathbb{P}(T^* \leq \zeta, \Delta = k | \mathbf{X} = \mathbf{x}) \\
&\quad + \ell \left( \hat{F}_k(\zeta | \mathbf{x}), 0 \right) \left( \mathbb{P}(T^* \leq \zeta, \Delta \neq k | \mathbf{X} = \mathbf{x}) + \mathbb{P}(T^* > \zeta | \mathbf{X} = \mathbf{x}) \right)
\end{aligned} \tag{45}$$

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

**Definition C.3.**

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

$$F_k^*(\zeta|\mathbf{x}) = \mathbb{P}(T^* \leq \zeta, \Delta = k | \mathbf{X} = \mathbf{x}) = \mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x}) \quad (46)$$

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

$$\mathbb{E}_{T^*, C, \Delta | \mathbf{X} = \mathbf{x}} \left[ \mathbb{L}_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}), (T^*, C, \Delta)) \right] = \mathbb{E}_{T, \Delta | \mathbf{X} = \mathbf{x}} [\mathbb{L}_{k,\zeta}(\hat{p}_{k,\zeta}, (T, \Delta))] \quad (47)$$

$$= \ell(\hat{p}_{k,\zeta}, 0) \mathbb{P}(Y_{k,\zeta} = 0 | \mathbf{X} = \mathbf{x}) + \ell(\hat{p}_{k,\zeta}, 1) \mathbb{P}(Y_{k,\zeta} = 1 | \mathbf{X} = \mathbf{x}) \quad (48)$$

$$= \mathbb{E}_{Y_{k,\zeta}} [\ell(\hat{p}_{k,\zeta}, Y_{k,\zeta}) | \mathbf{X} = \mathbf{x}] \quad (49)$$

$$\leq \mathbb{E}_{Y_{k,\zeta}} [\ell(p_{k,\zeta}, Y_{k,\zeta}) | \mathbf{X} = \mathbf{x}] \quad (50)$$

$$\leq \mathbb{E}_{T^*, C, \Delta | \mathbf{X} = \mathbf{x}} [\mathbb{L}_{k,\zeta}(\mathbb{P}(Y_{k,\zeta} = 1), (T, \Delta))] \quad (51)$$

$$\leq \mathbb{E}[\mathbb{L}_{k,\zeta}(F_k^*(\zeta|\mathbf{x}), (T, \Delta))] \quad (52)$$

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

574

575 So, we obtain that  $\forall \zeta, \forall k \in \llbracket 1, K \rrbracket$ ,  $\mathbb{L}_{k,\zeta}(\hat{F}_k(\zeta|\mathbf{x}), (T, \Delta))$  is a proper scoring rule of  $F_k^*(\zeta|\mathbf{x})$ .  
 576 □

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

579 *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  $\sum_{k=1}^K \mathbb{L}_{k,\zeta}$  is proper on the time horizon  $\zeta$  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:

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

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

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

$$\text{IL}(\hat{F}_1(\cdot|\mathbf{x}), \dots, \hat{F}_K(\cdot|\mathbf{x}), (T, \Delta)) \stackrel{\text{def}}{=} \mathbb{E}_Z \left[ \sum_{k=1}^K \mathbb{L}_{k,Z}(\hat{F}_k(Z|\mathbf{x}), (T, \Delta)) | \mathbf{X} = \mathbf{x} \right] \quad (53)$$

$$= \sum_{k=1}^K \underbrace{\mathbb{E}_Z \left[ \mathbb{L}_{k,Z}(\hat{F}_k(Z|\mathbf{x}), (T, \Delta)) | \mathbf{X} = \mathbf{x} \right]}_{\stackrel{\text{def}}{=} \text{IL}_k(\hat{F}_k(\cdot|\mathbf{x}), (T, \Delta))} \quad (54)$$

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

585 *Proof.* We have already proven that  $L_{k,\zeta} : \mathbb{R} \times \mathcal{D} \rightarrow \mathbb{R}$  is a (strictly) proper scoring rule. Using the  
 586 monotonicity /positivity of the expectation, the result is immediate.

$$\mathbb{E}_{T^*, C, \Delta | \mathbf{X} = \mathbf{x}, Z = \zeta} \left[ \text{IL}_k(\hat{F}_k(\zeta|\mathbf{x}), (T, \Delta)) \right] = \mathbb{E}_{T^*, C, \Delta | \mathbf{X} = \mathbf{x}, Z = \zeta} \left[ \mathbb{L}_k(\hat{F}_k(\zeta|\mathbf{x}), (T, \Delta)) \right] \quad (55)$$

$$\leq \mathbb{E}_{T^*, C, \Delta | \mathbf{X} = \mathbf{x}, Z = \zeta} \left[ \mathbb{L}_k(F_k^*(\zeta|\mathbf{x}), (T, \Delta)) \right] \quad (56)$$

$$\leq \mathbb{E}_{T^*, C, \Delta | \mathbf{X} = \mathbf{x}, Z = \zeta} \left[ \text{IL}_k(F_k^*(\zeta|\mathbf{x}), (T, \Delta)) \right] \quad (57)$$

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

$$\mathbb{E}_{T^*, C, \Delta} [\text{IL}_k(\hat{F}_k(Z|\mathbf{x}), (T, \Delta)) | \mathbf{X} = \mathbf{x}] \leq \mathbb{E}_{T^*, C, \Delta} [\text{IL}_k(F_k^*(Z|\mathbf{x}), (T, \Delta)) | \mathbf{X} = \mathbf{x}] \quad (58)$$

588 This allows us to consider the IL as a global proper scoring rule to compare different competing risks  
589 models.  $\square$

## 590 D The Yanagisawa (2023) scoring rule for survival

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

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

$$\begin{aligned} S_{Cen-log-simple}(\hat{F}, (t, \delta); \{\zeta_i\}_{i=0}^B) \stackrel{\text{def}}{=} & \\ & - \delta \sum_{i=0}^{B-1} \mathbb{1}_{\zeta_i < t \leq \zeta_{i+1}} \log(\hat{F}(\zeta_{i+1}) - \hat{F}(\zeta_i)) \\ & - (1 - \delta) \sum_{i=0}^{B-1} \mathbb{1}_{\zeta_i < t \leq \zeta_{i+1}} \log(1 - \hat{F}(\zeta_{i+1})) \quad (59) \end{aligned}$$

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

## 602 E Pseudo-code

---

### Algorithm 2 IPCW Computer

---

```

Input:  $\mathbf{x}, \delta, t, \hat{G}$ 
 $y \leftarrow \delta \mathbb{1}_{t \leq \zeta}$  ▷Computing the target
 $w \leftarrow 0$ 
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
   $w \leftarrow \frac{1}{\hat{G}(t|\mathbf{x})}$ 
end if
return  $y_i, w_i$ 

```

---

## 603 F Additional results for competing risk experiments

### 604 F.1 Results on synthetic dataset

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

---

<sup>2</sup>We use  $B = 32$ , as in the experiments in Yanagisawa (2023)

---

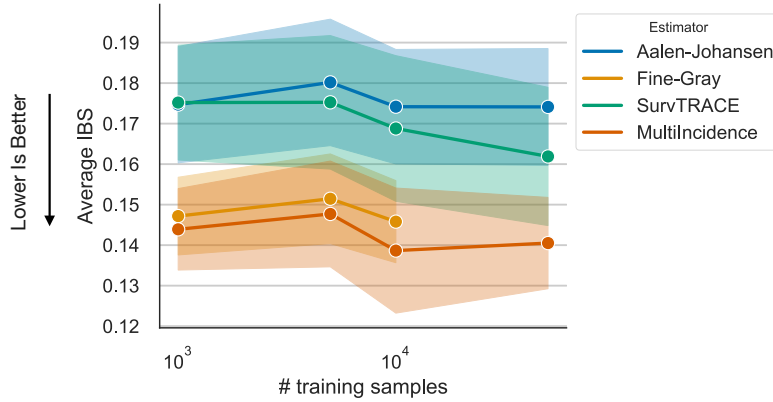
**Algorithm 3** Censoring Feedback Loop - One Iteration
 

---

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

---

**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.



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

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

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

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



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

| EVENT                   | 1             | 2             | 3             |
|-------------------------|---------------|---------------|---------------|
| AALLEN-JOHANSEN         | 0.1209        | 0.2832        | 0.0834        |
| FINE & GRAY             | 0.1055        | 0.0281        | 0.0822        |
| RANDOM SURVIVAL FORESTS | <b>0.0825</b> | 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          | <u>0.0832</u> | <b>0.0273</b> | <b>0.0757</b> |

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

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

## 638 F.2 Results for the SEER Dataset

639 **Learning curves** We ran the experiments while varying the number of training points. In doing so,  
 640 we measured the KM-adjusted Integrated Brier Score for each event. We also average it to have one  
 641 global metric. We see in Figure S7 that our model of the global evaluation metric is quite stable and  
 642 lower than the average Integrated Brier Score on SurvTRACE for any number of training points. We  
 643 expanded the Integrated Brier Score for each event while training on the whole dataset except for the  
 644 Random Survival Forests we trained with 100k data points and Fine and Gray with 10k data points  
 645 because the last two methods could not handle such an amount of data. In Table S2, we compare our  
 646 method with the other models. We see that our model MultiIncidence outperforms the other methods.  
 647 Furthermore, figure 3 shows that the models with the best average IBS are also the fastest to train.

648  **$C_\zeta$ -index** The  $C$ -index measures whether the ranking of the risk of the different samples is in  
 649 agreement with the order of the times in which the event of interest happens (Harrell et al., 1982). It  
 650 is originally a metric for survival settings but is often adapted to competing risks settings where it  
 651 is applied independently to each event (Uno et al., 2011). In such settings, it is biased and does not  
 652 control for the probabilities of the events. However, as it is a popular metric, we have included it in  
 653 our experiments.

654 We give tables below for the  $C_\zeta$ -index toward time for the three events S3. At a fixed time horizon  $\zeta$ ,  
 655 we compute the  $C_\zeta$ -index for each class (corresponding to the ROC-AUC where we handled censored  
 656 observations). The time horizons  $\zeta$  are selected based on the any-event distribution, representing  
 657 quantiles, indicating that at the time corresponding to 0.25, 25% of events have already occurred.  
 658 These results differ from those in the SurvTRACE paper (Wang & Sun, 2022) for two reasons: 1)  
 659 The available code online only implements one of their losses, 2) they treated the SEER dataset with

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

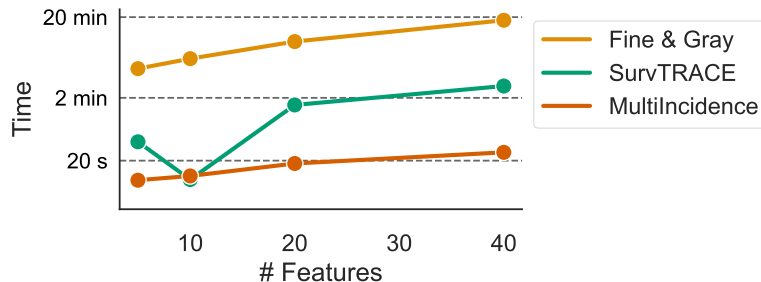


Figure S3: **Integrated Brier Score vs Censoring Rate** Integrated Brier Score for the synthetic dataset with 10,000 training points when we vary the censoring rate. Shaded areas represent the standard deviation across the different seeds. We used the Oracle censoring distribution to compute the weights

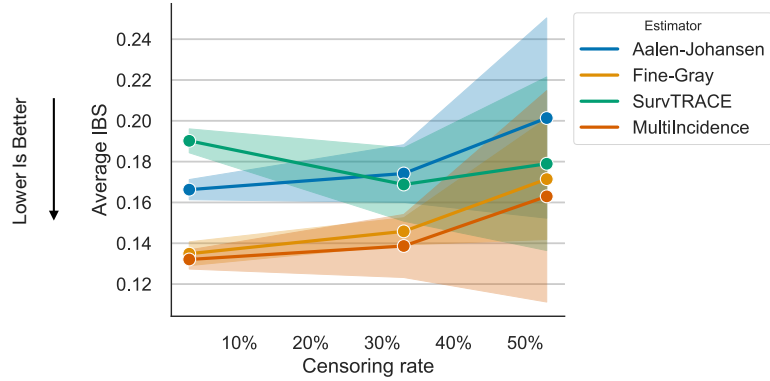


Figure S4: **Brier Score in time** Evolution of the Brier Score for the synthetic dataset I.2 with 20,000 training points with 50% of censoring. The weights are computed with the Oracle censoring distribution.

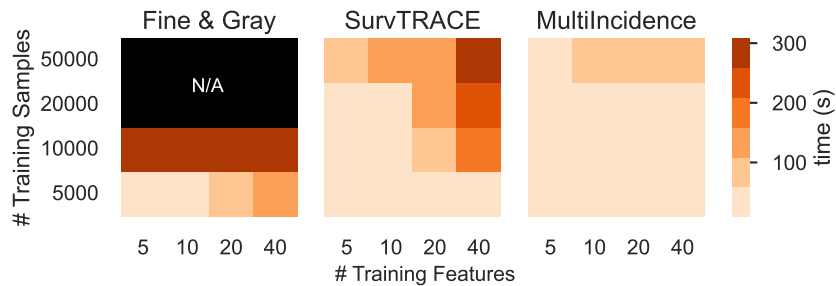
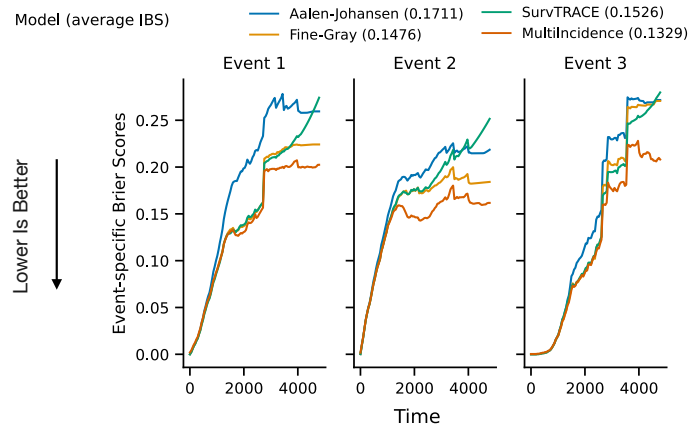


Figure S5: **Fit time for competing risks models.** We have measured the time to fit for each of them depending on the number of training points and the number of features.

660 two competing risks, and any other event was classified as censored, instead of collapsing them in a  
 661 third competing event.

## 662 G Additional results for survival experiments

### 663 G.1 Metrics for the survival analysis

### 664 G.2 Trade-off between training time and performances

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

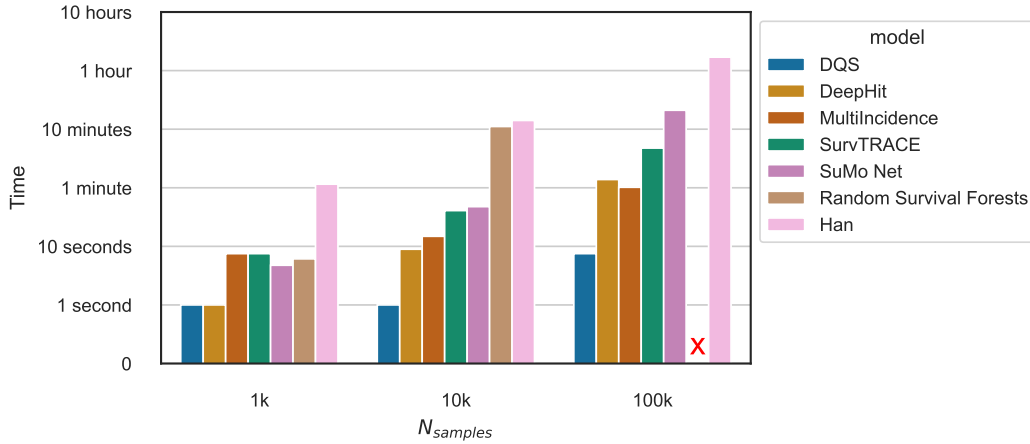


Figure S6: **Synthetic Dataset, training time for survival** Time to fit each survival method while varying the number of samples generated.

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 | <b>0.89</b> | 0.79        | 0.79        | <b>0.87</b> | <b>0.78</b> | <b>0.77</b> | <b>0.85</b> | <b>0.77</b> | <b>0.77</b> |
| DEEPHIT                 | 0.83        | <b>0.86</b> | <b>0.85</b> | 0.75        | 0.75        | <u>0.75</u> | 0.73        | 0.75        | <u>0.75</u> |
| DSM                     | 0.88        | 0.85        | 0.84        | 0.77        | 0.74        | <u>0.75</u> | 0.76        | 0.75        | <u>0.75</u> |
| DESURV                  | 0.83        | 0.82        | 0.81        | 0.72        | 0.70        | 0.71        | 0.74        | 0.73        | 0.73        |
| SURVTRACE               | <u>0.88</u> | 0.78        | 0.77        | <u>0.86</u> | <u>0.76</u> | <u>0.75</u> | <u>0.84</u> | <u>0.76</u> | <u>0.75</u> |
| MULTIINCIDENCE          | <u>0.88</u> | 0.79        | 0.77        | 0.85        | 0.72        | 0.71        | 0.81        | 0.66        | 0.62        |

## 667 H Implementation Details

### 668 H.1 Reference of used implementations for baselines

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

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.

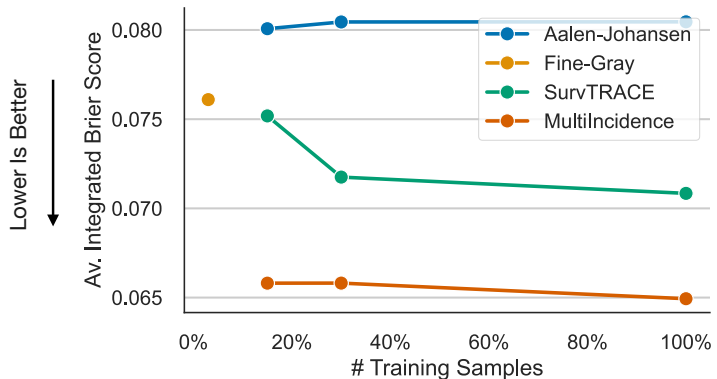


Table S4: METABRIC: Integrated Brier Score,  $S_{Cen-log-simple}$  and c-index at 50%

| 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        | <b>0.639±0.024</b> | 0.613±0.016        | 0.180±0.014        | 2.271±0.019          |
| PCHAZARD                | 0.595±0.088        | <b>0.639±0.019</b> | <b>0.639±0.014</b> | 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        | <u>2.205±0.044</u>   |
| SUMO NET                | <b>0.660±0.022</b> | 0.634±0.017        | 0.589±0.015        | <u>0.169±0.009</u> | <u>2.302±0.059</u>   |
| SURVTRACE               | 0.589±0.082        | 0.627±0.015        | 0.629±0.007        | <b>0.168±0.011</b> | 2.270±0.034          |
| MULTIINCIDENCE          | <u>0.627±0.016</u> | <u>0.636±0.015</u> | <u>0.635±0.011</u> | <b>0.168±0.019</b> | <b>2.169±0.056</b>   |

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        | <u>0.609±0.004</u> | 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                     | <b>0.601±0.019</b> | 0.598±0.012        | 0.592±0.009        | 0.201±0.007        | 1.987±0.069          |
| SUMO NET                | <u>0.590±0.016</u> | 0.589±0.016        | 0.589±0.015        | <u>0.194±0.010</u> | <b>1.721±0.016</b>   |
| SURVTRACE               | 0.578±0.008        | <u>0.609±0.005</u> | <u>0.610±0.006</u> | <u>0.194±0.005</u> | 1.870±0.018          |
| MULTIINCIDENCE          | 0.572±0.019        | <b>0.618±0.007</b> | <b>0.615±0.007</b> | <b>0.191±0.006</b> | <u>1.740±0.020</u>   |

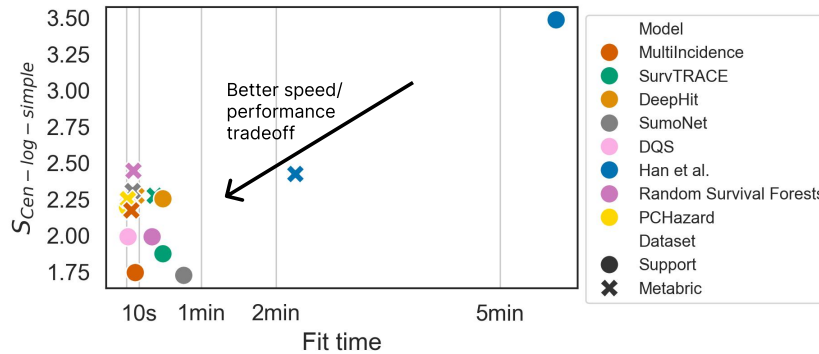


Figure S8: Trade-off between the performances and the training time for the  $S_{Cen-log-simple}$  for the survival model over METABRIC and SUPPORT

## 671 H.2 GridSearch Parameters

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

## 674 I Distribution of the competing risks datasets

### 675 I.1 SEER Distribution of events

676 Here, we present the distributions of the event of the SEER Dataset. We can highlight that the  
677 censoring distribution is non-uniform in time. The change in the censoring distribution from the 48<sup>th</sup>  
678 month may be hard to learn for some methods.

Table S6: Characteristics of used baselines.

| Name                    | Competing risks | Proper loss | Implementation   | Reference                    |
|-------------------------|-----------------|-------------|--|------------------------------|
| SurvTRACE               | ✓               |             | ours   | Wang & Sun (2022)            |
| DeepHit                 | ✓               |             | <a href="https://github.com/havakv/pycox">github.com/havakv/pycox</a>  | Lee et al. (2018)            |
| DSM                     | ✓               |             | <a href="https://autonlab.github.io/DeepSurvivalMachines">autonlab.github.io/DeepSurvivalMachines</a>  | Nagpal et al. (2021)         |
| DeSurv                  | ✓               |             | <a href="https://github.com/djdanks/DeSurv">github.com/djdanks/DeSurv</a>  | Danks & Yau (2022a)          |
| Random Survival Forests | ✓               |             | <a href="https://scikit-survival.readthedocs.io/">scikit-survival.readthedocs.io/</a> for survival, and <a href="http://www.randomforests.org/">www.randomforests.org/</a> for competing risks | Ishwaran et al. (2008, 2014) |
| Fine & Gray             | ✓               |             | <a href="https://cran.r-project.org/package=cprsk">cran.r-project.org/package=cprsk</a>  | Fine & Gray (1999)           |
| Aalen-Johansen          | ✓               |             | ours   | Aalen et al. (2008)          |
| Han et al.              |                 |             | <a href="https://github.com/rajesh-lab/Inverse-Weighted-Survival-Games">github.com/rajesh-lab/Inverse-Weighted-Survival-Games</a>  | Han et al. (2021)            |
| PCHazard                |                 |             | <a href="https://github.com/havakv/pycox">github.com/havakv/pycox</a>  | Kvamme & Borgan (2019b)      |
| SumoNet                 |                 | ✓           | <a href="https://github.com/MrHuff/Sumo-Net">github.com/MrHuff/Sumo-Net</a>  | Rindt et al. (2022)          |
| DQS                     |                 | ✓           | <a href="https://ibm.github.io/dqs/">ibm.github.io/dqs/</a>  | 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 | $\{2, 3\}$                      |

679 **I.2 Example of distribution of one synthetic dataset**

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

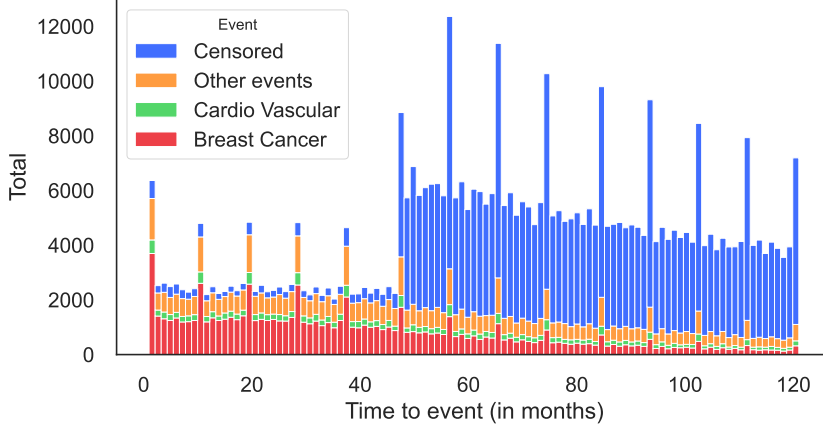


Figure S9: **SEER Dataset Distributions** The censoring rate is around 63%. The prevalence of the different events is 18% for Breast Cancer, 4.5% for Cardio Vascular events, and 10% for other events.

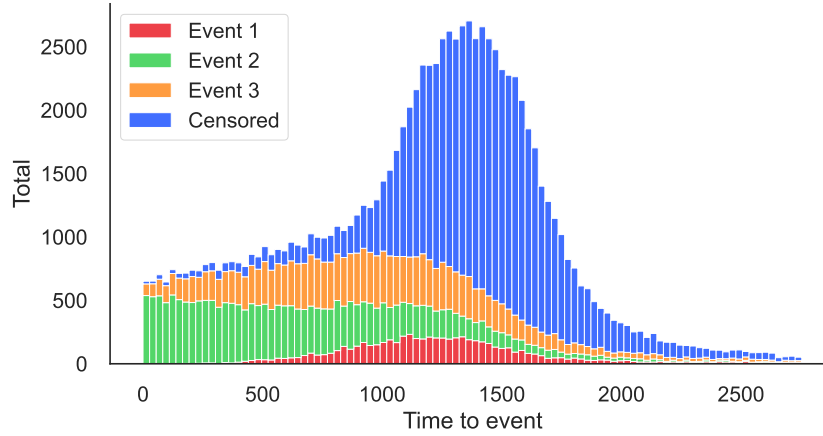


Figure S10: **Synthetic Dataset Distributions** Duration distributions of the synthetic dataset when censoring is dependent on  $X$ , censoring rate 69%. The events are stacked.

## 685 J Corollary: Bregman divergence

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

688 **Definition J.1.** Considering  $U : \mathbb{R}^d \rightarrow \mathbb{R}$  strictly convex and differentiable,

$$\text{Bregman divergence} \quad D_U(p, q) = U(p) - U(q) - \langle \nabla U(q), p - q \rangle. \geq 0 \quad (60)$$

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

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

690 Meanwhile, because  $U$  is strictly convex and differentiable:

$$\forall p, \hat{p}, \quad U(p) > U(\hat{p}) + \langle \nabla U(\hat{p}), p - \hat{p} \rangle \quad (61)$$

$$-U(\hat{p}) + \langle \nabla U(\hat{p}), \hat{p} - p \rangle > -U(p) \quad (62)$$

691 This implies:

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

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

## 693 K Examples

### 694 K.1 Brier Score

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

**Definition K.1.**

$$\begin{aligned}
\forall \zeta, \forall k \in [1, K], \\
\text{BS}_k(\hat{F}_k(\zeta, \mathbf{x}), \delta, t, \zeta, \mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}_{t_i \leq \zeta, \delta_i = k} \left( 1 - \hat{F}_k(\zeta|\mathbf{x}_i) \right)^2}{G^*(t_i|\mathbf{x}_i)} + \frac{\mathbb{1}_{t_i > \zeta} \left( \hat{F}_k(\zeta|\mathbf{x}_i) \right)^2}{G^*(\zeta|\mathbf{x}_i)} \\
+ \frac{\mathbb{1}_{t_i \leq \zeta, \delta_i \neq 0, \delta_i \neq k} \left( \hat{F}_k(\zeta|\mathbf{x}_i) \right)^2}{G^*(t_i|\mathbf{x}_i)} \quad (63)
\end{aligned}$$

### 697 K.2 Binary cross entropy loss

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

$$\begin{aligned}
\forall k \in [1, K], \\
l_k(\hat{F}_k(\zeta, \mathbf{x}), \delta, t, \zeta) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}_{t_i \leq \zeta, \delta_i = k} \log \left( \hat{F}_k(\zeta|\mathbf{x}_i) \right)}{G^*(t_i|\mathbf{x}_i)} + \frac{\mathbb{1}_{t_i \leq \zeta, \delta_i \neq 0, \delta_i \neq k} \log \left( 1 - \hat{F}_k(\zeta|\mathbf{x}_i) \right)}{G^*(t_i|\mathbf{x}_i)} \\
+ \frac{\mathbb{1}_{t_i > \zeta} \log \left( 1 - \hat{F}_k(\zeta|\mathbf{x}_i) \right)}{G^*(\zeta|\mathbf{x}_i)} \quad (64)
\end{aligned}$$