On the Calibration of Survival Models with Competing Risks

Julie Alberge
Inria de Saclay

Tristan Haugomat
Inria de Saclay, Drees

Gaël Varoquaux
Inria de Saclay, :probabl.

Judith Abécassis
Inria de Saclay

Abstract

Survival analysis deals with modeling the time until an event occurs, and accurate probability estimates are crucial for decision-making, particularly in the competing-risks setting where multiple events are possible. While recent work has addressed calibration in standard survival analysis, the competing-risks setting remains under-explored as it is harder (the calibration applies to both probabilities across classes and time horizon). We show that existing calibration measures are not suited to the competing-risk setting and that recent models do not give well-behaved probabilities. To address this, we introduce a dedicated framework with two novel calibration measures that are minimized for oracle estimators (*i.e.*, both measures are proper). We also introduce some methods to estimate, test, and correct the calibration. Our recalibration methods yield better probabilities while preserving discrimination.

1 INTRODUCTION: *FAITHFUL* PROBABILITIES MATTER

Is my patient more likely to die of a heart attack or breast cancer in the coming year? In various application domains, including health or marketing, accurate probabilities are crucial. Competing risks analysis focuses on modeling several mutually exclusive *time-to-event outcomes*. More specifically, one wants to predict which of the events of interest will happen first, and when. An application could be distinguishing between different causes of death. One challenge is right censoring: certain events are not observed. Practically, competing risks methods estimate the cumulative distribution functions of each event of interest for each individual, for example, the probabilities of dying from each cause at any time for a given patient. Initially, simple estimators were proposed, such as Aalen and Johansen (1978) (see Definition in Appendix A.4), a marginally consistent estimator that returns the incidence function for each event type for the entire population but does not predict individualized cumulative incidence functions. Recent research has developed more and more complex models using covariates to individualize the predictions of the cumulative incidence functions for each patient (Fine and Gray, 1999; Ishwaran et al., 2008; Lee et al., 2018; Wang and Sun, 2022). Those models estimate conditional probabilities that can be integrated over the whole population to obtain a marginal prediction. Some of these models, in particular, based on machine learning approaches like forests (Ishwaran et al., 2008) or deep neural networks (Lee et al., 2018; Wang and Sun, 2022) exhibit excellent discriminative performance, but their calibration is not studied.

Indeed, while discrimination between patients is useful, probabilities are key to decision making (Van Calster et al., 2019; Perez-Lebel et al., 2025), *e.g.*, to arbitrage between treatments depending on the probabilities of different causes of death. But to be useful, these probabilities must be *faithful* in some sense. *Calibration* is one way to assess their faithfulness, comparing the average individualized predicted probabilities with population-level predictions from a marginally consistent estimator (*e.g.* Aalen and Johansen, 1978). On the METABRIC dataset, following various causes of death (Rueda et al., 2019), this comparison reveals that existing methods produce predictions that, on average, deviate from the marginally consistent estimator, suggesting poor calibration (Figure 1).

For survival analysis, a simpler setting with a single event of interest, different measures of calibration have recently been studied (Haider et al., 2020; Qi et al., 2024). 1-calibration focuses on predictions at a given time point to recover the usual binary setting (Haider et al., 2020); it can be extended to time-to-event regression by integrating over time, leading to the KM-calibration (Qi et al., 2024). D-calibration focuses on the properties of the cumulative distribution function

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

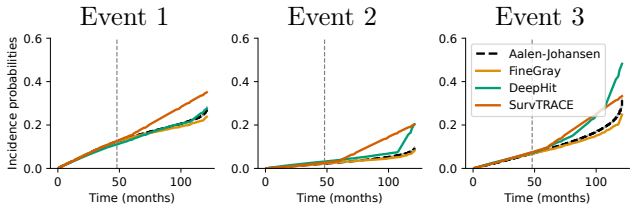

Figure 1: **Marginal probabilities of SOTA models do not behave as expected:** We compare the marginal probabilities of the different SOTA models on the SEER Dataset with 100k training samples. All these marginals should be close to the Aalen and Johansen (1978) estimator, which is marginally consistent.

Table 1: **Inadequacy of survival-only calibration scores for competing risks models.** On METABRIC (10 seeds), survival calibration metrics fail to capture the consistency of the Aalen and Johansen (1978) estimator. While D-calibration (Haider et al., 2020) suggests calibration for event 1 (death from breast cancer) but not event 2 (death from other causes), KM-calibration (Qi et al., 2024) incorrectly indicates poor calibration for both.

Event	D-calibration (mean±std)	KM-calibration (mean±std)
1 (breast-cancer)	1.0±.0	.0012±.0011
2 (other)	0.2±.4	.0393±.0113

of the survival function (Haider et al., 2020).
Survival-analysis calibration cannot be directly applied to competing-risks setting (see Figure 5 and Table 1); the multiclass aspect needs to be taken into account. Practitioners in need of ad-hoc solutions have extended 1-calibration to calibration plots (checking the calibration of different classes at one given time point) (Huang et al., 2020). But work is needed to define an actually *proper* metric that addresses both the multiple classes and the whole period of prediction. The challenge is indeed integrating in both directions (events and time), as illustrated in Figure 2.
This paper defines proper measures of calibration for competing risks, as well as estimators for those measures, by combining the notion of multi-class calibration and survival analysis calibration.

Contributions After highlighting that state-of-the-art (SOTA) competing-risks methods do not return marginal meaningful probabilities, we contribute a full calibration framework for competing risks:

- We extend D-calibration to competing-risks settings, proposing the CR D-calibration. We prove that this metric is proper: the oracle functions are CR D-calibrated. We also prove that any marginally consistent estimator is asymptotically CR D-calibrated, assuming the strict monotonicity of the cumulative incidence functions. We propose a consistent estimator of this metric and a statistical test.

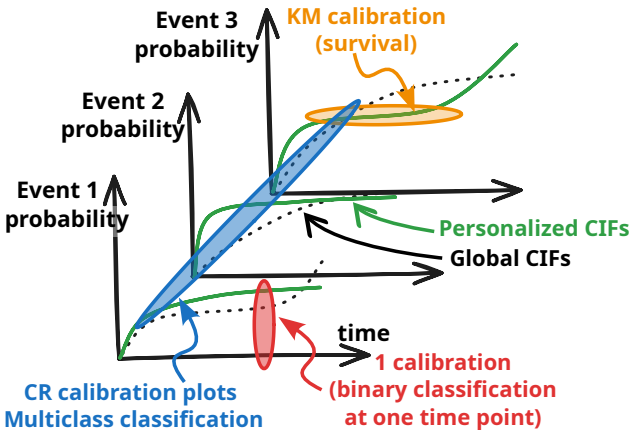

Figure 2: **Existing calibration measures cast to the competing-risks setting:** The CIF (Cumulative Incidence Function) for an individual specifies, for each event type, the probability of experiencing that event before time τ. In this space, different existing calibration measures correspond to projecting the vector-valued CIF onto a specific "direction": 1-calibration (binary calibration at a fixed time), KM-calibration (survival/Kaplan–Meier calibration), multiclass CR calibration plots, or calibration of global (population-level) CIFs. Each of these captures only one marginal aspect of the problem and fails to assess the joint calibration of the full CIF vector across event types.

- We introduce the cal_K^α-calibration metric that measures the difference between the marginal oracle function at any time point τ (*i.e.* $F_k(\tau)$) with the probabilities obtained by the estimator integrated over the samples. We prove that this metric well behaved (being equals to 0 for the oracle functions and calibrated ones), give a consistent estimator and propose a test.

- We propose two post-hoc re-calibration methods for competing risks models. We show that they maintain the same discrimination (*i.e.*, unchanged C-index), while also making models more calibrated. We study the impact of the recalibration on other competing risks performance metrics.

2 BACKGROUND

2.1 Problem formulation: competing-risks setting

We consider $K \in \mathbb{N}^*$ mutually exclusive competing events. For $k \in [\![1, K]\!]$, we denote $T_k^* \in \mathbb{R}_+$ the event time of the event k, depending on the covariates $\mathbf{X}$. We also denote $T^* \in \mathbb{R}_+$, the time of the first event that occurs, $T^* = \min_{k \in [\![1,K]\!]}(T_k^*)$ and $\Delta^* = \arg\min_{k \in [\![1,K]\!]}(T_k^*)$. We observe $(\mathbf{X}, T, \Delta) \sim \mu$, where μ denotes the distribution followed by $(\mathbf{X}, T, \Delta)$, with $T = \min(T^*, C)$ where $C \in \mathbb{R}_+$ is the censoring time, which can depend on $\mathbf{X}$, and $\Delta \in [\![0, K]\!], \Delta = \arg\min_{k \in [\![0,K]\!]}(T_k^*)$, where 0 denotes a censored observation. However, we are primarily interested in the distribution of the uncensored data, $(\mathbf{X}, T^*, \Delta^*) \sim \mu^*$, particularly the conditional distribution of $T^*, \Delta^* | \mathbf{X} = \mathbf{x}$. We index the observations by i, and $\mathbf{x}_i$ denotes the covariates associated with the ith observation. The outcome is represented by (t_i, δ_i), where t_i is the observed time, and $\delta_i \in [\![0, K]\!]$ is the event indicator. For $k \in [\![1, K]\!]$, $\delta_i = k$ indicates that the event of interest k was observed at time t_i, while $\delta_i = 0$ indicates that the observation was censored at time t_i. We divide our data set $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$ into a training set $\mathcal{D}_{train}$, a calibration data set $\mathcal{D}_{cal}$, and $\mathcal{D}_{test}$. We introduce the different quantities estimated by competing risks models:

Definition 2.1 (*Quantities of interest*)**.**

CIF (Cumulative Incidence Function):

$$F^*(\tau|\mathbf{x}) = \mathbb{P}(T^* \leq \tau | \mathbf{X} = \mathbf{x})$$

CIF of the kth event:

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

Survival Function to any event:

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

In the remainder of this work, we will use the following assumptions:

Assumption 2.1 (*Continuity of the CIFs and Survival Function*)**.** We assume that each cumulative incidence function F_k and the survival function are continuous.

Assumption 2.2 (*Non-informative censoring*)**.** We make the classic assumption in survival analysis or in the competing-risks setting that the censoring is non-informative with respect to covariates:

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

Assumption 2.3 (*Independence*)**.** We suppose that the samples are drawn i.i.d.

Assumption 2.4 (*Each event may occur eventually*)**.** For each event of interest, we make the minor assumption that the event may happen to each individual one day, *i.e.* $\mathbb{P}(\Delta^* = k|\mathbf{X}) \neq 0$.

While in survival analysis, one assumes that the event will occur one day, *i.e.* $\mathbb{P}(T^* < \infty|\mathbf{X}) = 1$, in the competing-risks setting, we make an equivalent assumption here, which will only be necessary in the definition of the CR-D-calibration (Def. 3.1):

Assumption 2.5 (*One event will happen a.s.*)**.** For all individuals, we assume that one event will happen almost surely one day, for a sufficiently long time, *i.e.* $\sum_{k=1}^{K} \mathbb{P}(T^* < \infty, \Delta^* = k|\mathbf{X}) = 1$.

2.2 Related work

The importance of competing risks The competing-risks setting accounts for multiple, mutually exclusive event types in time-to-event studies. For example, consider estimating the number of breast cancer-related deaths in a population. Focusing on one possible cause of death while ignoring other potential causes leads to biased estimates (Gorfine and Hsu, 2011; Wolbers et al., 2009).While the competing-risks setting is more complex than typical survival analysis, various approaches exist to tackle this scenario (Aalen and Johansen, 1978; Fine and Gray, 1999; Lee et al., 2018; Van Calster et al., 2019; Nagpal et al., 2021; Alberge et al., 2025).

Calibration metrics in survival analysis settings Introduced by Haider et al. (2020), D-calibration in the survival analysis setting is based on a property of the cumulative distribution function. They assume that the event of interest will happen almost surely, *i.e.* $\forall x, F(\infty|\mathbf{x}) = \mathbb{P}(T^* < \infty \mid \mathbf{X} = \mathbf{x}) = 1$, where F is the cumulative distribution function of T^*. By the probability integral transform (David and Johnson, 1948; Wikipedia, 1984), it follows that $F(T^*) \sim \mathcal{U}(0, 1)$ (as detailed in Casella and Berger (2002). Haider et al. (2020) extend this idea and incorporate censored individuals, given that the patients have survived up to their censoring time. A formal definition of the D-calibration in the survival analysis setting can be found in Appendix A.1. The other main calibration metric in survival analysis is the 1-calibration (Huang et al., 2020), which evaluates the model calibration at a fixed time point by considering the problem as a binary classification task. It comes with its extension: the Kaplan and Meier (1958) calibration (Chapfuwa et al., 2023).

Calibration in multi-class settings In classification without censoring, multi-class calibration can be characterized in several ways, some stronger than others (Kull et al., 2019; Perez-Lebel et al., 2023; Vaicenavicius et al., 2019). Top-label calibration (Guo et al., 2017)

considers that a model is calibrated if the highest predicted class corresponds to the actual class. Class-wise calibration, more stringent, controls the probability of each class (Zadrozny and Elkan, 2002). Finally, joint calibration evaluates the calibration of the entire prediction vector across all classes simultaneously (Kull et al., 2019). Post-hoc methods can recalibrate a multiclass classifier (Filho et al., 2023). Isotonic regression (Zadrozny and Elkan, 2002) is a classic technique, but it does not control for multiclass probabilities. These are often cobbled together using one-versus-rest (Berta et al., 2025; Scikit-learn, 2025). Temperature scaling, using a softmax parameterization, is better behaved (Guo et al., 2017).

Calibration for competing risks models The notion of calibration for competing risks lacks theoretical grounding. Nevertheless, as Wolbers et al. (2009) show in their review, most works tackle it using calibration plots at a given fixed time point using pseudo-observations to handle censoring, or with a marginal estimator (*e.g.* Aalen and Johansen, 1978; Zhang et al., 2018). The calibration plot can also be constructed using more complex methods (Austin et al., 2022; Gerds et al., 2014). Booth et al. (2023) used shrinkage to recalibrate competing risks methods as a post-hoc operation. But, to our knowledge, there is no real-valued global calibration metric in the competing-risks setting that has been formally defined and that addresses the multiple dimensions of the problem together, as illustrated graphically in Figure 2.

3 MEASURING CALIBRATION IN COMPETING-RISKS SETTINGS

3.1 Distribution calibration in the competing-risks setting (CR D-cal)

Without censoring For now, we suppose there are no censored individuals. Importantly, the D-calibration for survival analysis introduced in section 2.2 does not easily carry over to the competing-risks setting. Indeed, in the competing-risks setting, a patient does not have a 100% chance of contracting one particular event and not the other (Assumption 2.4). As a result, the key property for D-calibration in survival analysis *i.e.* $\forall x, F(\infty|\mathbf{x}) = \mathbb{P}(T^* < \infty \mid \mathbf{X} = \mathbf{x}) = 1$ does not extend to competing risks: for each event k, $F_k(T^*, \mathbf{X}) \nsim \mathcal{U}(0,1)$ because $F_k(\infty|\mathbf{X}) = \mathbb{P}(T^* < \infty, \Delta^* = k \mid \mathbf{X} = \mathbf{x}) = \mathbb{P}(\Delta^* = k|\mathbf{X} = \mathbf{x}) \neq 1$. But as a direct extension, we see almost surely that $\forall k \in [\![1, K]\!]$, $F_k(T^*|\Delta^* = k, \mathbf{X}) \sim \mathcal{U}(0, \mathbb{P}(\Delta^* = k|\mathbf{X}))$. However, this uniformity cannot be generalized to the population level, but an easier way to see this quantity is to take the ratio $F_k(T^*|\mathbf{X})/F_k(\infty|\mathbf{X})$ which simply follows a uniform distribution between 0 and 1, *i.e.*, $(F_k(T^*|\mathbf{X})/F_k(\infty|\mathbf{X})|\Delta^* = k, \mathbf{X}) \sim \mathcal{U}(0,1)$ (property of a cumulative distribution function, see Lemma C.1, Appendix C). Conditioning on the event of interest and taking the ratio of the cumulative incidence functions allows us to use the probability integral transform and define the CR-D-calibration in the case where there is no censoring (illustrated on fig. 3). In the following, we will write $F_{k/\infty}(t \mid \mathbf{x}) \stackrel{\text{def}}{=} \frac{F_k(t|\mathbf{x})}{F_k(\infty|\mathbf{x})}$.

Incorporating the censoring mechanism Competing risks are concerned with censored individuals. Incorporating censored samples is more subtle than in the single-event case. For a censored individual, the potential event type is unobserved, and the subject will experience *exactly one* of the competing events. Hence, unlike standard survival analysis, we cannot directly impute the conditional distribution of T^* given $\mathbf{X}$ and $\Delta = 0$, as the terminal event remains unknown. Nonetheless, under the assumption that each individual will experience an event (*i.e.*, $\sum_k F_k(\infty \mid \mathbf{X}) = 1$) (Assumption 2.5), and that the individual has survived up to the censoring time C, we define the competing risks D-calibration (CR D-calibration) by assigning, for each cause k, the distribution of $\mathbb{P}(T^* \leq \tau, \Delta^* = k|\mathbf{X}, T^* > C) = \mathbb{P}(C \leq T^* \leq \tau, \Delta^* | \mathbb{X})/\mathbb{P}(T^* > C)$ (Definition 3.1) hence attributing a contribution of each censored observation to all possible events following their relative probability distribution. This construction enables accounting for censored observations by distributing each event-specific risk beyond C in a way that respects the model's marginal predictions. With this added term, we recover a uniform distribution (see Theorem 1). To compute the CR-D-calibration, we construct buckets (to compare them with those from

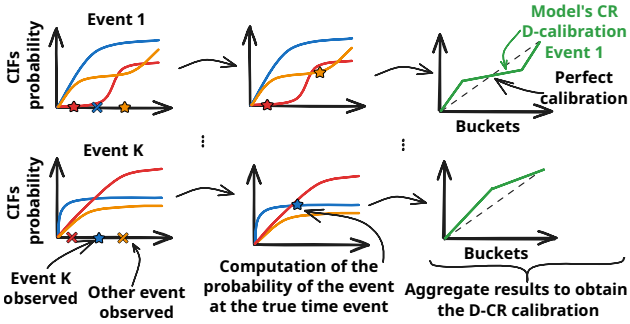

Figure 3: **CR D-calibration without censoring:** For each event, we compute the individualized CIFs. Then, for each observed individual $\mathbf{x}_i$ with event k, we compute $F_k(t_i|\mathbf{x}_i)$ and $F_k(\infty|\mathbf{x}_i)$. Following theorem 1, we compare the $F_k(t_i|\mathbf{x}_i)/F_k(\infty|\mathbf{x}_i)$ cumulative empirical distribution to an uniform cumulative distribution function. Summing this metric over all event types the results gives a measure of the CR D-calibration.

a uniform distribution), $B_{[0,\rho]}$ with $\rho \in [0,1]$. For a given bucket, we compute two terms. In the first one, we measure the patients who have experienced the event of interest $\mathbb{1}_{\hat{F}_{k/\infty}(T|\mathbf{X})\in[0,\rho],\ \Delta=k}$. The second term represents the patient that have been censored ($\mathbb{1}_{C\leq T^*,\ \hat{F}_{k/\infty}(C|\mathbf{X})\in[0,\rho]}$), this adjustment term is based on the expected value given their known censoring time: $\hat{F}_k(\infty|\mathbf{X})\rho - \hat{F}_k(C|\mathbf{X}) / \hat{S}(C|\mathbf{X})$.

Definition 3.1 (CR D-calibration)**.** For any estimator of the CIFs, we measure the following quantity, for a given $\alpha \in \mathbb{R}_+, \alpha > 1$ where α represents the α-norm. The numerator estimates the measure of the bucket, and the denominator normalizes by the total predicted mass of event (since some individuals may never experience k).

$$\text{CR D-cal} \quad D_\alpha^{CR} = \sum_{k=1}^{K} \left(\int_0^1 \Big| \underbrace{\frac{\mathbb{E}_{T,\Delta}(B^k_{[0,\rho]})}{\mathbb{E}_X(\hat{F}_k(\infty|X))}}_{\stackrel{\text{def}}{=} b^k_{[0,\rho]}} - \rho \Big|^\alpha d\rho \right)^{\frac{1}{\alpha}}$$

where $\forall \rho \in [0,1]$

Observed samples

Probability after censoring

$$B^k_{[0,\rho]} = \mathbb{1}_{\hat{F}_{k/\infty}(T|\mathbf{X})\in[0,\rho],\ \Delta=k} + \frac{\hat{F}_k(\infty|\mathbf{X})\rho - \hat{F}_k(C|\mathbf{X})}{\hat{S}(C|\mathbf{X})} \mathbb{1}_{C\leq T^*, \hat{F}_{k/\infty}(C|\mathbf{X})\in[0,\rho]}$$

A model is perfectly CR D-calibrated when $D_\alpha^{CR} = 0$which implies that the empirical distribution corresponds to the expected one. We can also define the intervals differently, between two reals $0 \leq a < b \leq 1$ instead of $[0,\rho]$, $B^k_{[a,b]} = B^k_{[0,b]} - B^k_{[0,a]}$. This will add a new term to the construction of the interval $B_{[a,b]}$ detailed in Appendix D.2 (eq. 31).

Theorem 1 (Properness)**.** *Under assumptions 2.1, 2.2, 2.3, 2.4, the expectation of the bucket for the oracle function can be computed as:* $\mathbb{E}(B_{[a,b]} \mid \mathbf{X}) = (b-a)F_k(\infty \mid \mathbf{X})$. *This implies that the oracle functions are CR D-calibrated (Def. 3.1)* - i.e., $\forall \alpha \in \mathbb{R}^*$, D_α^{CR} *(eq. 1) is minimized for the oracle CIFs.*
Under the same assumptions, a marginally consistent estimator is asymptotically CR D-calibrated.

Proof sketch. We start by proving the stated expectation value (Lemma D.2) by studying the oracle function in those cases: if the patient is censored or if the event has been observed. Then, the result is straightforward with the Lemma D.2. With Lebesgue's dominated convergence theorem, we show that a marginally consistent estimator is asymptotically CR D-calibrated. For more details, see Appendix D.2. □

Corollary 3.1. *As the Aalen-Johansen estimator is a consistent (see Lemma D.1) marginal estimator of the oracle functions, it is also asymptotically CR D-calibrated.*

Estimator of the CR D-calibration Extending Definition 3.1 to the empirical measure, for a given population and a given model, we define the estimation of the CR D-calibration on given dataset (named $\hat{D}_\alpha^{CR}$).

Definition 3.2 (CR $\hat{D}_\alpha^{CR}$-calibration)**.** Thus, we define, for a fixed $\rho \in [0,1]$, and event type k:

$$\hat{b}^k_{[0,\rho]} = \frac{\text{card}\left\{ i \;\middle|\; \delta_i = k, \hat{F}_{k/\infty}(t_i \mid \mathbf{x}_i) \in [0,\rho] \right\}}{\sum_i \{\hat{F}_k(\infty \mid \mathbf{x}_i)\}} + \frac{\sum_i \left\{ \frac{\hat{F}_k(\infty|\mathbf{x}_i)\rho - \hat{F}_k(t_i|\mathbf{x}_i)}{\hat{S}(t_i|\mathbf{x}_i)} \;\middle|\; \begin{array}{l} \delta_i = 0 \\ \hat{F}_{k/\infty}(t_i \mid \mathbf{x}_i) \in [0,\rho]) \end{array} \right\}}{\sum_i \{\hat{F}_k(\infty \mid \mathbf{x}_i)\}}$$

Where the first term counts individuals who experienced event k and whose normalized prediction falls in $[0,\rho]$, scaled by the total predicted risk. The second term accounts for censored patients, estimating their contribution. Then, we can define an estimator of the CR D_α^{CR}-calibration as:

$$\hat{D}_\alpha^{CR}\text{-cal estimator} \quad \hat{D}_\alpha^{CR} \stackrel{\text{def}}{=} \sum_{k=1}^{K} \left(\int_0^1 \left| \hat{b}^k_{[0,\rho]} - \rho \right|^\alpha d\rho \right)^{\frac{1}{\alpha}}$$

Proposition 3.1 (Consistency of the estimator)**.** *As the sample size (m) of the calibration set $\mathcal{D}_{cal}$ tends toward infinity (*i.e. $m \to \infty$*), the estimator CR $\hat{D}$-calibration (Def. 3.2) converges almost surely toward CR D-calibration (Def. 3.1), making it a consistent estimator of the CR D-calibration.*

Proof Sketch. The proof relies primarily on the strong law of large numbers, which ensures that the consistent estimator converges almost surely to its expected value, the CR-D-calibration. □

3.2 cal_K^α-calibration, another calibration metric using marginal probabilities

While CR D-calibration focuses on the properties of the cumulative distribution function, we can define calibration in another way. As illustrated in Figure 1, a desirable property of a model is to have globally faithful probabilities (*i.e.*, $\forall \tau \in \mathbb{R}_+$, $\forall k \in [\![1, K]\!]$, $\hat{F}_k(\tau) \approx F_k(\tau)$). In other words, the predicted cumulative incidence functions $\hat{F}_k(\tau)$ should approximate the oracle marginal function $F_k(\tau)$. The following can be seen as the extension of the 1-calibration (Haider et al., 2020) in survival analysis to the multi-class setting. Given the presence of multiple competing events, we adopt the terminology of multiclass calibration (as defined in Perez-Lebel et al., 2023). We introduce three notions of calibration in the competing-risks setting, focusing on the marginal probabilities, ordered from the weakest to the strongest one:

Definition 3.3 (cal_K^α-calibration(s))**.** Given a time τ and an event k, we introduce the $\text{cal}_k(\tau)$ as:

$$\text{cal}_k(\tau)\text{-calibration:} \quad \text{cal}_k(\tau) = \left| F_k(\tau) - \mathbb{E}_{\mathbf{X}}(\hat{F}_k(\tau|\mathbf{X})) \right|$$

With the previous definition, we define the cal_k^α-calibration, that coincides with classwise-calibration in multiclass settings (Zadrozny and Elkan, 2002):

cal_k^α-calibration: (Classwise-Calibration)

$$\text{cal}_k^\alpha = \left(\int_0^{t_{max}} \text{cal}_k(\tau)^\alpha \mathrm{d}\tau \right)^{1/\alpha}$$

Extending the sum over the different events, we obtain a jointly calibrated measure (Kull et al., 2019; Vaicenavicius et al., 2019):

$$\text{cal}_K^\alpha\text{-calibration: (Jointly Calibration)} \quad \text{cal}_K^\alpha = \sum_{k=1}^{K} \text{cal}_k^\alpha$$

A model is perfectly cal_K^α-calibrated (resp. cal_k^α-calibrated, $\text{cal}_k(\tau)$-calibrated) if $\text{cal}_K^\alpha = 0$ (resp. $\text{cal}_k^\alpha = 0$, $\text{cal}_k(\tau) = 0$). Assessing the calibration should be aligned with the specific research question. When the focus is on understanding the behavior of the model at a specific time point τ, the $\text{cal}_k(\tau)$-calibration will be most useful. To evaluate the calibration of the model for a particular, potentially rare, risk event over time, the cal_k^α-calibration offers a better perspective. For a global assessment of a model's calibration, aggregating all available information, the cal_K^α-calibration is the appropriate metric. Each of these distinct calibration measures provides a specialized viewpoint, but in the following part of this paper, we will then focus on the cal_K^α-calibration. The choice of focusing on the cal_K^α-calibration can be justified by the fact that if a model is cal_K^α-calibrated, so it will be cal_k^α-calibrated for each event of event and $\text{cal}_k(\tau)$-calibrated for any fixed τ.

By definition, the oracle function is calibrated according to the different distributions.
Any consistent marginal estimator is also asymptotically calibrated according to all definitions 3.1.

Plug-in (PI) estimators for the cal_K^α-calibration
From a population point of view, these calibration quantities can be estimated using any marginally consistent estimator. To remain general, we present the definitions using an abstract "plug-in" estimator, but this plug-in estimator can be implemented as the Aalen and Johansen (1978) estimator in practice (see Appendix A.4 for its formal definition). For each event type k, we denote the chosen marginally consistent estimator (PI) of cumulative distribution function by $\hat{F}_{\text{PI}}^k$.

Our plug-in cal_k^α-calibration estimator simply takes such an arbitrary marginal estimator $\hat{F}_{\text{PI}}^k$ (e.g., Aalen–Johansen) and plugs it into the definition of cal_K^α-calibration to obtain an empirical estimate of the calibration error. Intuitively, this is natural because, under mild assumptions, $\hat{F}_{\text{PI}}^k$ is a consistent estimator of the true marginal CIFs; hence, if a model is well cal_K^α-calibrated, its predicted marginal CIFs should match $\hat{F}_{\text{PI}}^k$ in the limit, which is exactly what the plug-in estimator tests.

Definition 3.4 (Plug-in calibrations)**.** The $\text{cal}_k(\tau)$ calibration can be approximated by:

Plug-in-$\text{cal}_k(\tau)$ calibration:

$$\text{PI}_k(\tau) = \left| \hat{F}_{\text{PI}(\mathcal{D}_{cal})}^k(\tau) - \frac{1}{|\mathcal{D}_{cal}|} \sum_{\mathbf{x}_i \in \mathcal{X}_{cal}} (\hat{F}_k(\tau|\mathbf{x}_i)) \right|$$

$$\text{Plug-in-}k^\alpha\text{ calibration:} \quad \text{PI}_k^\alpha = \left(\int_0^{t_{max}} \text{PI}_k(\tau)^\alpha \mathrm{d}\tau \right)^{1/\alpha}$$

Where in practice, we compute the integral by a Riemman In the same vein, we can introduce the approximation of the joint calibration as:

$$\text{Plug-in-cal}_K^\alpha\text{-calibration:} \qquad \text{PI}_{cal}^\alpha = \sum_{k=1}^{K} \text{PI}_k^\alpha$$

We obtain an equivalent proposition as 3.1 here with the PI-calibrations:

Proposition 3.2 (Consistency of the PI-cal_K^α-calibration)**.** *As the sample size (m) of the calibration set $\mathcal{D}_{cal}$ goes to infinity, for a fixed α the Plug-in-cal_K^α-calibration (Def. 3.4) PI_{cal}^α converges almost surely toward cal_K^α-calibration (Def. 3.3) cal_K^α, making it a consistent estimator of the cal_K^α-calibration.*

Proof. The proof is straightforward using the strong law of large numbers. □

Using the Aalen and Johansen (1978) estimator as the plug-in estimator, we define the *AJ-k-calibration* for each event k and the *AJ-cal_K^α-calibration.*

Plug-in-cal_K^α-calibration vs CR D-calibration
These calibrations differ in the sense that the cal_K^α-calibration (Def. 3.4) compares the marginal probabilities in time and the CR D-calibration (Def. 3.1) assesses whether the estimator returns well-behaved cumulative distribution functions. Nonetheless, we can derive several asymptotic results that relate both calibrations and D-calibration (Haider et al., 2020, and Appendix A.1). Appendix 6 gives an explanatory diagram with these relationships.

Theorem 2 (Equivalence between calibrations)**.** *If a model is* cal_K^α*-calibrated, then it is CR D-calibrated.*
Asymptotically, if a model is cal_K^α*-calibrated, it implies that it is CR D-calibrated.*
Under the assumption that each $\hat{F}_k$ *is strictly increasing and continuous, a model that is CR D-calibrated is also* cal_K^α*-calibrated.*
Additionally, if a model is CR D-calibrated, it is also calibrated according to the calibration plots.

Proof. Using the idea that when a model is cal_K^α-calibrated, the estimated marginal probabilities are equal to the oracle ones and that a consistent marginal estimator is marginally calibrated (Theorem 1), we obtain that if a model is cal_K^α-calibrated, it implies that the model will be D-calibrated.
For the other way, see proof in Appendix D.3. □

Choice of α: We use $\alpha = 2$ as the standard norm, but acknowledge that a more suitable value for α may exist depending on the specific use-case and could be found through further experimentation.

Test if a model is calibrated While the associated metrics give a comprehensive evaluation, a statistical test can conclude on whether observed decalibration is explained by sampling noise (limited size of calibration set). The presented CR-D-Calibration and PI-cal_K^α-calibration can be tested by applying a Kolmogorov (1933)-Smirnov (1948) (KS) test (with multiplicity adjustment) to check for the uniform distribution or to compare the Aalen and Johansen (1978) estimator with predictions, respectively (details in Appendix E).

4 HOW TO RECALIBRATE A MODEL?

Having defined and estimated two calibration measures for competing-risks models, we now ask: How can we recalibrate these models while preserving desirable properties?

4.1 Formulating recalibration as a minimization problem

To recalibrate a competing-risks model $(\hat{F}_k)_k$, typically trained on a dataset $\mathcal{D}_{\text{train}}$, on a calibration set $\mathcal{D}_{\text{cal}} = \{(\mathbf{x}_i, t_i, \delta_i)\}_i$, a classical post-hoc procedure consists in learning a transformation of the model outputs that minimizes a calibration loss on $\mathcal{D}_{\text{cal}}$.
Formally, let ℓ be a loss function that takes as first argument the predicted CIFs on the calibration set, $(\hat{F}_k(\cdot \mid \mathbf{x}_i))_{i,k}$, and as second argument the corresponding observed outcomes, $(t_i, \delta_i)_i$. In what follows, we take ℓ to be one of the calibration metrics introduced in Section 3. Let $\mathcal{G}$ be a family of recalibration maps g acting on the predicted CIFs. For any $g \in \mathcal{G}$, we define the recalibrated model by

$$\hat{F}_k^{(g)}(\tau \mid \mathbf{x}) \stackrel{\text{def}}{=} g\big(\tau, k, \hat{F}_k(\tau \mid \mathbf{x})\big),$$

and we then choose

$$g^* \stackrel{\text{def}}{=} \arg\min_{g \in \mathcal{G}} \ \ell\Big((\hat{F}_k^{(g)}(\cdot \mid \mathbf{x}_i))_{i,k},\ (t_i, \delta_i)_i\Big).$$

The final calibrated estimator is $\hat{F}_k^{\text{cal}} \stackrel{\text{def}}{=} \hat{F}_k^{(g^*)}$.

4.2 Recalibration using a PI-estimator (AJ-recalibration)

We introduce *AJ-recalibration*: a post-hoc recalibration method using the AJ-k calibration (Def. 3.4) as the function ℓ as detailed below. After training a model on $\mathcal{D}_{train}$, for each event k, we compute the AJ-k calibration at different times τ on $\mathcal{D}_{cal}$. At calibration time, the predictions are recalibrated simply by computing the AJ-k calibrations at different times so that g is simply $g\big(\tau, k, \hat{F}_k(\tau \mid x)\big) = \hat{F}_k(\tau \mid x) - \text{AJ}_k(\tau)$ (taking Aalen and Johansen (1978) as the PI estimator in Def.3.4). This recalibration method gives us nice asymptotic guaranties at population level (Appendix F.1.1) and detail more the method in Appendix F.1.

Discrimination is unchanged As mentioned before, discrimination of patients is an important property in the competing-risks setting. We use the definition of the C-index designed for the competing-risks setting introduced by Wolbers et al. (2009) and recall its formula in Appendix A.2. With our recalibration method, we can prove the following:

Theorem 3. *The C-index in the competing-risks setting is not affected by the AJ recalibration.*

Proof. The proof is inspired by Qi et al. (2024) in the survival setting. The comparison between samples is performed at a fixed time point τ, and we apply the same transformation at every point at τ. Thus, the order is intact and the C-index is unchanged. □

4.3 Competing-risks temperature scaling

Temperature scaling, proposed by Guo et al. (2017), is a recalibration method adapted to multiclass classification. Due to the presence of time and censored people, we have adapted the temperature scaling to the competing-risks setting. Indeed, we propose here a slightly more complex post-hoc recalibration method using the temperature scaling framework.
We recalibrate independently at each prediction time τ using the general setup of Section 4.1, taking as

loss function ℓ the plug-in $\text{cal}_k(\tau)$-calibration $\text{PI}_k(\tau)$, summed over all events k. As recalibration maps, we consider $\mathcal{G} = \{g_\beta\}_{\beta>0}$, where for any vector of positive probabilities p, $g_\beta(p) \stackrel{\text{def}}{=} \text{softmax}\big(\beta \cdot \text{logit(p)}\big)$.
In the competing-risks setting, we naturally extend this construction by applying g_β at each time τ to the vector of CIFs values $(\hat{F}_k(\tau \mid \mathbf{x}))_k$, with the softmax taken over the event index k.

Definition 4.1 (Temperature scaling)**.** We define, for each τ, temperature-scaling recalibration in the competing-risks setting as:

$$\left(\hat{F}_k^{\text{TS-cal}}(\tau)\right)_k \stackrel{\text{def}}{=} g_{\beta^*(\tau)}\left(\hat{F}_\cdot(\tau)\right)$$

$$\text{with} \quad \beta^*(\tau) \stackrel{\text{def}}{=} \arg\min_{\beta>0} \sum_{k=1}^{K} \text{PI}_k(\tau)\left(g_\beta\left(\hat{F}_\cdot(\tau)\right)\right).$$

Instead of relying on a direct target (as done in the multiclass setting), our method utilizes a marginally consistent estimator of the whole population. We show in Appendix F.2 that, with the true probabilities or a consistent estimator of the CIFs, the probabilities remain unchanged ($\forall\tau, \beta(\tau) = 1$). Finally, this design inherently ensures that the sum of probabilities for each individual equals 1.

5 EXPERIMENTAL STUDY

The code can be found as part of the hazardous library.

Baselines We compare different state-of-the-art competing risks models. All models have been trained with their default parameters on an internal cluster (40 CPUs, 252 GB of RAM, 4 NVIDIA Tesla V100 GPUs). We used the Aalen and Johansen (1978) estimator. We also used a linear model, the Fine and Gray (1999) estimator, which assumes proportional hazards (*i.e.*, the instantaneous risk of every event) as Cox (1972)). We also trained a tree-based model (Random Survival Forests, Ishwaran et al., 2008) that optimizes the discrimination of patients. We also studied DeepHit (Lee et al., 2018), a neural network designed to optimize a loss function that emphasizes discrimination and negative cross-entropy. We also added SurvivalBoost (Alberge et al., 2025), a tree-based model trained with a proper scoring rule. Finally, SurvTRACE (Wang and Sun, 2022) is a transformer model that models the observed quantiles of the events.

Datasets The experiments were made on a synthetic dataset and two real-life datasets (see Appendix I):

- Synthetic: We design a synthetic dataset with linear features with dependent censoring, and use it in the following with three competing events.

Table 2: **Competing risks calibrations:** CR-$\hat{D}_{cal}$ and AJ-cal$_K^\alpha$-calibrations computed on the SEER dataset with 10k training samples with $\alpha = 2$ over 5 random seeds. We recover that Aalen and Johansen (1978) is the most calibrated model while SurvTRACE has the worst metrics (expected given Figure 1). The table also reports the proportion of seeds passing the 5% calibration test (Appendix E), with higher values indicating better calibration.

	AJ-K-cal		CR-$\hat{D}_{cal}$	
Model	mean ± std	test	mean ± std	test
AalenJohansen	0.00 ± 0.00	100%	0.05 ± 0.01	100%
DeepHit	0.18 ± 0.13	0%	1.13 ± 0.56	0%
FineGray	0.00 ± 0.00	100%	1.18 ± 2.30	20%
RSF	0.00 ± 0.00	40%	0.09 ± 0.04	80%
SurvTRACE	0.24 ± 0.17	0%	1.34 ± 0.06	0%
SurvivalBoost	0.01 ± 0.00	0%	0.09 ± 0.01	100%

- SEER (2025): This real-life dataset contains 470,000 patients with a 10-year follow-up. After preprocessing, it yields three competing events: death from breast cancer, cardiac events, and other causes.

- METABRIC (Rueda et al., 2019): With 1,700 patients, this real-life dataset contains two competing events, with a follow-up period of up to 30 years.

$\hat{D}$-calibration and AJ-calibration on real-life datasets Table 2 shows the CR $\hat{D}$-calibration and the AJ-k-calibration for for the SEER dataset trained over 10k samples. We also show whether the KS test assessed the calibration of the model. In this context, we see that the model that is the most calibrated overall is the Aalen and Johansen (1978) estimator, and that DeepHit and SurvTRACE are the models that are less calibrated. Other results (METABRIC, SEER 100k training datapoints, and the synthetic dataset) can be found in Appendix H.2. The method least CR $\hat{D}$-calibrated, SurvTRACE, is also the farthest from the Aalen and Johansen (1978) in Fig. 1. Across all datasets (Tables 6 through 10), the computation of the CR-$\hat{D}$-calibration for each event reveals a consistent trend: most models, excepted DeepHit, exhibit their highest calibration error for the minority class. This is a concerning issue in the competing risks setting.

Recalibrations: impact on the metrics We study the probabilities before and after the recalibration. For this, we use the Integrated Brier score (IBS, Graf et al., 1999; Schoop et al., 2011). This is the sum over time of the Brier Score re-weighted to account for the censoring. It is a strictly proper scoring rule: its minimum gives the true probabilities. Recalibration might influence the IBS as it alters the probabilities for

a poorly calibrated model. Figure 4 shows the metrics on the SEER dataset (Appendix H gives the results on the METABRIC dataset, the synthetic dataset, and for different training sizes of the SEER dataset). Figure 4a shows the IBS (lower is better) before and after recalibration over 5 different seeds. It shows that, after recalibration, probability errors for the different models are smaller, while having the AJ-cal$_K^\alpha$-calibration and the D-CR-calibration smaller (Fig. 4b 4c).
As mentioned before, both recalibrations, applied to predicted probabilities, leave discrimination power (*i.e.* C-index) unchanged. Nevertheless, if the corresponding input probabilities do not initially sum to 1 across predicted classes for each individual, their renormalization does change discrimination (Figure 15).

Comparison between recalibration methods The best recalibration method depends on the dataset size and desired performance. Across our experiments (Section H), AJ-recalibration yielded superior calibration metrics universally and lower IBS on smaller datasets (Fig. 4a and Fig. 9). Conversely, the Temperature Scaling approach seems more effective for larger datasets (Fig. 27) and has the added benefit of producing meaningful individual probabilities.

6 DISCUSSION AND CONCLUSION

Limitations We have introduced 2 metrics, with complementary strengths and weaknesses. A limitation of CR-D-calibration, as in Haider et al. (2020), is that we assume a prediction horizon $t_{\max}$ large enough for cumulative incidence functions to approach the probability of the event of interest (*i.e.* $\forall x, \hat{F}_k(t_{max}|\mathbf{x}) \approx \hat{F}_k(\infty|\mathbf{x})$), which may not hold in practice. In contrast, the PI-cal$_K^\alpha$-calibration does not have this limitation by not relying on the behavior of the cumulative incidence function at $t_{\max}$. However, it requires sufficiently large data to ensure that the PI estimator closely approximates the true marginal distribution.
A natural direction for future work is a dedicated simulation study under known ground truth that systematically investigates the finite-sample behavior of our calibration estimators and recalibration procedures (varying sample sizes, as well as covariate, event-time, and censoring distributions) and assesses whether pathologies similar to those reported for classification calibration errors by Gruber and Buettner (2022) arise in our setting, in which case it would be important to explore more robust metrics and tests for competing risks.

Conclusion We addressed a critical gap in survival analysis with two novel, oracle-minimized calibration metrics for competing-risks settings: the *CR-D-calibration* and the cal$_K^\alpha$*-calibration*, proving their equivalence under a mild assumption. For practical use, we provided for both consistent estimators and corresponding recalibration formulation and procedures (one derived from AJ-k-calibration and an adaptation of the temperature scaling). We have shown that those methods improve model calibration while preserving discriminative capacity and increasing their abilities to predict individualized probabilities.
Our application of these new metrics to SOTA competing risks models on both synthetic and real datasets revealed heterogeneity in their calibration, increasing the urgent necessity of this evaluation criteria. Ultimately, relying on poorly calibrated predictions risks inadequate decisions in critical applications, emphasizing the need to consider our proposed calibration measures alongside other performance criteria.

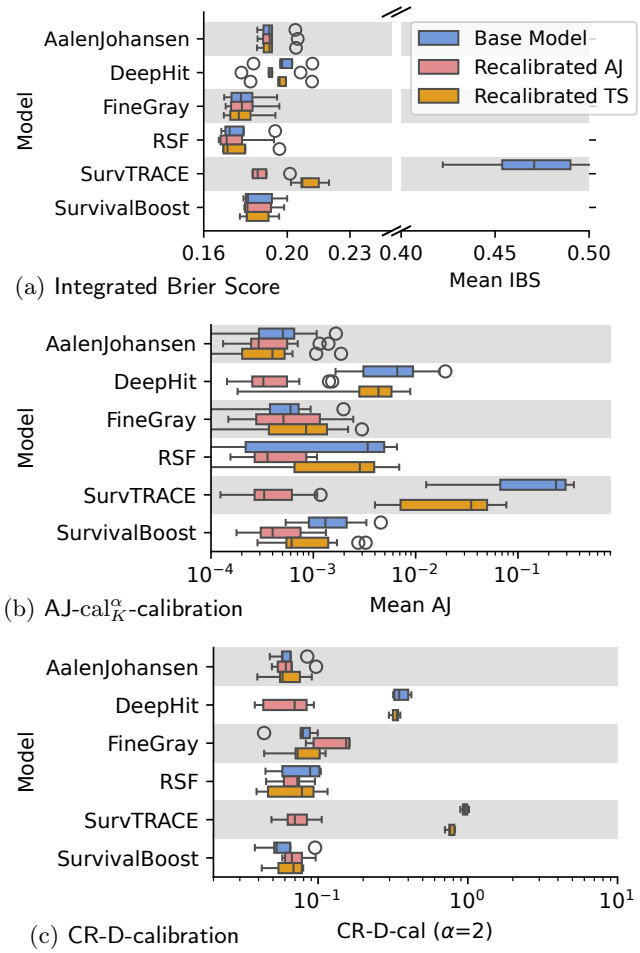

Figure 4: **IBS and calibration on METABRIC.** Boxplots over 5 seeds of the integrated Brier score (IBS), AJ-cal$_K^\alpha$-calibration (with parameter $\alpha = 2$), and CR-D-calibration ($\alpha = 2$) for each model and its AJ-recalibration and temperature-scaling (TS) recalibration. Results on the synthetic and SEER datasets are reported in Appendix H.

Acknowledgments JA, JA, and GV acknowledge funding from the European Union's Horizon Europe research and innovation program under grant agreement No 1010954433 (Intercept-T2D). TH acknowledge funding from the French Ministry of Health.

References

Aalen, O. O. and Johansen, S. (1978). An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. *Scandinavian Journal of Statistics*, 5(3):141–150. Publisher: [Board of the Foundation of the Scandinavian Journal of Statistics, Wiley].

Alberge, J., Maladière, V., Grisel, O., Abécassis, J., and Varoquaux, G. (2025). Survival Models: Proper Scoring Rule and Stochastic Optimization with Competing Risks. *AIStats*.

Austin, P. C., Putter, H., Giardiello, D., and Van Klaveren, D. (2022). Graphical calibration curves and the integrated calibration index (ICI) for competing risk models. *Diagnostic and Prognostic Research*, 6(1):2.

Berta, E., Holzmüller, D., Jordan, M. I., and Bach, F. (2025). Rethinking Early Stopping: Refine, Then Calibrate. arXiv:2501.19195 [cs].

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*.

Booth, S., Mozumder, S. I., Archer, L., Ensor, J., Riley, R. D., Lambert, P. C., and Rutherford, M. J. (2023). Using temporal recalibration to improve the calibration of risk prediction models in competing risk settings when there are trends in survival over time. *Statistics in Medicine*, 42(27):5007–5024. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.9898.

Candès, E. (2023). Conformalized survival analysis. *Statistical Methodology*, 85(1).

Casella, G. and Berger, R. L. (2002). Theorem 2.1.10, p.54. *tatistical Inference (2nd ed.)*.

Chapfuwa, P., Tao, C., Li, C., Khan, I., Chandross, K. J., Pencina, M. J., Carin, L., and Henao, R. (2023). Calibration and Uncertainty in Neural Time-to-Event Modeling. *IEEE Transactions on Neural Networks and Learning Systems*, 34(4):1666–1680. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

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

David, F. N. and Johnson, N. L. (1948). The probability integral transformation when parameters are estimated from the sample. *Biometrika*, 35(1/2):182–190.

Farina, R., Kuchibhotla, A. K., and Tchetgen, E. J. T. (2025). Doubly Robust and Efficient Calibration of Prediction Sets for Censored Time-to-Event Outcomes. arXiv:2501.04615 [stat].

Filho, T. S., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., and Flach, P. (2023). Classifier Calibration: A survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9):3211–3260. arXiv:2112.10327 [cs].

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

Gerds, T. A., Andersen, P. K., and Kattan, M. W. (2014). Calibration plots for risk prediction models in the presence of competing risks. *Statistics in Medicine*, 33(18):3191–3203. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.6152.

Gorfine, M. and Hsu, L. (2011). Frailty-Based Competing Risks Model for Multivariate Survival Data. *Biometrics*, 67(2):415–426.

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

Gruber, S. and Buettner, F. (2022). Better uncertainty calibration via proper scores for classification and beyond. In *Advances in Neural Information Processing Systems*, volume 35, pages 8618–8632.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. arXiv:1706.04599 [cs].

Haider, H., Hoehn, B., Davis, S., and Greiner, R. (2020). Effective Ways to Build and Evaluate Individual Survival Distributions.

Huang, Y., Li, W., Macheret, F., Gabriel, R. A., and Ohno-Machado, L. (2020). A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, 27(4):621–633.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3). arXiv:0811.1645 [stat].

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

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari. 4: 83–91.*

Kull, M., Perello-Nieto, M., Kängsepp, M., Filho, T. S., Song, H., and Flach, P. (2019). Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. arXiv:1910.12656 [cs].

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

Nagpal, C., Li, X., and Dubrawski, A. (2021). Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3163–3175.

Perez-Lebel, A., Morvan, M. L., and Varoquaux, G. (2023). Beyond calibration: estimating the grouping loss of modern neural networks. arXiv:2210.16315 [cs, stat].

Perez-Lebel, A., Varoquaux, G., Koyejo, S., Doutreligne, M., and Morvan, M. L. (2025). Decision from suboptimal classifiers: Excess risk pre-and post-calibration. *AIStats*.

Qi, S.-a., Yu, Y., and Greiner, R. (2024). Conformalized Survival Distributions: A Generic Post-Process to Increase Calibration. arXiv:2405.07374 [cs].

Qin, J., Piao, J., Ning, J., and Shen, Y. (2024). Conformal predictive intervals in survival analysis: a resampling approach. arXiv:2408.06539 [stat].

Rueda, O. M., Sammut, S.-J., Seoane, J. A., Chin, S.-F., Caswell-Jin, J. L., Callari, M., Batra, R., Pereira, B., Bruna, A., Ali, H. R., et al. (2019). Dynamics of breast-cancer relapse reveal late-recurring er-positive genomic subgroups. *Nature*, 567(7748):399–404.

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

Scikit-learn (2025). Multi-class support in isotonic recalibration. https://scikit-learn.org/stable/modules/calibration.html#multiclass-support. [Online; accessed May-2025].

SEER (2025). SEER database. https://seer.cancer.gov/causespecific. [Online; accessed May-2025].

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*.

Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., and Schön, T. B. (2019). Evaluating model calibration in classification. arXiv:1902.06977 [cs].

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., Steyerberg, E. W., diagnostic tests, T. G. E., and prediction models' of the STRATOS initiative Bossuyt Patrick Collins Gary S. Macaskill Petra McLernon David J. Moons Karel GM Steyerberg Ewout W. Van Calster Ben van Smeden Maarten Vickers Andrew J. (2019). Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):230.

Wang, J.-G. (1987). A note on the uniform consistency of the kaplan-meier estimator. *The Annals of Statistics*, pages 1313–1316.

Wang, Z. and Sun, J. (2022). SurvTRACE: Transformers for Survival Analysis with Competing Events. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–9. arXiv:2110.00855 [cs, stat].

Wikipedia (1984). Probability integral transform.

Wolbers, M., Koller, M. T., Witteman, J. C. M., and Steyerberg, E. W. (2009). Prognostic Models With Competing Risks: Methods and Application to Coronary Risk Prediction. *Epidemiology*, 20(4):555–561.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, Edmonton Alberta Canada. ACM.

Zhang, Z., Cortese, G., Combescure, C., Marshall, R., Lee, M., Lim, H. J., and Haller, B. (2018). Overview of model validation for survival regression model with competing risks using melanoma study data. *Annals of Translational Medicine*, 6(16):325.

Checklist

1. For all models and algorithms presented, check if you include:

(a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.
Yes, sections 2 and 3 present the calibration measures and the recalibration methods.

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
Yes, the properties of the presented methods have been tackled in sections 2 and 3.

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
Yes, the code is provided in the supplemental material.

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results.
Yes, sections 2 and 3 present the calibration measures and the recalibration methods and their theoretical properties.

(b) Complete proofs of all theoretical results.
Yes, all proofs can be found in the appendix, and sketch proofs in the main manuscript.

(c) Clear explanations of any assumptions.
Yes, each assumption made is also explained in plain English.

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).
Yes, all code is provided in the supplemental material.

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
Yes, all the training details can be found in the appendix.

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
Yes, in each experiment, error bars are provided.

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
Yes, computing architecture can be found in the appendix.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets.
Yes, all models and datasets that were used are mentioned and cited.

(b) The license information of the assets, if applicable.
Not Applicable

(c) New assets either in the supplemental material or as a URL, if applicable.
Yes, in the code provided in the supplemental material.

(d) Information about consent from data providers/curators.
Not Applicable

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content.
Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots.
Not Applicable

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.
Not Applicable

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation.
Not Applicable

SUPPLEMENTARY MATERIALS

A FORMAL DEFINITIONS

Definition A.1 (D-calibration: Survival setting)**.** For a dataset $D = \{[\mathbf{x}_i, t_i, \delta_i] | i = 1, ..., n\}$, and any interval $[a, b] \subset [0, 1]$:

$$D([a, b]) = \{[\mathbf{x}_i, t_i, \delta_i = 1] \in D \quad | \quad \hat{S}(t_i|\mathbf{x}_i) \in [a, b]\} \tag{1}$$

Definition A.2 (C-index CR Wolbers et al. (2009))**.** We recall the definition of the C-index at time τ for the kth competing event Wolbers et al. (2009) as:

$$\mathrm{C}(\tau) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(A_{ij}\hat{W}_{ij,1}^{-1} + B_{ij}\hat{W}_{ij,2}^{-1})Q_{ij}(\tau)\mathbb{1}_{t_i \leq \tau, \delta_i = k}}{\sum_{i=1}^{n}\sum_{j=1}^{n}(A_{ij}\hat{W}_{ij,1}^{-1} + B_{ij}\hat{W}_{ij,2}^{-1})\mathbb{1}_{t_i \leq \tau, \delta_i = k}} \tag{2}$$

where:

$$A_{ij} = \mathbb{1}_{t_i < t_j \cup (t_i = t_j \cap \delta_j = 0)} \tag{3}$$

$$B_{ij} = \mathbb{1}_{t_i \geq t_j, \delta_j \neq k, \delta_j \neq 0} \tag{4}$$

$$\hat{W}_{ij,1} = \hat{G}(t_i|\mathbf{X} = \mathbf{x}_i)\hat{G}(t_i|\mathbf{X} = \mathbf{x}_j) \tag{5}$$

$$\hat{W}_{ij,2} = \hat{G}(t_i|\mathbf{X} = \mathbf{x}_i)\hat{G}(t_j|\mathbf{X} = \mathbf{x}_j) \tag{6}$$

$$Q_{ij}(t) = \mathbb{1}_{F_k(\tau|\mathbf{X}=\mathbf{x}_i) > F_k(\tau|\mathbf{X}=\mathbf{x}_j)} \tag{7}$$

where $\hat{G}$ is one estimator of the censoring function, $G(\tau|\mathbf{X} = \mathbf{x}) = \mathbb{P}(C > \tau|\mathbf{X} = \mathbf{x})$.

A.1 Marginal estimators in survival analysis and in the competing risks setting

Here is the definition of the Kaplan and Meier (1958) estimator, a marginal estimator in the survival analysis setting,

Definition A.3 (Kaplan and Meier (1958) estimator)**.**

$$\textbf{Kaplan-Meier (KM) estimator:} \qquad \hat{F}_{\mathrm{KM}}(\tau) = 1 - \hat{S}_{KM}(\tau) = 1 - \prod_{t_i \leq \tau} \frac{Y(t_i) - d(t_i)}{Y(t_i)} \tag{8}$$

$d(t)$ is the number of events at time t_i and $Y(t_i)$ is the number of individuals at risk at time t_i.

And now, in the competing risks setting, where there exist more than one event of interest, we write the definition of the Aalen and Johansen (1978) estimator.

Definition A.4 (Aalen and Johansen (1978) estimator)**.**

$$\textbf{Aalen-Johansen (AJ) estimator:} \qquad \hat{F}_{\mathrm{AJ}}^{k}(\tau) = \sum_{t_i \leq \tau} \hat{S}(t_i^-)\frac{d_k(t_i)}{Y(t_i)} \tag{9}$$

where $\hat{S}(t_i^-)$ is the survival function estimated from the Kaplan and Meier (1958) (see A.3) estimator, $d_k(t)$ is the number of events of type k at time t_i and $Y(t_i)$ is the number of individuals at risk at time t_i.

B SURVIVAL ANALYSIS RECALIBRATION METHOD

Here is the result of the recalibration using Qi et al. (2024)'s method. Using the D-calibration in survival analysis, onto the SEER dataset with 10k training samples and 3 competing risks, we clearly see that the method has a negative impact on the marginal calibration of the Aalen and Johansen (1978) estimator.

C DISTRIBUTION CALIBRATION IN THE COMPETING-RISKS SETTING (CR D-Cal)

In this part, we reuse the notations that were explained in the main paper.

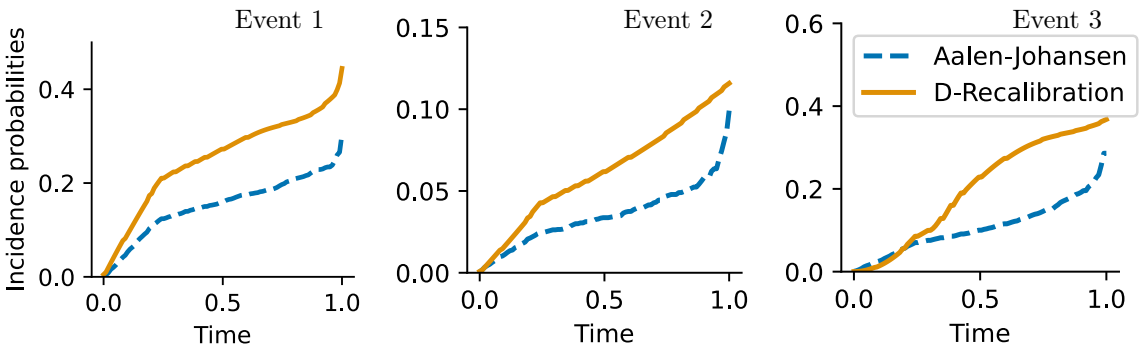

Figure 5: **Survival-analysis recalibration of good competing risk probabilities breaks them**. On this real-life dataset (SEER, 10k training samples), we apply the D-recalibration method in survival analysis Qi et al. (2024) to Aalen and Johansen (1978) by treating the risks independently (cause-specific). It incorrectly changes the marginal probabilities, though these were correct, as we used a consistent marginal estimator. The root of the problem is that survival-risk recalibration is applied to events independently, while their probabilities are tied Gorfine and Hsu (2011).

C.1 Without censoring

For now, we suppose there are no censored individuals. It is crucial to understand the hurdles in extending the D-calibration into the competing-risks setting. As mentioned earlier, in the survival analysis 2.2, the authors assume that the event of interest will almost surely happen, *i.e.* $F(\infty|\mathbf{x}) = 1$, they had that $F(T^*) \sim \mathcal{U}(0,1)$. Whereas, in the competing setting, that for each event k, $F_k(T^*, \mathbf{X}) \nsim \mathcal{U}(0,1)$ because $F_k(T^*, \mathbf{X}) \leq F_k(\infty|\mathbf{X}) = \mathbb{P}(\Delta^* = k|\mathbf{X} = \mathbf{x}) < 1$.
To overcome this issue, we introduce the D-calibration without any censored individuals as C.1:

Definition C.1 (CR D-calibration - No Censoring)**.** Thus, in the case of non-censorship, we define the Competing Risks Distribution calibration (CR D-calibration NC) as:

$$\text{CR D-calibration NC} \qquad D_{cal}^{NC} = \sum_{1}^{K} \left\| \frac{\mathbb{E}_{T^*,\Delta^*}(B_{[0,\rho]}^k)}{\mathbb{E}_{\mathbf{X}}(\hat{F}(\infty|\mathbf{X}))} - \rho \right\|_{\alpha} \tag{10}$$

$$\text{where} \qquad B_{[0,\rho]}^k = \mathbb{1}_{\hat{F}_{k/\infty}(T^*|\mathbf{X}) \in [0,\rho]} \mathbb{1}_{\Delta^* = k} \tag{11}$$

If the kth incidence function is D-calibrated, we have $B_{[0,\rho]}^k \simeq \rho$.
This distance corresponds to integrating the differences between the uniform ratio and the cumulative incidence function of a uniform law.

The CR D-calibration only have a meaning if the oracle functions (*i.e.* each of the CIF) are calibrated according to our definition. The following lemma C.1 helps us to overcome this issue:

Lemma C.1 (Uniform Ratio CDF)**.** *For the oracle function, conditioning of* Δ^* *and* $\mathbf{X}$*, this ratio is following a uniform distribution,* i.e.*:*

$$F_{k/\infty}(T^*|\mathbf{X}) \mid \Delta^* = k, \mathbf{X} \quad \sim \quad \mathcal{U}(0,1) \tag{12}$$

Proof Sketch. Using the Uniform CDF theorem Wikipedia (1984) and the conditioning over Δ^* and X. The whole proof can be found in Appendix. □

Theorem 4 (CR D-calibration - No Censoring)**.** *With lemma C.1, the oracle functions are CR D-calibrated. This proves that the result that* D_{NC}^{CR} *is proper.*
A marginally consistent estimator is asymptotically NC CR D-calibrated.

Proof sketch. For the oracle functions, the result is straightforward with the Lemma C.1. With Lebesgue's dominated convergence theorem, we show that a marginally consistent estimator is asymptotically CR D-calibrated. For more details, see below. □

Corollary C.1. *Because the Aalen-Johansen estimator is a consistent (see Lemma D.1) marginal estimator of the oracle functions, it is also asymptotically NC CR D-calibrated.*

For a given population and a given model, we also have to define the estimation of the D-calibration according to the population (named $\hat{D}$-calibration).

Definition C.2 (CR $\hat{D}$-calibration - No Censoring)**.** Thus, in the case of non-censorship, for a given population, we define:

$$b^k_{[0,\rho]} = \frac{\text{card}\{i \mid \delta_i = k \ \text{ and } \ \hat{F}_{k/\infty}(t_i \mid \mathbf{x}_i) \in [0,\rho]\}}{\text{sum}\{\hat{F}_k(\infty \mid \mathbf{x}_i)\}} \tag{13}$$

Then, we can define the estimation of the D-calibration as:

$$\hat{D} - \text{calibration estimation} \qquad \hat{D} \stackrel{\text{def}}{=} \sum_{k=1}^{K} \left\| b^k_{[0,\rho]} - \rho \right\|_\alpha \tag{14}$$

D PROOFS

For the following proofs, we will use the definition of the Aalen and Johansen (1978) estimator (written in Appendix A.4) given an event of interest. We first recall that, for each event of interest k, the Aalen and Johansen (1978) estimator is consistent.

Lemma D.1. *For each event of interest $k \in [\![1, K]\!]$, $\hat{F}^k_{\text{AJ}}$ is consistent.*

Proof. We fix an event of interest k. We consider the training dataset $\mathcal{D}_{train}$ with n samples. In this proof, we write the Aalen and Johansen (1978) estimator for an event of interest as F^k_n (resp. S_n).
Given the definition of Aalen and Johansen (1978), the proof can be detailed in 3 main points.

1. **Consistency of the survival function:** Kaplan and Meier (1958) estimator is consistent Wang (1987) in the sense that the survival function converges in probability for the uniform norm toward the oracle function: for all M such that $\mathbb{P}(T^* \geq M,\ \Delta^* = k) > 0$,

$$\forall \varepsilon > 0, \lim_{n\to\infty} \mathbb{P}(\sup_{\tau \leq M} |S_n(\tau) - S(\tau)| \leq \varepsilon) = 1.$$

2. **Convergence of the instant risk rate.** With the law of large numbers, we have: $\frac{d_k(t_i)}{Y(t_i)}$ that converges almost surely toward the instant risk of contracting the event of interest, *i.e.*, almost surely

$$\frac{d_k(t_i)}{Y(t_i)} \longrightarrow \lambda_k(t_i)dt = dt \lim_{dt \longrightarrow 0} \frac{\mathbb{P}(t_i \leq T^* \leq t_i + dt, \Delta = k) | T^* \geq t_i)}{dt}$$

3. **Convergence of the sum to the integral.** Finally, because we have an integrable function (Riemman), the sum converges toward the integral.

With those steps, we obtain the convergence in probabilities:

$$\hat{F}^k_{\text{AJ}}(\tau) = \sum_{t_i \leq \tau} \hat{S}(t_i^-) \frac{d_k(t_i)}{Y(t_i)} \longrightarrow \int_0^\tau \hat{S}(t)\lambda_k(t)dt = F^k(\tau)$$

Finally, we have obtain that

$$\lim_{n\to\infty} \mathbb{P}(|F^k_n(\tau) - F_k(\tau)| = 0) = 1.$$

The Aalen-Johansen estimator is marginally consistent. □

D.1 CR D-calibration without censoring mechanism

Lemma C.1 (Uniform Ratio CDF)**.** *For the oracle function, conditioning of Δ^* and $\mathbf{X}$, this ratio is following a uniform distribution,* i.e.*:*

$$F_{k/\infty}(T^*|\mathbf{X}) \mid \Delta^* = k, \mathbf{X} \quad \sim \quad \mathcal{U}(0,1) \tag{12}$$

Proof: distribution of F_k/F_k^∞. We prove that $F_{k/\infty}(T^* \mid \mathbf{X})|\Delta^* = k, \mathbf{X} \sim \mathcal{U}(0,1)$.

$$\forall t \in [0,1], \tag{15}$$

$$\mathbb{P}(\frac{F_k(T^*|\mathbf{X})}{F_k(\infty|\mathbf{X})} \leq t|\Delta^* = k, \mathbf{X}) = \mathbb{P}(T^* \leq F_k^{-1}(t\ F_k(\infty|\mathbf{X})|\mathbf{X})|\Delta^* = k, \mathbf{X}) \tag{16}$$

$$= \mathbb{P}(T^* \leq F_k^{-1}(t\ F_k(\infty|\mathbf{X})|\mathbf{X})|\Delta^* = k, \mathbf{X}) \tag{17}$$

$$= \frac{\mathbb{P}(T^* \leq F_k^{-1}(t\ F_k(\infty|\mathbf{X})|\mathbf{X}), \Delta^* = k|\mathbf{X})}{\mathbb{P}(\Delta^* = k|\mathbf{X})} \tag{18}$$

$$= \frac{F_k(F_k^{-1}(t\ F_k(\infty|\mathbf{X})|\mathbf{X}))}{F_k(\infty|\mathbf{X})} \tag{19}$$

$$= t \tag{20}$$

Thus, this result is constant for all $\mathbf{X}$. □

Theorem 4 (CR D-calibration - No Censoring)**.** *With lemma C.1, the oracle functions are CR D-calibrated. This proves that the result that D_{NC}^{CR} is proper.*
A marginally consistent estimator is asymptotically NC CR D-calibrated.

Proof. The proof is a special case of the setting with censoring. See Appendix D.2 for the whole proof. □

D.2 CR D-calibration with the censoring mechanism

Let the bucket $B_{[a,b]}$ be an extension of $B_{[0,\rho]}$ defined by:

$$B_{[a,b]} \stackrel{\text{def}}{=} B_{[0,b]} - B_{[0,a]}.$$

Lemma D.2 (Expectation with the oracle function)**.** *For the oracle function, we prove:*

$$\mathbb{E}(B_{[a,b]} \mid \mathbf{X}) = (b-a)F_k(\infty \mid \mathbf{X}).$$

Proof. Starting from

$$\mathbb{P}(T^* \leq t \mid \Delta^* = k,\ \mathbf{X}) = \frac{F_k(t \mid \mathbf{X})}{F_k(\infty \mid \mathbf{X})} \tag{21}$$

$$\mathbb{P}\left(\frac{F_k(T^* \mid \mathbf{X})}{F_k(\infty \mid \mathbf{X})} \in [a,b] \mid \Delta^* = k,\ \mathbf{X}\right) = b - a \quad \text{Assumption 2.1, Lemma C.1} \tag{22}$$

$$\text{And} \quad \mathbb{P}\left(\frac{F_k(T^* \mid \mathbf{X})}{F_k(\infty \mid \mathbf{X})} \in [a,b],\ \Delta^* = k \mid \mathbf{X}\right) = (b-a)F_k(\infty \mid \mathbf{X}) \quad \text{Bayes' Theorem} \tag{23}$$

We can't work with (28) because we don't have access to $\mathcal{L}(T, \Delta \mid \Delta^* = k)$ because of the censored individuals, so we'll work with (29).

First, we study the left part of the equality, we will study several cases: $T^* < C$ (we have access to the expression) and $C \leq T^*$ (we will estimate the expression). To make the computations more readable, we will also separate

the cases: when $\frac{F_k(C|\mathbf{X})}{F_k(\infty|\mathbf{X})} < a$ and $\frac{F_k(C|\mathbf{X})}{F_k(\infty|\mathbf{X})} \in [a,b]$, because $\frac{F_k(C|\mathbf{X})}{F_k(\infty|\mathbf{X})} > b$ is not possible (as $\frac{F_k(T^*|\mathbf{X})}{F_k(\infty|\mathbf{X})} \in [a,b]$) *i.e.*, we can decompose the computation as:

$$\begin{aligned}\mathbb{P}(\frac{F_k(T^* \mid \mathbf{X})}{F_k(\infty \mid \mathbf{X})} \in [a,b],\ \Delta^* = k \mid \mathbf{X}) = &\underbrace{\mathbb{P}(\cdots,\ T^* < C \mid \mathbf{X})}_{P_1} \\ &+ \underbrace{\mathbb{P}(\cdots,\ C \leq T^*, \frac{F_k(C \mid \mathbf{X})}{F_k(\infty \mid \mathbf{X})} \in [a,b] \mid \mathbf{X})}_{P_2} \\ &+ \underbrace{\mathbb{P}(\cdots,\ C \leq T^*, \frac{F_k(C \mid \mathbf{X})}{F_k(\infty \mid \mathbf{X})} < a \mid \mathbf{X})}_{P_3} \end{aligned} \tag{24}$$

Now, P_1 is just:

$$P_1 = \mathbb{P}(\frac{F_k(T \mid \mathbf{X})}{F_k(\infty \mid \mathbf{X})} \in [a,b],\ \Delta = k \mid \mathbf{X})$$

For the other parts P_2 and P_3, we write the shifted distribution $T^* \mid \Delta^* = k, X$:

$$\mathbb{P}[T^* \leq t,\ \Delta^* = k \mid T^* > s,\ \mathbf{X}] = \frac{F_k(t \mid \mathbf{X}) - F_k(s \mid \mathbf{X})}{S^*(s \mid \mathbf{X})}$$

And:

$$\begin{aligned}\mathbb{P}(F_k(T^* \mid \mathbf{X}) \leq u,\ \Delta^* = k \mid T^* > s,\ \mathbf{X}) &= \mathbb{P}(T^* \leq F_k^{-1}(u|\mathbf{X}),\ \Delta^* = k \mid T^* > s,\ \mathbf{X}) \quad \text{2.1} && (25)\\ &= \mathbb{P}(s \leq T^* \leq F_k^{-1}(u|\mathbf{X}),\ \Delta^* = k \mid \ \mathbf{X})\frac{1}{S(s|\mathbf{X})} && (26)\\ &= \frac{u - F_k(s \mid \mathbf{X})}{S(s \mid \mathbf{X})} \quad \text{if} \quad F_k(s \mid \mathbf{X}) \leq u \leq F_k(\infty \mid \mathbf{X}) && (27)\end{aligned}$$

We rewrite P_2 conditioning by C and $C \leq T^*$ and we use Assumption 2.2 *i.e.* $T^* \perp\!\!\!\perp C \mid \mathbf{X}$:

$$\begin{aligned}P_2 &= \mathbb{P}\left(\frac{F_k(T^* \mid \mathbf{X})}{F_k(\infty \mid \mathbf{X})} \leq b,\ \Delta^* = k, C \leq T^*, \frac{F_k(C \mid \mathbf{X})}{F_k(\infty \mid \mathbf{X})} \in [a,b] \mid \mathbf{X}\right) && (28)\\ &= \mathbb{E}\left(\mathbb{P}(\frac{F_k(T^* \mid \mathbf{X})}{F_k(\infty \mid \mathbf{X})} \leq b, \Delta^* = k \mid C, C \leq T^*, \mathbf{X})\mathbb{1}(C \leq T^*, \frac{F_k(C \mid \mathbf{X})}{F_k(\infty \mid \mathbf{X})} \in [a,b]) \mid \mathbf{X}\right) && (29)\\ &= \mathbb{E}\left(\frac{F_k(\infty \mid \mathbf{X})b - F_k(C \mid \mathbf{X})}{S^*(C \mid \mathbf{X})}\mathbb{1}(C \leq T^*, \frac{F_k(C \mid \mathbf{X})}{F_k(\infty \mid \mathbf{X})} \in [a,b]) \mid \mathbf{X}\right) && (30)\end{aligned}$$

Idem, with P_3, we obtain that:

$$P_3 = \mathbb{E}(\frac{F_k(\infty \mid \mathbf{X})(b-a)}{S^*(C \mid \mathbf{X})}\mathbb{1}(C \leq T^*, \frac{F_k(C \mid \mathbf{X})}{F_k(\infty \mid \mathbf{X})} < a) \mid \mathbf{X})$$

So, when we recall the definition of $B_{[a,b]}$:

$$\begin{aligned}B_{[a,b]} \stackrel{\text{def}}{=} &\mathbb{1}(\frac{F_k(T \mid \mathbf{X})}{F_k(\infty \mid \mathbf{X})} \in [a,b],\ \Delta = k) \\ &+ \frac{F_k(\infty \mid \mathbf{X})b - F_k(C \mid \mathbf{X})}{S^*(C \mid \mathbf{X})}\mathbb{1}(C \leq T^*, \frac{F_k(C \mid \mathbf{X})}{F_k(\infty \mid \mathbf{X})} \in [a,b]) \\ &+ \frac{F_k(\infty \mid \mathbf{X})(b-a)}{S^*(C \mid \mathbf{X})}\mathbb{1}(C \leq T^*, \frac{F_k(C \mid \mathbf{X})}{F_k(\infty \mid \mathbf{X})} < a) \in \sigma(T, \Delta, \mathbf{X})\end{aligned} \tag{31}$$

We do obtain that:

$$\mathbb{E}(B_{[a,b]} \mid \mathbf{X}) = (b-a)F_k(\infty \mid \mathbf{X})$$

□

Theorem 1 (Properness)**.** *Under assumptions 2.1, 2.2, 2.3, 2.4, the expectation of the bucket for the oracle function can be computed as:* $\mathbb{E}(B_{[a,b]} \mid \mathbf{X}) = (b-a)F_k(\infty \mid \mathbf{X})$*. This implies that the oracle functions are CR D-calibrated (Def. 3.1)* - i.e., $\forall \alpha \in \mathbb{R}^*$, D_α^{CR} *(eq. 1) is minimized for the oracle CIFs.*
Under the same assumptions, a marginally consistent estimator is asymptotically CR D-calibrated.

Proof. For the oracle functions, we have Lemma D.2: $\mathbb{E}(B_{[0,\rho]} \mid \mathbf{X}) = \rho F_k(\infty \mid \mathbf{X}) \implies \mathbb{E}_{\mathbf{X}}(\mathbb{E}(B_{[0,\rho]} \mid \mathbf{X})) = \rho\mathbb{E}(F_k(\infty)|\mathbf{X})$. So:

$$D^{CR} = \sum_1^K \left\| \frac{\mathbb{E}_{T,\Delta}(B^k_{[0,\rho]})}{\mathbb{E}_{\mathbf{X}}(\hat{F}_k(\infty|\mathbf{X}))} - \rho \right\|_\alpha \tag{32}$$

$$= \sum_1^K \left\| \frac{\rho\mathbb{E}_{\mathbf{X}}(F_k(\infty|\mathbf{X}))}{\mathbb{E}_{\mathbf{X}}(F_k(\infty|\mathbf{X}))} - \rho \right\|_\alpha \tag{33}$$

$$= 0 \tag{34}$$

Let's be $\hat{F}_k^n(\tau)$ a marginally consistent estimator of the cumulative incidence functions trained of n samples.
Let's suppose the bigger assumption that $C \perp\!\!\!\perp T^*$. Because we have $\forall k, \forall \tau, \hat{F}_k^n(\tau) \to F_k(\tau)$, this convergence is point-wise.
Moreover, we have $\forall k, \forall \tau, \mid \hat{F}_k^n(\tau) \mid \leq 1$ by definition of $\hat{F}_k^n(\tau)$.
So, by Lebesgue's dominated convergence theorem, we obtain that:

$$\forall k, \forall \tau, \quad \lim_{n\to\infty} \mathbb{E}(|\hat{F}_k^n(\tau) - F_k(\tau)|) = 0$$

With that, we obtain exactly the same computations as before with the marginal estimator. □

D.3 Links between calibration measures

Theorem 2 (Equivalence between calibrations)**.** *If a model is* cal_K^α*-calibrated, then it is CR D-calibrated.*
Asymptotically, if a model is cal_K^α*-calibrated, it implies that it is CR D-calibrated.*
Under the assumption that each F_k *is strictly increasing and continuous, a model that is CR D-calibrated is also* cal_K^α*-calibrated.*
Additionally, if a model is CR D-calibrated, it is also calibrated according to the calibration plots.

Proof. We will do the proof of the second part in the asymptotic case ; the non-asymptotic case is similar but simpler.
Consider the sequence of estimators indexed by $n \to \infty$, which is asymptotically CR D-calibrated. The asymptotic CR D-calibrated refers to the property that

$$\forall \rho \in [0,1], \mathbb{E}(B^{k,n}_{[0,\rho]})/\hat{F}_k^n(\infty) \longrightarrow_{n\to\infty} \rho.$$

A straightforward computation following the same ideas as in Lemma D.2 gives us

$$\mathbb{E}(B^{k,n}_{[0,\rho]}) = \mathbb{P}(\hat{F}^n_{k/\infty}(T^*) \leq \rho, \Delta^* = k),$$

so we obtain

$$\forall \rho \in [0,1], \mathbb{P}(\hat{F}^n_{k/\infty}(T^*) \leq \rho, \Delta^* = k)/\hat{F}_k^n(\infty) \longrightarrow_{n\to\infty} \rho. \tag{35}$$

Applying (35) to $\rho = 1$ yields $\mathbb{P}(\Delta^* = k)/\hat{F}_k^n(\infty) \to 1$ and so $\hat{F}_k^n(\infty) \to \mathbb{P}(\Delta^* = k)$ as $n \to \infty$. Hence, substituting the limit of the denominator into (35),

$$\forall \rho \in [0,1], \mathbb{P}(\hat{F}^n_{k/\infty}(T^*) \leq \rho \mid \Delta^* = k) \longrightarrow_{n\to\infty} \rho. \tag{36}$$

Since the CIF of $T^* \mid \Delta^* = k$ is $F_{k/\infty}$ and the function $\hat{F}^n_{k/\infty}$ is stricly increasing, the equation (36) is equivalent to

$$\forall \rho \in [0,1], F_{k/\infty} \circ (\hat{F}^n_{k/\infty})^{-1}(\rho) \longrightarrow_{n\to\infty} \rho. \tag{37}$$

Using the continuity of $F^{-1}_{k/\infty}$, we have that $(\hat{F}^n_{k/\infty})^{-1}$ converges pointwise to $F^{-1}_{k/\infty}$. By the classical Dini's theorem on the uniform convergence of monotone functions, this convergence is in fact uniform. Therefore, we can invert the functions and obtain the convergence of $\hat{F}^n_{k/\infty}$ toward $F_{k/\infty}$. This concludes the proof.

□

$$\begin{tikzcd}
\text{PI } K-cal \arrow[r, "m\to\infty"'] & K-cal \arrow[d, leftrightarrow, "n\to\infty"] \\
\text{CR } \hat{D}_{\text{cal}} \arrow[r, "m\to\infty"'] & \text{CR } D_{\text{cal}}
\end{tikzcd}$$

Figure 6: **Links between all calibrations.** n represents the number of samples in the training set and m represents the number of samples the calibration set. The equivalence between both measures is true with a mild assumption.

E TESTING THE CALIBRATION

Here, we describe the process that we used to design, compute, and understand the limitations of the different tests.

E.1 Designing the tests

As proven before (Th. 1), the event-specific bucket $B^k_{[0,.]}$ should follow a uniform distribution. With this, we can evaluate the CR-D-calibration using a (Kolmogorov, 1933)-(Smirnov, 1948) (KS) test (with our definition of CR–D–Calibration, taking $\alpha = \infty$ reduces the criterion to the KS statistic). An inspirational alternative, in Haider et al. (2020), the authors discretize $[0,1]$ into a fixed number of buckets and apply a Pearson χ^2 test. Because testing is conducted separately for each competing event k, we adjust for multiplicity across events (*e.g.*, Bonferroni (1936)), acknowledging this yields a conservative decision rule.
In the same vein, for the PI-calibrations, we can assess the model's calibration with the KS test between the (Aalen and Johansen, 1978) estimator and the marginal predictions of the estimator.
More powerful tests could be defined by exploring the limit distribution, and/or integrating bootstrap techniques to better address the censoring distribution.

E.2 Computing the tests

To compute the proposed calibration tests, we employ the same methodologie for CR-D-calibration and PI-cal$^\alpha_K$-calibration.

For both calibrations, we use a KS test. Here is the description on the scheme for the CR-D-calibration. This can be adapted to the PI-cal$^\alpha_K$-calibration, using the marginal predictions of the PI-estimator (in our case the Aalen and Johansen (1978) estimator) instead of the Uniform distribution.

- We predict the CIFs for each event on the $\mathcal{D}_{test}$.
- We compute the buckets from the CR-D)calibration. The buckets should follow a uniform distribution.
- We make a KS test onto each event to test the null hypothesis that they are drawn from a standard uniform distribution Uniform(0,1).

- Since a separate test is conducted for each competing event k, the resulting p-values must be adjusted for multiplicity (using the Bonferroni correction) to maintain the overall significance level across all tests.
- The test is considered has passed if the p-values are higher than $0.05 \times K$, with K the number of competing events.

E.3 Limitations

For the tests, we want to acknowledge that the tests are not that powerful. Indeed, we see that most models may pass one or several tests while being insufficiently calibrated (as shown in Figure 1). We really want to recommend using the metrics instead of the tests to understand which model is more calibrated, etc. But, in a primary approach, the calibration tests can be applied to have a more precise idea.

F RECALIBRATION

F.1 Recalibration using AJ-k calibration: Framework

We describe here our process to recalibrate competing risks models. The same process can be applied to survival models with only one event of interest.

We have $\mathcal{D} = (x_i, t_i, \delta_i)_{1 \leq i \leq N}$. Our dataset is split into three different sets: a training set $\mathcal{D}_{train}$, a testing set $\mathcal{D}_{test}$, and a third set: a calibration set $\mathcal{D}_{cal}$.
Our recalibration process can be explained as:

- We train a given model onto $\mathcal{D}_{train}$.
- At calibration time, we denote $\mathcal{D}_{cal} = (\mathcal{X}_{cal}, \mathcal{Y}_{cal})$. For each event (resp. survival to any event), we predict the incidence function for each individual $\mathbf{x}_i \in \mathcal{X}_{cal}$ at d fixed times $(\tau_j)_{1 \leq j \leq d}$. The different incidence functions (resp. survival to any event) for each individual $\hat{F}_k(.|\mathbf{X} = \mathbf{x}_i)$ (resp. $\hat{F}_0(.|\mathbf{X} = \mathbf{x}_i) \stackrel{\text{def}}{=} \hat{S}(.|\mathbf{X} = \mathbf{x}_i)$) are called individualized incidence functions (IIFs) in the following. We choose the different times $(\tau_\nu)_{1 \leq \nu \leq d}$ as the quantiles of the duration of the calibration set. We train an Aalen and Johansen (1978) estimator onto $\mathcal{D}_{cal}$ and predict the IIFs at the same times $(\tau_\nu)_{1 \leq \nu \leq d}$.
- We compute the marginal differences between the Aalen and Johansen (1978) estimator and the model on $\mathcal{D}_{cal}$. *I.e.* $\forall k \in [0, K], \forall \tau_j$, $d_j^k = \hat{F}_k^{AJ}(\tau_j) - \frac{1}{|\mathcal{D}_{cal}|} \sum_{\mathbf{x} \in \mathcal{X}_{conf}} \hat{F}_k(\tau_j|\mathbf{X} = \mathbf{x}_i)$ to approximate $\hat{F}_k^{AJ}(\tau_j) - \int_{\mathbf{x} \in \mathcal{X}_{cal}} \hat{F}_k(\tau_j|\mathbf{X} = \mathbf{x}) d\mathbf{x}$.
- Then, on a new set of data $\mathcal{D}_{test}$, we predict the IIFs with the trained model at $(\tau_j)_{1 \leq j \leq \nu}$ and obtain $(\hat{F}_k(\tau_j|\mathbf{X} = \mathbf{x}_i))_{1 \leq j \leq \nu, \mathbf{x}_i \in \mathcal{X}_{test}, k \in [\![0, K]\!]}$. Then, for each x_i, we return the re-calibrated probabilities at each time τ_j defined as
$$\forall k \in [\![0, K]\!], \tilde{F}_k(\tau_j|\mathbf{X} = \mathbf{x}_i) = \hat{F}_k(\tau_j|\mathbf{X} = \mathbf{x}_i) - d_j^k.$$

This framework is close to the one used in conformalized survival analysis Qi et al. (2024); Farina et al. (2025).

F.1.1 Theoritical asymptotic guaranties

The AJ recalibration allows us to obtain asymptotic conformal intervals that we defined here.

Asymptotic Conformal Intervals To our knowledge, no formal definition of conformal prediction has been given in the competing-risks setting. For a level $\gamma \in [0, 1]$ and an event $k \in [\![1, K]\!]$, following the works of Booth (Booth et al., 2023; Candès, 2023; Farina et al., 2025; Qi et al., 2024; Qin et al., 2024) on conformal prediction in survival analysis, we define an asymptotically marginally calibrated Upper Predictive Bound (UPB) as a function $\hat{L}$ estimated from $\mathcal{D}_{cal}$ satisfying

$$\liminf_{|\mathcal{D}_{cal}| \to \infty} \mathbb{P}\big(T^* \leq \hat{L}(\mathbf{X}) \mid \Delta^* = k\big) \geq 1 - \gamma,$$

where $(\mathbf{X}, T^*, \Delta^*) \sim \mathcal{D}$ is an independent observation drawn from the same distribution.

Lemma F.1 (Asymptotic Conformal intervals)**.** *Under Assumption 2.3, if a model is AJ-recalibrated, it implies that it has naturally asymptotically marginally calibrated UPBs.*

Proof. By combining the consistency and the Gaussian Central limit theorem for the Aalen–Johansen estimator established in Aalen and Johansen (1978), we directly derive UPBs whose width depends on the sample size of $\mathcal{D}_{cal}$. □

F.2 Recalibration using temperature scaling

Here, we explain how we adapted the temperature scaling framework Guo et al. (2017) for recalibration in the competing risks setting, re-using the AJ-k recalibration framework.

This method works as follows:

- We train a given model onto $\mathcal{D}_{cal}$
- We take the quantile of the duration distribution on $\mathcal{D}_{cal}$), that gives us a time grid on which we will perform the recalibration. This recalibration will be performed independently at each of these times.
- At each fixed time point τ, a separate temperature scaling model is trained. This typically involves learning a single scalar parameter, $\beta(\tau)$ (the "temperature"), which is applied multiplicatively to the logits (or linear scores) of the original model's output before the final activation function (a softmax).
- The true or "actual proportion" of events at the chosen time point is estimated using a non-parametric method, such as the (Aalen and Johansen, 1978) estimator or another form of marginally consistent PI estimator. This estimator provides the gold-standard, non-parametric survival probability at that specific time, serving as the calibration target for the temperature scaling model.
- The temperature scaling model is trained to minimize the difference (*e.g.*, using a cross-entropy loss) between the original model's marginal predictions and the PI-estimator's output for that same time τ.
- To summarize, you are not scaling each of the cumulative incidence function with one temperature; you are finding an optimal temperature parameter $\beta(\tau)$ for each time τ to align the model's predicted marginal probability with the non-parametric marginal probability.

G COMPUTATION OF THE CR-D-CALIBRATION: ALGORITHM

To highlight how we compute the D-CR-calibration, here is the algorithm of the computation:

H EXPERIMENTAL PART

H.1 Experimental details

As said in the main part, we compare different state-of-the-art competing risks models. All models have been trained with their default parameters on an internal cluster (40 CPUs, 252 Gb of RAM, 4 NVIDIA Tesla V100 GPUs). The implementation was done in Python, using the lifelines library for the Aalen and Johansen (1978) estimator, the Pycox library `https://github.com/havakv/pycox` for the DeepHit model Lee et al. (2018), the "cmprsk" from the R library using rpy2 to train the Random Survival Forests model Ishwaran et al. (2008) and the hazardous Python library `https://github.com/soda-inria/hazardous` to train SurvTRACE Wang and Sun (2022) and compute the C-index, and the IBS. The calibration metrics in survival analysis were computed thanks to the SurvivalEval Python library `https://github.com/shi-ang/SurvivalEVAL`.
Following Candès (2023), to run the experiments, we have divided our dataset into three different dataset: 40% was used for the training set ($\mathcal{D}_{train}$), 40% to re-calibrate the models ($\mathcal{D}_{cal}$) and 20% for the test set ($\mathcal{D}_{test}$).

Algorithm 1 Computation of $\hat{D}_{\alpha}^{CR}$-Calibration

Require: $\{(t_i, \delta_i, \mathbf{x}_i)\}_{i=1}^{n}$, $\{\hat{F}_k\}_{k=1}^{K}$, $\hat{S}$, α, M.
  $\hat{D}_{\alpha}^{CR} \leftarrow 0$
  for $k \leftarrow 1$ to K **do**
    $w_{total} \leftarrow \sum_{i=1}^{N} \hat{F}_k(\infty \mid \mathbf{x}_i)$ # Denominator
    $I_k \leftarrow 0$ # Initialize integral
    for $\rho \leftarrow 1/M$ to 1 with step $1/M$ **do**
      $Count_k \leftarrow 0$
      $Sum_{cens} \leftarrow 0$
      for patient $i \leftarrow 1$ to N **do**
        # Compute normalized prediction
        $\hat{F}_{k/\infty} \leftarrow \hat{F}_k(t_i \mid \mathbf{x}_i)/\hat{F}_k(\infty \mid \mathbf{x}_i)$
        if $\hat{F}_{k/\infty} \in [0, \rho]$ **then**
          if $\delta_i = k$ **then**
            $Count_k \leftarrow Count_k + 1$
          if $\delta_i = 0$ **then**
            # Adjust for censored samples
            $cens \leftarrow \frac{\left(\hat{F}_k(\infty|\mathbf{x}_i)\cdot\rho - \hat{F}_k(t_i|\mathbf{x}_i)\right)}{\hat{S}(t_i|\mathbf{x}_i)}$
            $Sum_{cens} \leftarrow Sum_{cens} + cens$
      # Calculate bias for current ρ
      $\hat{b}_{[0,\rho]}^{k} \leftarrow \frac{Count_k + Sum_{cens}}{w_{total}}$
      $I_k \leftarrow I_k + \frac{\left|\hat{b}_{[0,\rho]}^{k} - \rho\right|^{\alpha}}{M}$
    $\hat{D}_{\alpha}^{CR} \leftarrow \hat{D}_{\alpha}^{CR} + (I_k)^{1/\alpha}$
  return $\hat{D}_{\alpha}^{CR}$

H.2 More results

Below are all the experimental results for the different datasets. For each dataset, we will show the calibration metrics, and then the IBS and the C-index before and after recalibration.

For the Table 9, we see one limitation of the tests. Although the AJ-cal$_K^{\alpha}$-calibration is really small (around 10^{-4}, Fig 27), not all tests are positive because of the sample size. A slight difference has an enormous impact on the test result.

Tables 8 and 10 reveal that, with the exception of DeepHit, the minority class (event 2) of the SEER dataset (Fig. 34) exhibits the largest calibration error across all models. This heterogeneity suggests that current competing risks models generally struggle to maintain good calibration on less frequent events, which represents a significant potential issue.

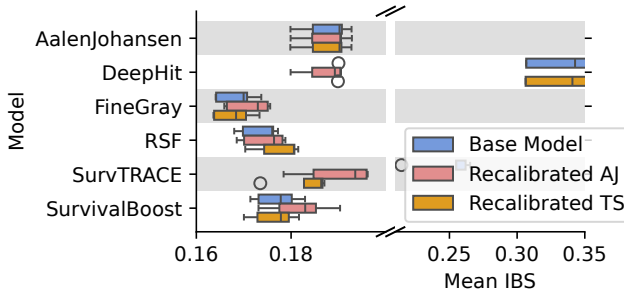

Figure 7: Mean Integrated Brier Score

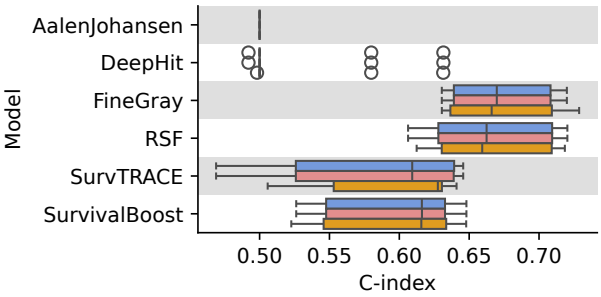

Figure 8: C-index

Figure 9: **Synthetic Dataset**, before and after recalibration, usual metrics.

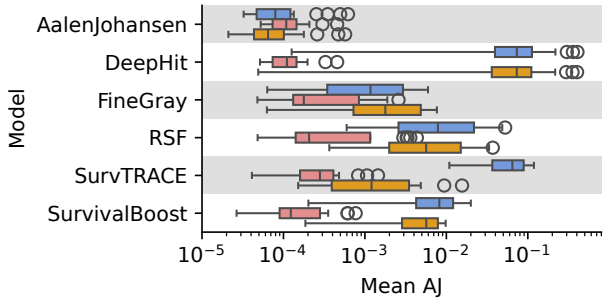

Figure 10: AJ Calibration

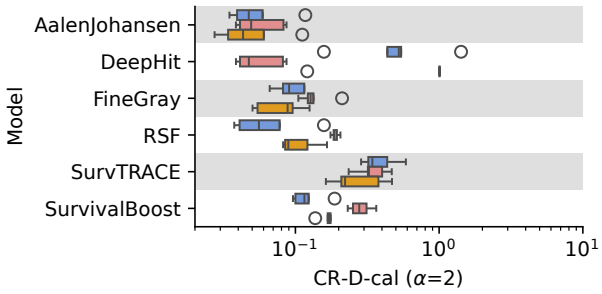

Figure 11: CR-D-calibration

Figure 12: **Synthetic Dataset**, before and after recalibration, calibration metrics.

Table 3: **Synthetic Dataset** KS test of D-CR-calibration with the multi-test correction. Computations are made over 5 seeds and we show the percentage of tests that have passed.

model	recalibration	Test D-CR-calibration	Test AJ-cal$_K^\alpha$-calibration
AalenJohansen	Base Model	100%	100%
AalenJohansen	Recalibrated AJ	60%	100%
AalenJohansen	Recalibrated TS	100%	100%
DeepHit	Base Model	0%	20%
DeepHit	Recalibrated AJ	40%	100%
DeepHit	Recalibrated TS	0%	20%
FineGray	Base Model	80%	100%
FineGray	Recalibrated AJ	0%	100%
FineGray	Recalibrated TS	80%	100%
RSF	Base Model	80%	0%
RSF	Recalibrated AJ	0%	100%
RSF	Recalibrated TS	80%	0%
SurvTRACE	Base Model	0%	0%
SurvTRACE	Recalibrated AJ	0%	100%
SurvTRACE	Recalibrated TS	0%	60%
SurvivalBoost	Base Model	80%	0%
SurvivalBoost	Recalibrated AJ	0%	100%
SurvivalBoost	Recalibrated TS	20%	0%

Table 4: **Synthetic Dataset**: CR-$\hat{D}$-calibration per event. We show the $\hat{D}_2^{CR}$.

Model	Event 1	Event 2	Event 3
AalenJohansen	0.03 ± 0.0	0.05 ± 0.03	0.04 ± 0.02
DeepHit	0.31 ± 0.42	0.17 ± 0.22	0.2 ± 0.25
FineGray	0.03 ± 0.01	0.05 ± 0.03	0.08 ± 0.02
RSF	0.04 ± 0.01	0.05 ± 0.04	0.06 ± 0.03
SurvTRACE	0.3 ± 0.14	0.34 ± 0.11	0.35 ± 0.13
SurvivalBoost	0.09 ± 0.01	0.08 ± 0.04	0.06 ± 0.02

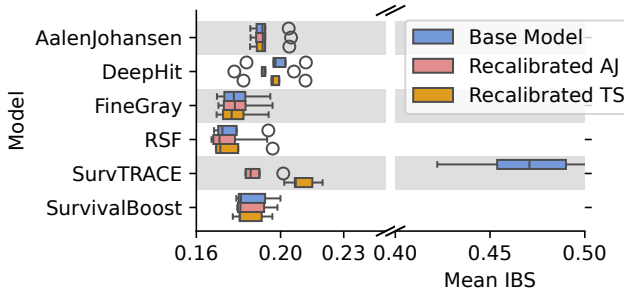

Figure 13: Mean Integrated Brier Score

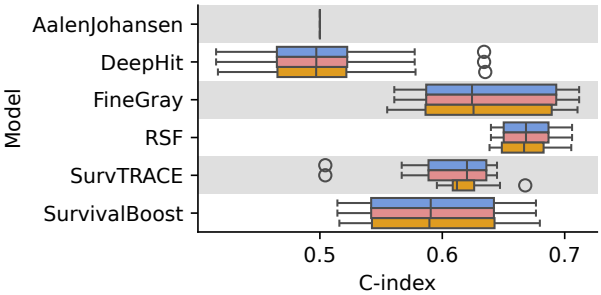

Figure 14: C-index

Figure 15: **METABRIC**, before and after recalibration, usual metrics.

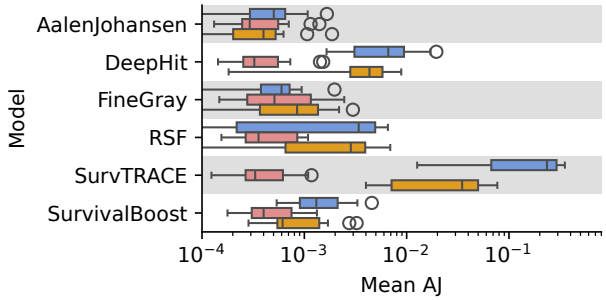

Figure 16: AJ Calibration

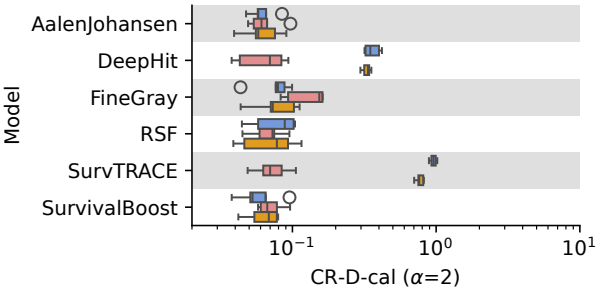

Figure 17: CR-D-calibration

Figure 18: **METABRIC**, before and after recalibration, calibration metrics.

Table 5: **METABRIC** KS test of calibrations with the multi-test correction. Computations are made over 5 seeds and we show the percentage of tests that have passed.

Model	Recalibration	Test CR-D-Calibration	Test AJ-cal$_K^\alpha$-calibration
AalenJohansen	Base Model	100%	100%
AalenJohansen	Recalibrated AJ	100%	100%
AalenJohansen	Recalibrated TS	100%	100%
DeepHit	Base Model	0%	0%
DeepHit	Recalibrated AJ	100%	100%
DeepHit	Recalibrated TS	0%	80%
FineGray	Base Model	100%	100%
FineGray	Recalibrated AJ	20%	100%
FineGray	Recalibrated TS	20%	100%
RSF	Base Model	100%	0%
RSF	Recalibrated AJ	60%	100%
RSF	Recalibrated TS	100%	0%
SurvTRACE	Base Model	0%	0%
SurvTRACE	Recalibrated AJ	100%	100%
SurvTRACE	Recalibrated TS	0%	0%
SurvivalBoost	Base Model	100%	80%
SurvivalBoost	Recalibrated AJ	60%	100%
SurvivalBoost	Recalibrated TS	100%	80%

Table 6: **METABRIC**: CR-$\hat{D}$-calibration per event. We show the $\hat{D}_2^{CR}$.

Model	Event 1	Event 2
AalenJohansen	0.05 ± 0.02	0.05 ± 0.01
DeepHit	0.34 ± 0.06	0.14 ± 0.03
FineGray	0.05 ± 0.02	0.07 ± 0.02
RSF	0.04 ± 0.01	0.08 ± 0.03
SurvTRACE	0.58 ± 0.06	0.77 ± 0.04
SurvivalBoost	0.05 ± 0.01	0.04 ± 0.02

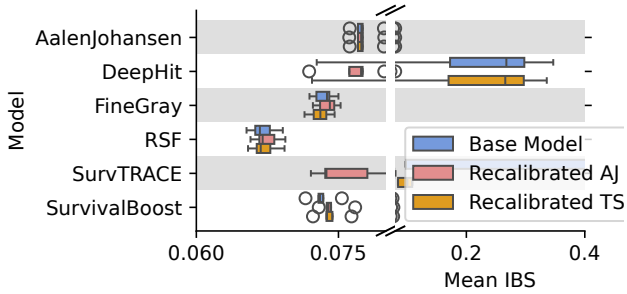

Figure 19: Mean Integrated Brier Score

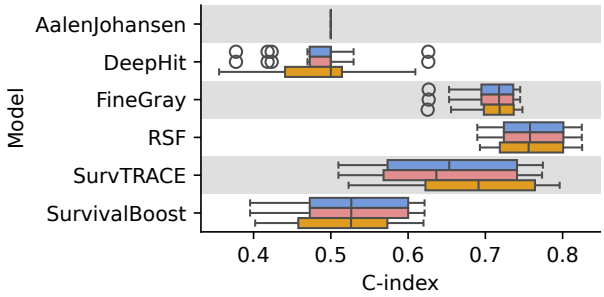

Figure 20: C-index

Figure 21: **SEER 10k training samples**, before and after recalibration, usual metrics.

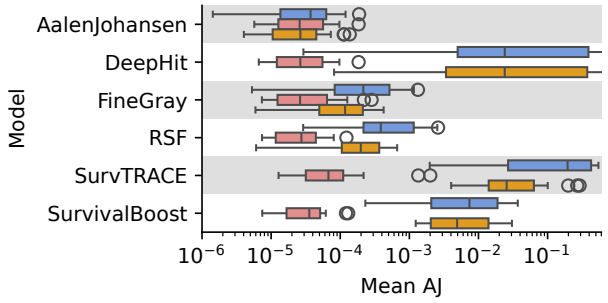

Figure 22: AJ Calibration

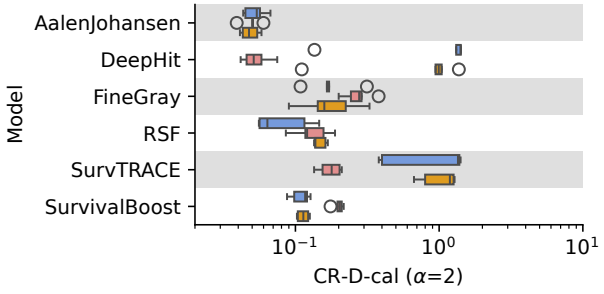

Figure 23: CR-D-calibration

Figure 24: **SEER 10k training samples**, before and after recalibration, calibration metrics.

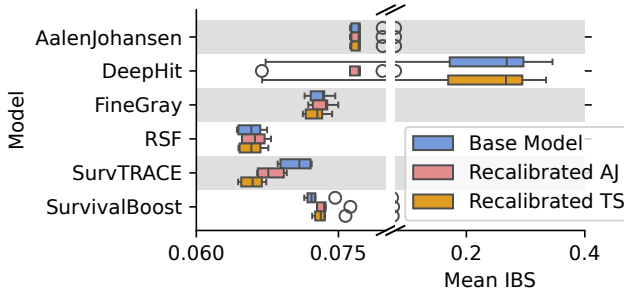

Figure 25: Mean Integrated Brier Score

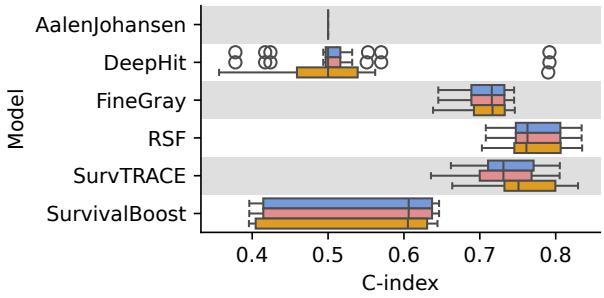

Figure 26: C-index

Figure 27: **SEER 100k training samples**, before and after recalibration, usual metrics.

Table 7: **SEER 10k training samples**, KS test of the calibrations with the multi-test correction. Computations are made over 5 seeds and we show the percentage of tests that have passed.

Model	Recalibration	Test D-CR-Calibration	Test AJ-cal$_K^\alpha$-calibration
AalenJohansen	Base Model	100%	100%
AalenJohansen	Recalibrated AJ	100%	100%
nAalenJohansen	Recalibrated TS	100%	100%
DeepHit	Base Model	0%	0%
DeepHit	Recalibrated AJ	80%	100%
DeepHit	Recalibrated TS	20%	0%
FineGray	Base Model	20%	100%
FineGray	Recalibrated AJ	0%	100%
FineGray	Recalibrated TS	0%	100%
RSF	Base Model	80%	40%
RSF	Recalibrated AJ	0%	100%
RSF	Recalibrated TS	0%	100%
SurvTRACE	Base Model	0%	0%
SurvTRACE	Recalibrated AJ	0%	100%
SurvTRACE	Recalibrated TS	0%	0%
SurvivalBoost	Base Model	100%	0%
SurvivalBoost	Recalibrated AJ	0%	100%
SurvivalBoost	Recalibrated TS	100%	0%

Table 8: **SEER 10k training points**: CR-$\hat{D}$-calibration per event. We show the $\hat{D}_2^{CR}$.

Model	Event 1	Event 2	Event 3
AalenJohansen	0.03 ± 0.01	0.04 ± 0.01	0.03 ± 0.01
DeepHit	0.77 ± 0.42	0.73 ± 0.38	0.79 ± 0.43
FineGray	0.05 ± 0.03	0.15 ± 0.09	0.12 ± 0.06
RSF	0.04 ± 0.01	0.08 ± 0.04	0.04 ± 0.01
SurvTRACE	0.54 ± 0.28	0.61 ± 0.34	0.57 ± 0.31
SurvivalBoost	0.07 ± 0.01	0.06 ± 0.02	0.07 ± 0.01

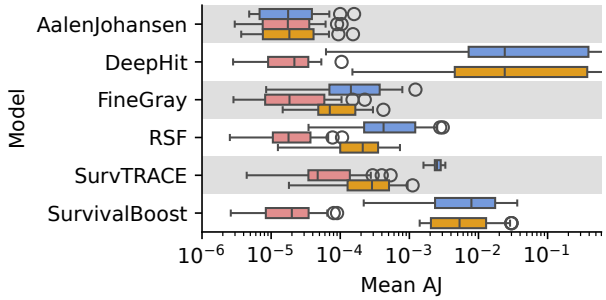

Figure 28: AJ Calibration

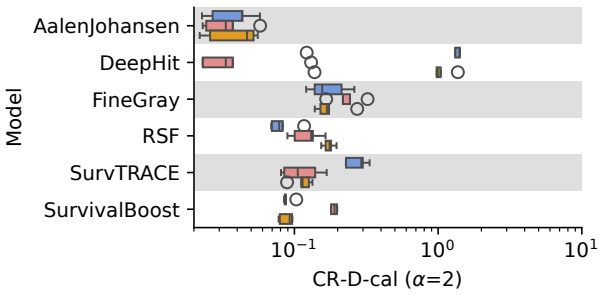

Figure 29: CR-D-calibration

Figure 30: **SEER 100k training samples**, before and after recalibration, calibration metrics.

Table 9: **SEER 100k training samples**, KS test of the calibrations with the multi-test correction. Computations are made over 5 seeds and we show the percentage of tests that have passed.

Model	recalibration	Test D-CR-Calibration	Test AJ-cal^2_K-calibration
AalenJohansen	Base Model	100%	100%
AalenJohansen	Recalibrated AJ	100%	100%
AalenJohansen	Recalibrated TS	100%	100%
DeepHit	Base Model	0%	0%
DeepHit	Recalibrated AJ	80%	100%
DeepHit	Recalibrated TS	0%	0%
FineGray	Base Model	20%	100%
FineGray	Recalibrated AJ	0%	100%
FineGray	Recalibrated TS	0.0 %	100%
RSF	Base Model	80%	20%
RSF	Recalibrated AJ	0%	100%
RSF	Recalibrated TS	0%	60%
SurvTRACE	Base Model	0%	40%
SurvTRACE	Recalibrated AJ	0%	100%
SurvTRACE	Recalibrated TS	20%	100%
SurvivalBoost	Base Model	100%	0%
SurvivalBoost	Recalibrated AJ	0%	100%
SurvivalBoost	Recalibrated TS	100%	0%

Table 10: **SEER 100k training points**: CR-$\hat{D}$-calibration per event. We show the $\hat{D}_2^{CR}$.

Model	Event 1	Event 2	Event 3
AalenJohansen	0.02 ± 0.01	0.03 ± 0.01	0.02 ± 0.01
DeepHit	0.77 ± 0.41	0.72 ± 0.39	0.79 ± 0.43
FineGray	0.04 ± 0.02	0.14 ± 0.05	0.1 ± 0.04
RSF	0.03 ± 0.0	0.08 ± 0.02	0.04 ± 0.02
SurvTRACE	0.11 ± 0.03	0.19 ± 0.04	0.16 ± 0.02
SurvivalBoost	0.04 ± 0.01	0.06 ± 0.01	0.05 ± 0.01

I DATASETS DESCRIPTION

For all datasets, the distributions of event types are summarized in Figure 34.

For the synthetic dataset, we consider a competing-risks setting with three causes and covariates $X = (\Lambda_1, \Lambda_3, S_1, S_2, S_3)$ with independent components drawn from uniform distributions:

- $\Lambda_1 \sim \mathcal{U}(0.4, 0.9)$,
- $\Lambda_2 = 1$ (constant),
- $\Lambda_3 \sim \mathcal{U}(1.2, 3)$,
- $S_1 \sim \mathcal{U}(1, 20)$,
- $S_2 \sim \mathcal{U}(1, 10)$ and
- $S_3 \sim \mathcal{U}(1.5, 5)$.

Conditionally on X, we generate three event times T_1, T_2, T_3 and a censoring time C as independent positive random variables with Weibull distributions with

- $\mathbb{P}(T_k \leq t \mid X) = 1 - \exp\big(-(t/\Lambda_k)^{S_k}\big)$ for $k = 1, 2, 3$, and
- $\mathbb{P}(C \leq t \mid X) = 1 - \exp\big(-t/\Lambda_0\big)$ with $\Lambda_0 = 1.5 \cdot \mathbb{E}[T^*]$,

where $T^* = \min(T_1, T_2, T_3)$ and $\Delta^* = \arg\min_k T_k$.

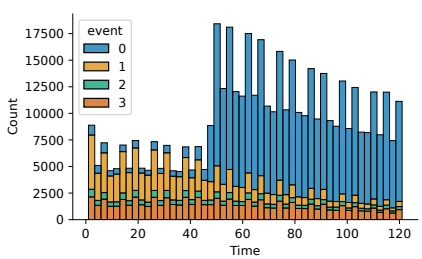

Figure 31: SEER

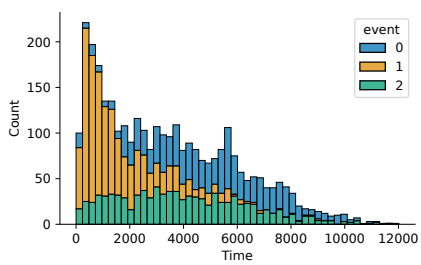

Figure 32: METABRIC

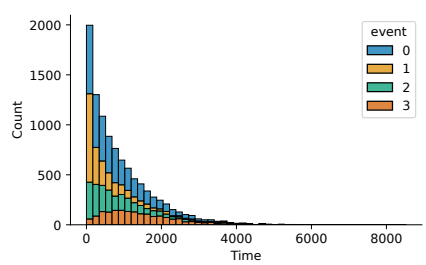

Figure 33: Synthetic Dataset

Figure 34: Histograms of the distribution of the events for each dataset

J NOTATIONS

Table 11: Notations used

Maths Symbol	Domain	Description
K	$\mathbb{N}^*$	number of competing events (events of interest)
$\mathbf{X}$	$\mathbb{R}^d$	the d covariates the individual
T_k^*	$\mathbb{R}_+$	random variable of the time-to-event for event k
C	$\mathbb{R}_+$	random variable of the time-to-censoring
T^*	$\mathbb{R}_+$	$\min(T_1^*, T_2^*, ..., T_K^*)$
T	$\mathbb{R}_+$	$\min(T, C)$
Δ^*	$[1, K]$	$\arg\min_{k\in[1,K]}(T_k^*)$
Δ	$[0, K]$	$\arg\min(C, T_1^*, T_2^*, ..., T_K^*)$
$\mu^\star$		Distribution of $(\mathbf{X}, (T^\star, \Delta^\star)))$
μ		Distribution of $(\mathbf{X}, (T, \Delta)))$
S	$\mathbb{R}^d \longrightarrow \mathbb{R}_+$	Survival function
F	$\mathbb{R}^d \longrightarrow \mathbb{R}_+$	Cumulative Incidence Function
$\mathcal{D}$		Dataset considered
n	$\mathbb{N}^*$	number of individuals in the training set
m	$\mathbb{N}^*$	number of individuals in the calibration set
$\mathbf{x}$		individuals observed
t	$\mathbb{R}_+^n$	time-to-event observed
δ	$[0, K]$	event observed, 0 indicates censoring
B	$\mathbb{N}^*$	Number of buckets
$I_k \stackrel{\text{def}}{=} [l_k, l_{k+1}]$		k^{th} interval
$F_T \stackrel{\text{def}}{=} F_k(T\|X)$	$\mathbb{R}_+$	
$F_{./\infty} \stackrel{\text{def}}{=} F_k(.\|X)$	$\mathbb{R}_+$	
PI		Plug-in estimator