Conformalized Survival Distributions:
A Generic Post-Process to Increase Calibration

Shi-ang Qi [1]   **Yakun Yu** [2]   **Russell Greiner** [1 3]

Abstract

Discrimination and calibration represent two important properties of survival analysis, with the former assessing the model's ability to accurately rank subjects and the latter evaluating the alignment of predicted outcomes with actual events. With their distinct nature, it is hard for survival models to simultaneously optimize both of them especially as many previous results found improving calibration tends to diminish discrimination performance. This paper introduces a novel approach utilizing *conformal regression* that can improve a model's calibration without degrading discrimination. We provide theoretical guarantees for the above claim, and rigorously validate the efficiency of our approach across 11 real-world datasets, showcasing its practical applicability and robustness in diverse scenarios.

1. Introduction

Survival analysis is a useful tool, especially in biomedical applications. Given a patient's covariates, we focus on survival models that can provide a time-to-event distribution (aka, survival distribution). Accurate estimates of these times-to-event can help clinicians prescribe personalized treatment strategies or allocate resources for the healthcare system. Learning such survival models is challenging as many training datasets include *censored* subjects – instances where we only know a lower bound of their time-to-event.

Many survival models seek to optimize discriminative performance, which prioritizes the ability to accurately rank subjects (Harrell Jr et al., 1996) in terms of their survival probabilities or time-to-event times. Discrimination mea-

[1]Computing Science, University of Alberta, Edmonton, Canada [2]Eletrical Computer Engineeering, University of Alberta, Edmonton, Canada [3]Alberta Machine Intelligence Institute, Edmonton, Canada. Correspondence to: Shi-ang Qi <shiang@ualberta.ca>, Russell Greiner <rgreiner@ualberta.ca>.

surements are useful if, for example we want to rank the severity levels of the patients on the waiting list so that we can prioritize the most severe to receive the earliest surgery.

While discrimination is helpful, this evaluation might not assess some important aspects of a prediction. For example, imagine a model estimates Mr. Smith's chance of survival as 90% within 1 year, 50% in 2 years, and 30% in 10 years. Discrimination does not reflect how well the model's predicted survival probabilities match the actual distribution of observations across a population – a concept known as calibration (D'Agostino & Nam, 2003; Haider et al., 2020). There are two main types of calibration measurements: distribution calibration (D-cal) and single-time calibration (1-cal), which offer distinct but related insights (see Section 2 for illustration and comparison). Calibrated predictions are essential as they ensure the reliability of survival probabilities, which is fundamental for both individual assessments and population-level decision-making strategies. By evaluating how well a model corresponds to the world, calibration aids in aligning model predictions closely with real-world outcomes. The importance of calibration is emphasized across various fields – *e.g.*, forecasting (DeGroot & Fienberg, 1983), Bayesian analysis (Dawid, 1982), and risk assessment (Pepe & Janes, 2013).

Many research projects attempt to enhance calibration performance by incorporating a calibration-related loss function into the optimization process. However, this approach presents several challenges during the model's training phase, including increased difficulty in model convergence and the need to tune more hyperparameters. More importantly, many of these approaches found that improving calibration tends to diminish discriminative performance (Goldstein et al., 2020; Chapfuwa et al., 2020b). This phenomenon, known as the *discrimination-calibration trade-off* (Kamran & Wiens, 2021), limits the applicability of these methods, particularly in contexts where discriminative accuracy remains important.

Contributions: To address this challenge, this paper introduces the *Conformalized Survival Distribution (CSD)* framework – a plug-in post-processing method designed to enhance the calibration of a survival distribution model, without compromising its discriminative power. This

model-agnostic framework is compatible with all statistical and machine-learning survival models as long as they can predict individual survival curves (ISD). These learned models first discretize predicted survival curves into percentile times (PCTs, defined in Section 3), followed by adjusting these PCTs using conformal regression (Romano et al., 2019). The principal contributions of our study are:

1. Developing the CSD framework, adapting conformal regression to censored data for making calibrated predictions.
2. Demonstrating theoretically and empirically that CSD not only enhances calibration but also preserves discriminative performance.
3. Empirically comparing CSD with other calibration-improving methods, showing that CSD is superior over many metrics.
4. Connecting the two calibration measures for survival analysis by proving that minimizing D-cal is asymptotically equivalent to minimizing integrated 1-cal across all time points.

From an application point of view, our findings suggest a shift in focus for researchers. Instead of concurrently optimizing discrimination and calibration during the training phase, they can now seek models with superior discriminative abilities and subsequently apply the CSD framework to improve calibration. This approach simplifies the model development process while ensuring robust performance across these key metrics.

Section 2 provides the background and introduces the definition of discrimination and calibration in survival analysis. Section 3 describes the CSD method and provides guarantees that CSD can not only enhance calibration but also preserve discriminative performance. Section 4 presents the extensive empirical analysis across 11 real-world survival datasets and shows the effectiveness of CSD. A Python implementation of CSD is available online at https://github.com/shi-ang/CSD, along with code to replicate all experiments.

2. Discrimination-Calibration Trade-Off

Survival prediction tools learn models that can predict survival for novel instances, from a survival dataset $\mathcal{D} = \{(\boldsymbol{x}_i, t_i, \delta_i)\}_{i=1}^N$, where each triplets corresponds to a subject, where $\boldsymbol{x}_i$ is the observed features for the i-th subject, $t_i \in \mathbb{R}_+$ is the time of the event or censoring, and δ_i is a binary indicator where $\delta_i = 1$ means the subject had the event at time t_i and $\delta_i = 0$ means the subject is right-censored (i.e., the subject has not experienced the event at time t_i). For each subject, there are two potential times of interest: the event time e_i and the censoring time c_i. However, only the earlier of these two times is observable. Thus, we define

$t_i \triangleq \min\{e_i, c_i\}$ and $\delta_i \triangleq \mathbb{1}[e_i \leq c_i]$, where $\mathbb{1}[\cdot]$ is the indicator function. In this study, we make two assumptions:

- *exchangeable*: the order of the subjects do not matter. For example, $\{(\boldsymbol{x}_i, t_i, \delta_i)\}_{i=1}^N$ are drawn i.i.d.
- *conditional independent censoring*: event time is independent of the censoring time, given a description of the subject, *i.e.*, $e_i \perp c_i \mid \boldsymbol{x}_i$.

Our goal is to estimate the individual survival distribution (ISD), $S(t \mid \boldsymbol{x}_i) = \mathbb{P}(e_i > t \mid \boldsymbol{x}_i)$, which represents the probability that the subject with feature $\boldsymbol{x}_i$ will survive beyond time t.

2.1. Discrimination

Discriminative performance measures how accurately a model ranks subjects based on predicted risk scores.[1] It is often measured by the concordance index (C-index) (Harrell Jr et al., 1996), which is the proportion of all comparable subject pairs whose predicted and outcome orders are concordant, defined as

$$\text{C-index}(\{\eta_i\}_{i=1}^N, \mathcal{D}) = \frac{\sum_{i,j} \mathbb{1}[t_i < t_j] \cdot \mathbb{1}[\eta_i > \eta_j] \cdot \delta_i}{\sum_{i,j} \mathbb{1}[t_i < t_j] \cdot \delta_i},$$
(1)

where η_i denotes the model's predicted risk score of i-th subject. The risk score can be defined as the negative of survival probability at a specified time or as the negative of predicted time-to-death. Many survival models aim to directly or partially optimize a C-index-related objective. These include Cox proportional hazards (CoxPH) (Cox, 1972), as well as some recent approaches such as DeepHit (Lee et al., 2018) and Diffsurv (Vauvelle et al., 2023).
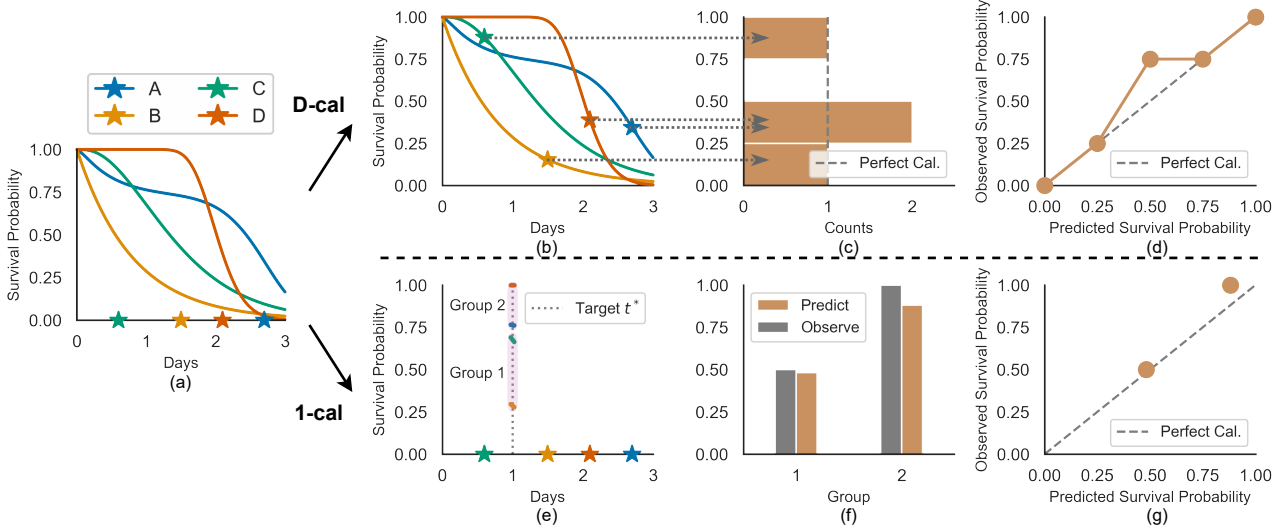
2.2. Calibration

Distribution Calibration Calibration measures how well the prediction fits a set of observations. Consider patient A whose predicted ISD (denoted $\hat{S}(t \mid \boldsymbol{x}_A)$) is showned in Figure 1(a). The predicted median survival time for A (the point at which the curve crosses 50%) is around 2 days. We say this ISD is calibrated if, hypothetically, among 100 similar patients in parallel universes, about 50 experience the event before 2 days, and the remaining 50 after 2 days. In reality, we cannot observe multiple outcomes for the same subject. Consequently, we shift our focus from subject-specific (conditional) calibration to a broader (marginal) calibration across the entire population. This leads to the concept of distribution calibration (D-cal) (Haider et al., 2020): for any probability interval $[a, b] \subset [0, 1]$, let

$$\hat{\mathcal{D}}(a, b) = \{[\boldsymbol{x}_i, t_i, \delta_i = 1] \in \mathcal{D} \mid \hat{S}(t_i \mid \boldsymbol{x}_i) \in [a, b]\}, \quad (2)$$

---

[1] Note this is fundamentally differ from discrimination in other fields such as fairness (Verma & Rubin, 2018).

Figure 1. Comparison of distribution calibration (upper part) and single-time calibration (lower part) using four uncensored patients. The leftmost (a) includes the predicted survival distributions (solid curves) and the true labels (stars) for four patients. **The upper section shows that D-cal process:** (b) obtains the predicted probabilities of each patient's event time; (c) constructs a histogram for these probabilities, where an ideally calibrated model would yield a uniformly shaped histogram; and (d) computes a probability–probability (P-P) plot assessing the histogram's uniformity. **The lower section (1-cal) process:** (e) obtains the predicted probabilities at a target time t^* and groups the patients based on these sorted probabilities; (f) calculates average predicted and observed probabilities for each group, which should show statistical similarity; and (g) computes a P-P plot visualizing the similarity between predicted and observed probabilities.

be the subset of uncensored subjects in $\mathcal{D}$ whose predicted probability at their event time, $S(t_i \mid \boldsymbol{x}_i)$, falls within $[a, b]$. A model is D-calibrated if the proportion of patients $\frac{|\hat{\mathcal{D}}(a,b)|}{|\mathcal{D}|}$ is similar to the length of the interval $b - a$. The upper section of Figure 1 demonstrates a D-cal assessment on 4 uncensored patients. We first gather the set of predicted probabilities at the event times of patients $\{\hat{S}(t_i \mid \boldsymbol{x}_i)\}_i$, and see if this set uniformly distributes across $[0, 1]$. The calibration level can be visualized using the histogram (Figure 1(c)) and P-P plot (Figure 1(d)), or quantified via a χ^2 test. To integrate censored patients, we can "split" each censored patient uniformly to the subsequent probability intervals after the survival probability at censor time $\hat{S}(c_i \mid \boldsymbol{x}_i)$ (Haider et al., 2020), see (10) in Appendix A.2.

Single-Time Calibration Another form of calibration is the single-time calibration (1 cal) (D'Agostino & Nam, 2003), which evaluates the calibration of the predictions at a specific time. Considering the same patient A and its ISD, we can obtain the predicted survival probability at 1 day, *e.g.*, $\hat{S}(t^* = 1 \mid \boldsymbol{x}_A) = 0.99$. This prediction implies that for 100 similar patients, around 99 are expected to survive beyond 1 day. To formally calculate 1-cal at t^*, we first sort the predicted probabilities at t^* for all patients and group them into K groups, $G_1, \ldots, G_K$, as shown in Figure 1(e) (using $K = 2$). Each group G_k's observed survival probability at t^* is determined using the Kaplan & Meier (1958) (KM) estimation, denoted as $S_{\mathrm{KM}(G_k)}(t^*)$ (not shown in the

figure).[2] We then compare the predicted survival probabilities, $\mathbb{E}_{\boldsymbol{x}_i \sim G_k}[\hat{S}(t^* \mid \boldsymbol{x}_i)]$, with the KM-estimated observed probabilities. The 1-cal assessment can be visualized using a histogram (Figure 1(f)), P-P plot (Figure 1(g)) or statistically evaluated via the Hosmer-Lemeshow test (Hosmer & Lemesbow, 1980).

KM Calibration 1-cal, however, is limited to a single time point, whereas ISDs span all future times. To address this, Chapfuwa et al. (2020b) proposed KM calibration (KM-cal), which compares the mean ISD curve for all subjects $\mathbb{E}_{\boldsymbol{x}_i \sim \mathcal{D}}[\hat{S}(t \mid \boldsymbol{x}_i)]$, against the overall KM survival curve. The closer these two curves align, the more KM-calibrated the model is considered. This method essentially integrates 1-cal across all time points, using only one group ($K = 1$) at each time. Appendix A.2 provides further details and illustrations.

Remark While calibration is clearly a desired property, it is important to recognize that focusing solely on calibration can be misleading, *e.g.*, while the KM estimator is robustly calibrated, it cannot discriminate subjects as it provides identical predictions for all, meaning it does not discriminate any pairs of subjects. In practice, the goal should be to identify a model that not only demonstrates good calibration

---

[2]When there is no censorship, it is essentially the proportion of instances experiencing the event over time.

but also possesses strong discriminative power.

2.3. Calibrated Survival Predictions

Several approaches have been developed to enhance the calibration of ISDs, usually by integrating a calibration-related loss function into the objective function. For example, models optimized for the continuous rank probability score (CRPS) (Avati et al., 2020) can achieve good performance on 1-cal. Goldstein et al. (2020) introduces X-cal to the objective function, which is a differentiable approximation of D-cal. Similarly, Chapfuwa et al. (2020b) incorporated the KM-cal term into the loss function, an approach called survival function matching (SFM).

However, discrimination and calibration serve distinct purposes – *i.e.*, KM is well calibrated but has the lowest possible C-index, whereas models with perfect C-index may have poor calibration (see Appendix A.3 for detailed explanation). When optimizing for multiple distinct objectives simultaneously, a model is prone to be dominated by one or several tasks, leading to potentially compromising performance in others (Tang et al., 2020). In fact, many (Goldstein et al., 2020; Chapfuwa et al., 2020b; Kamran & Wiens, 2021) found that improving the model's calibration performance will often harm the discrimination performance, sometimes significantly; see also our reproduced results in Figure 5 and Section 4.4.

In this paper, we introduce the CSD framework, which aims to reconcile this trade-off by disentangling calibration from discrimination in the optimization process. In particular, this approach means the learner can focus on achieving optimal discrimination without considering the calibration. Afterward, a post-processing CSD step will modify that model, to produce a new model with the same discrimination, but improved calibration. Both theoretical and empirical evidence suggest that this post-processing approach retains the model's initial discriminative power while achieving calibration performance comparable to the KM estimator.

The conformalized survival analysis (CSA) proposed by Candès et al. (2023) bears a superficial resemblance to our method. However, despite the similar name, our methods are fundamentally different. First and most importantly, their CSA is applicable only to datasets with known censored times for all subjects, including those who have experienced an event (Type-I censoring) – which is not available in many real-world situations. As right-censoring is a strict generalization of Type-I censoring, our CSD tackles a survival prediction task that is more general and realistic than CSA. Secondly, CSA focuses solely on calibrating 90% quantile predictions, which is less complex than our task of calibrating the entire ISD curve. Thirdly, CSA's real-world experiments did not account for censoring in the calibration calculation. Lastly, their study did not present

the discrimination performance of the final model.[3]

3. Conformalized Survival Distribution (CSD)

We now describe how Conformal Survival Distribution (CSD) modifies an ISD prediction in a way that can improve the calibration while maintaining discrimination. For simplicity, Section 3.1 will initially ignore the censored subjects. Section 3.2 will then integrate censored data into the CSD framework. Section 3.3 will discuss practical considerations and extensions of CSD.

3.1. Overview

Under the *exchangeable* assumption, the CSD begins by splitting the dataset into two subsets: a training set $\mathcal{D}_{\text{train}}$ and a conformal set $\mathcal{D}_{\text{con}}$. Then, a given survival prediction algorithm will learn a model $\mathcal{M}$ from $\mathcal{D}_{\text{train}}$. Then we apply this model to $\mathcal{D}_{\text{con}}$ to obtain initial predictions, *e.g.*, Figure 1(a) shows the ISDs predicted by the model for the four patients $\{A, B, C, D\}$.

In the next step, we discretize the ISD predictions at some specific percentile levels – Figure 2(a) shows 3 levels. For each patient $\boldsymbol{x}_i$, the corresponding times for these discretized percentile levels are defined as:

$$\hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i) \, = \, \inf\{\, t : \hat{S}_{\mathcal{M}}(t \mid \boldsymbol{x}_i) \leq \rho \,\} = \, \hat{S}_{\mathcal{M}}^{-1}(\rho \mid \boldsymbol{x}_i),$$

where $\hat{S}_{\mathcal{M}}^{-1}(\mathbb{P} = \rho \mid \boldsymbol{x}_i)$ is the inverse function of model's predicted ISD $\hat{S}_{\mathcal{M}}(t \mid \boldsymbol{x}_i)$, and ρ is the ρ-th survival percentile.[4] We refer to this $\hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i)$ as the predicted percentile time (PCT), which is represented by the hollow points in Figure 2. The motivation of the discretization is to facilitate the connection between the ISD predictions and D-cal calculation. Specifically, PCT can form a prediction interval with the maximum time, *i.e.*, $[\hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i), \infty]$. This interval indicates that there is ρ probability that the actual event time for the subject will fall within this range. Therefore, $\forall \, \rho \in [0, 1]$, if a model's predicted PCTs satisfy

$$\mathbb{P}(\, t_i \in [\hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i), \infty]\,) \, = \, \frac{|\hat{\mathcal{D}}(0, \rho)|}{|\mathcal{D}|} \, \approx \, \rho, \quad (3)$$

we can say that this model is D-calibrated, where $\hat{\mathcal{D}}(0, \rho)$ is calculated using (2).

Now we can adapt conformal regression (Romano et al., 2019) after discretizing the ISDs. This approach allows

---

[3]Due to their exclusive focus on calibration, they use only half of the dataset for training the model, leaving the other half to the conformalization; this will diminish the model's discrimination.

[4]Note that the ρ in our paper is defined slightly different from the conformal regression (Romano et al., 2019; Candès et al., 2023): our ρ represents the survival percentile, while in their papers α is the quantile for the cumulative density function (CDF). That means our defined ρ is complementary to theirs, $\rho = 1 - \alpha$.
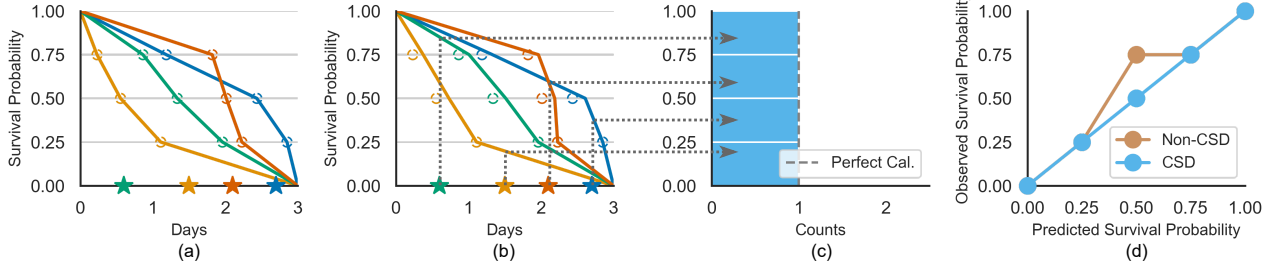
Figure 2. Example of using Conformal Survival Distribution (CSD) to make the prediction D-calibrated, using the same patients and predictions shown in Figure 1. (a) Discretize the predicted survival distributions at three percentile levels (25%, 50%, 75%); (b) Generate the new ISD by adjusting the PCTs, where the hollow points are the old PCTs; (c) Calculate the D-cal histogram using the adjusted ISDs; (d) P-P plot comparing the D-cal level between non-CSD and CSD predictions.

us to compute conformity scores that quantify the error made by PCT (the horizontal distance between the stars and hollow points in Figure 2(a)):

$$s_{j,\mathcal{M}}(\rho) = \hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_j) - t_j, \tag{4}$$

and

$$\mathcal{S}_{\mathcal{M}}(\rho) = \{ s_{j,\mathcal{M}}(\rho) \}_{j=1}^{|\mathcal{D}_{con}|},$$

for each subject j in the conformal set. To interpret the conformity score: high $s_{j,\mathcal{M}}$ represents a poor agreement between the predicted interval and the true event time. If t_j is outside the predicted interval, then s_j will be positive, which can be considered the magnitude of the error incurred by this mistake. If t_j correctly belongs to this interval, then s_j will be nonpositive.

Finally, we can adjust the PCT using the following equation:

$$\hat{q}'_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i) = \hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i) - \text{Quantile}\left[\,\rho; \mathcal{S}_{\mathcal{M}}(\rho)\,\right], \tag{5}$$

where $\text{Quantile}\left[\rho; \mathcal{S}_{\mathcal{M}}(\rho)\right]$ denotes the $\frac{\lceil \rho(|\mathcal{D}_{con}|+1) \rceil}{|\mathcal{D}_{con}|}$-th empirical quantile of $\mathcal{S}_{\mathcal{M}}(\rho)$.[5] This calculation ensures that, for every percentile ρ, there is only $\rho\%$ of the subjects got included in their post-adjust prediction intervals $[\hat{q}'_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i), \infty]$. The post-adjust PCTs are shown in Figure 2(b), where we can see that patient D (red), whose probability at event time was originally located in $[0.25, 0.5]$, now moves to $[0.5, 0.75]$, making the histogram (c) and the P-P plot (d) aligned with the perfect calibration line. Note that all 4 ISDs were shifted, however, the other 3 patients stayed within their original percentile intervals. Regarding the C-index, we can see that the *relative order* of the predicted median survival times – at the 50% percentile intersection – remains consistent before and after CSD.

We now prove that this generic CSD framework satisfied our proposed properties: (1) it does not diminish the discrimination performance; and (2) it improves the calibration performance (both D-cal and KM-cal).

---

[5]This is essentially the ρ-th quantile, with a small correction.

Theorem 3.1. *Applying the CSD adjustment to the percentile predictions does not affect the C-index of the model, regardless of whether to use the negative of the median or of the mean survival times as the predicted risk scores.*

Proof sketch. For median survival times, this obviously holds because the term $\text{Quantile}[\rho; \mathcal{S}_{\mathcal{M}}(\rho)]$ can be considered as a constant (independent) wrt to $\boldsymbol{x}_i$. Therefore, the predicted median survival times $\{\hat{q}'_{\mathcal{M}}(0.5 \mid \boldsymbol{x}_i)\}_i$ will remain in the same order as $\{\hat{q}_{\mathcal{M}}(0.5 \mid \boldsymbol{x}_i)\}_i$. For mean survival times, Appendix C.1 provides the full proof. Note that this theorem is applicable to both uncensored and censored cases (in the next section).

Theorem 3.2. *Under the exchangeable assumption, the percentile predictions constructed by CSD will exhibit exact distribution calibration, s.t., $\forall\, \rho \in [0, 1]$, we can have*

$$\rho \leq \mathbb{P}(\, t_i \in [\hat{q}'_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i), \infty] \mid \boldsymbol{x}_i) \leq \rho + \frac{1}{|\mathcal{D}_{con}| + 1}. \tag{6}$$

Proof sketch. The theorem comes directly from the definition and is inspired by a standard proof in conformal regression (Romano et al., 2019; Angelopoulos et al., 2023). For completeness, Appendix C.2 provides the full proof.

Lemma 3.3. *Under the exchangeable assumption, the percentile predictions constructed by CSD will asymptotically exhibit exact integrated calibration at all time points, which means that the prediction is KM-calibrated.*

Proof sketch. We derive this by first building a connection between D-cal and KM-cal. We prove that D-cal and KM-cal are asymptotically equal, see Appendix C.3. Then, under Theorem 3.2, CSD will asymptotically be KM-calibrated.

3.2. Handling Right-Censoring

3.2.1. UNCENSORED

A naive approach to handle censoring is to discard the censoring subjects in the conformal set and compute the conformity scores as (4). However, under the *conditional indepen-*

dent censoring assumption, it is easy to show that the data distribution with versus without censored subjects can be different, leading to non-calibration for the adjusted ISDs.

3.2.2. DECENSORING

Another option to handle censored subject is to assign a "best-guess" (BG) value to each censored subject using non-parametric methods. After using those BG values as a surrogate for their true event times, we can use (4) to calculate the conformity scores. Here, we consider two methods for calculating these surrogate values.

Margin (Haider et al., 2020) calculates the conditional expectation of the event time given a subject is censored at c_i wrt the KM estimation, *i.e.*, $\mathbb{E}_t[S_{KM}(t \mid t > c_i)]$.

Pseudo-observation (Andersen et al., 2003) (PO) uses a jackknife method to calculate the contribution of a censored subject to the group-level KM survival distribution.

Details of these two methods can be found in the Appendix D.1.

3.2.3. KM-SAMPLING

One problem with both of these decensoring methods is that they provide deterministic BG values as the surrogate times, which means the algorithm should be 100% certain that these censored subjects have the event in that time, which means the algorithm only works when the BG values are all correct.

Therefore, instead of a single BG value, here we provide a *BG distribution* for the event time. We first compute the KM estimation (Kaplan & Meier, 1958) for the training set, $S_{KM}(t) = \prod_{i:t_i \leq t}(1 - \frac{d_i}{n_i})$, as shown by the tan curve in Figure 3. Here t_i is a time where at least one event happened, d_i is the number of events that occur at t_i, and n_i is the count of at-risk individuals (who have not yet had an event) before t_i. Then we can calculate the conditional group-level survival distribution, given a subject is censored at c_j in the conformal set:

$$S_{KM}(t \mid t > c_j) = \min\left\{\frac{S_{KM}(t)}{S_{KM}(c_j)}, 1\right\}. \quad (7)$$

The blue curve in Figure 3 is this conditional distribution, associated with $c_i = 4000$.[6] To incorporate this BG distribution into the conformity score calculation, we can use the idea of sampling, *i.e.*, repeating a censored subject R times, and for each, compute the conformity score by sampling an event time from the BG distribution. Formally, for the r-th

---

[6]If the KM does not reach 0% at the end of the curve, we adopt a linear extrapolation method (Haider et al., 2020) to extend survival curves, using the starting point $[0, 1]$ and ending point $[t_{max}, S_{KM}(t_{max})]$, see the dotted extensions in Figure 3.
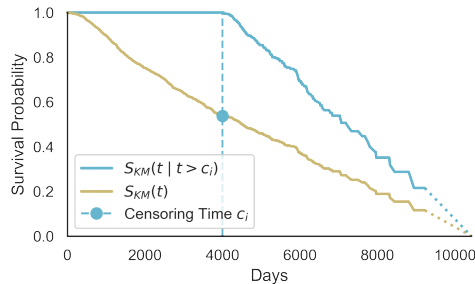

Figure 3. An example of KM sampling from `METABRIC` dataset.

sampling, the conformity score is

$$s_{j,\mathcal{M}}^r(\rho) = \hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_j) - t_j^r, \quad t_j^r \sim S_{KM}(t \mid t > c_j). \quad (8)$$

After each censored subject is sampled multiple times, the number of censored subjects in the conformal set will be significantly increased. Correspondingly, we need to balance the uncensored subjects. Therefore, for each uncensored subject, we need to repeat it R times as well, to maintain the same event-censoring ratio – that is, we also sample R times for each uncensored subject, using the Heaviside function at its event time. In this study, we use $R = 1000$ for sampling.

The pseudo-algorithm for computing the CSD with the KM-sampling process is presented at Algorithm 1 in Appendix D.2.

Remark 3.4. This method depends on the KM model for calculating the BG distributions. Technically, we can use any survival estimator as the prior distribution for the subsequent sampling operation. However, KM is guaranteed to be asymptotically calibrated (both D-cal and KM-cal, see Theorem B.2 and B.1). In contrast, other survival algorithms that predict ISDs (*e.g.*, CoxPH) do not provide this guarantee, meaning they might perform very poorly here. Our empirical observations also align with this theoretical claim. That is, we could not find any ISD model that can consistently exhibit the same level of calibration performance as KM – see the calibration results in Section 4.4 and Figure 9.

Remark 3.5. On the positive side, the KM curve does not make any assumption on the distribution. However, its applicability is limited to datasets with a sufficient number of events for accurate estimation. Our ablation study (Section 4.4) reveals that for small datasets ($N < 1000$), the performance of the KM-sampling method is similar to both decensoring methods. CSD also has other constraints, *e.g.*, it cannot perform the conformalization step in batch mode due to Quantile operation, which means it is computationally inefficient. We provide the computational analysis in Section 4.4 and Appendix F.4.

Remark 3.6. CSD method makes two assumptions: *exchangeability* and *conditional independent censoring*.

Exchangeability is a standard assumption in machine learning. For example, if the average life expectancy is around 70

for the first 100 instances (collected now), then we assume the average will still be around 70 for the next 100 instances, etc. If this assumption is violated, then any model trained on the current dataset may be ineffective and non-calibrated when it is tested on future instances. We have endeavored to uphold this assumption as rigorously as possible during our experiments.

Conditional independent censoring is the fundamental assumption in survival analysis. This assumption underpins the development of many survival models and the KM-sampling method. While this assumption may not always hold in practical scenarios, we intend to include a broad array of real-world datasets to mitigate the impact of any potential assumption violations.

3.3. Dataset Splitting

As mentioned earlier, CSD begins by splitting the data into a training and a conformal set. Previous methods (Romano et al., 2019; Candès et al., 2023) recommend allocating 50% of the data for training and the rest 50% for conformal, ensuring sufficient samples in the calibration set.

However, this policy can be problematic: reducing the training set size by half may cause underfitting of the model, resulting in suboptimal discriminative ability. Since the post-conformal step does not enhance the model's discrimination ability, this method essentially sacrifices some discrimination performance for improved calibration, which is not our intended outcome.

To address this, we explore two alternative policies: (1) Using the validation set, typically required for hyperparameter tuning or early stopping, as the conformal set; (2) Combining the validation and training sets to form a larger conformal set, where the training set would be a subset of the conformal set. For models that do not require a separate validation set, we propose reusing the training set as the conformal set. Section 4.4 discusses the performance of these two policies.

4. Experiments and Results

4.1. Datasets

To evaluate the efficacy of CSD, we trained the non-CSD baselines and their post-CSD competitors in 11 datasets. Table 1 offers an overview of the dataset statistics; see Appendix E.1 for details. We deliberately select a wide range of datasets with small-to-large sample sizes, small-to-large censoring rates, and small-to-large feature-sample ratios.

We split each dataset into a training set (90%) and a testing set (10%) using a stratified splitting procedure that balances both the time t_i and the censor indicator δ_i. For algorithms that require a validation set to tune hyperparameters or to

Table 1. Summary of the datasets. We categorize datasets into *small*, *medium*, and *large* groups based on sample sizes, using thresholds of 1,000 and 10,000 samples. [†]Number of raw features, with (number of features after one-hot encoding) in parentheses.

Dataset	#Sample	%Censored	Max t	#Feature[†]
VALCT	137	6.57%	999	6 (8)
DLBCL	240	42.50%	21.80	7399
PBC	418	61.48%	4,795	17
GBM	595	17.23%	3,881	8 (10)
METABRIC	1,981	55.17%	9218	79
GBSG	2,232	43.23%	87.36	7
NACD	2,396	36.44%	84.30	48
SUPPORT	9,105	31.89%	2,029	26 (31)
SEER-brain	73,703	40.12%	227	10
SEER-liver	82,841	37.57%	227	14
SEER-stomach	100,360	43.40%	227	14

early stop, we partition another balanced 10% validation set from the training set. We run each dataset using 10 different random seeds and compute the mean and 95% confidence interval (CI) on all evaluation metrics across the 10 splits.

4.2. Evaluation Metrics

Our goal is to enhance calibration performance while preserving the existing level of discriminative ability. To test that, we use a discrimination metric (C-index) and two calibration metrics (D-cal and KM-cal) as the primary metrics.

For the C-index, we use the negative value of the predicted median survival time as the risk score in (1). The D-cal metric is calculated by quantifying the mean squared distances between the observed survival probability and the predicted survival probability at each percentile (the distances between the dashed perfect calibration line and the model's solid P-P line in Figure 1(d)). Formally, for a set of percentiles $\mathcal{P} = \{0.1, 0.2, \ldots, 0.9\}$, D-cal is calculated by

$$\text{D-cal} = \frac{1}{|\mathcal{P}|} \sum_{\rho \in \mathcal{P}} \left(\frac{|\hat{\mathcal{D}}_{\text{test}}(0, \rho)|}{|\mathcal{D}_{\text{test}}|} - \rho \right)^2,$$

where $\hat{\mathcal{D}}_{\text{test}}(0, \rho)$ is calculated using (2). This score can also be interpreted as the area between the perfect calibration line (dashed line) and the model's P-P line (solid line) in Figure 1(d). The KM-cal is calculated using

$$\text{KM-cal} = \frac{1}{t_{\max}} \int_{t=0}^{t_{\max}} \left(S_{\text{KM}(\mathcal{D}_{\text{test}})}(t) - \mathbb{E}_{\boldsymbol{x}_i}[\hat{S}(t \mid \boldsymbol{x}_i)] \right)^2 dt,$$

where $t_{\max}$ is the maximum time in the testing set. This score is the integrated discrepancy between the KM curve and the expected model prediction over time.

In addition to the primary metrics, we include two other metrics to assess survival prediction performance: the integrated
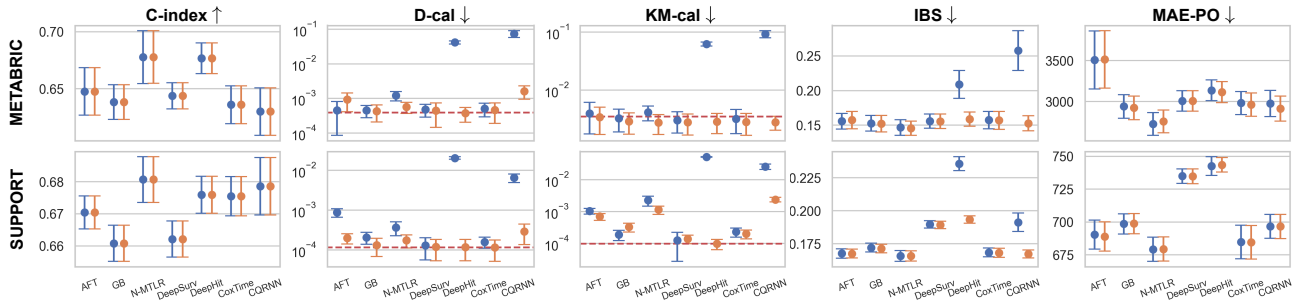
Figure 4. Parts of the empirical results. The error bars represent mean and 95% CI over ten runs, with blue denoting the non-CSD baseline and orange for the CSD-version. The red dashed lines represent the mean calibration performance for KM estimator, serving as an empirical lower-limit. A higher C-index score indicates better performance, whereas lower scores are preferable for the other metrics.

Brier score (IBS) (Graf et al., 1999) measures the predicted probability accuracy across all times; and the mean absolute error with pseudo-observation (MAE-PO) (Qi et al., 2023a) measures the accuracy of predicting the time to an event.

4.3. Models

We compare the results using 7 survival algorithms: Accelerate Failure Time (AFT) (Stute, 1993) with Weibull distribution, Gradient Boosting (GB) (Hothorn et al., 2006) with component-wise least squares as base learner, Deep-Surv (Katzman et al., 2018) (with Breslow extension), Neural Multi-Task Logistic Regression (N-MTLR) (Fotso, 2018), DeepHit (Lee et al., 2018), CoxTime (Kvamme et al., 2019), and censored quantile regression neural networks (CQRNN) (Pearce et al., 2022). Appendix E.3 describes the implementation details and hyperparameter settings.

Additionally, we also include a "dummy" model – KM estimator (Kaplan & Meier, 1958) – which uses the group-level KM for the training set as the same prediction to all test samples. Theoretically, it represents an extreme in our model spectrum, offering minimal discriminative performance (C-index$= 0.5$) while theoretically achieving perfect calibration (Theorems B.1 and B.2). However, as evident in the results (Figure 4 and Appendix F), this model still registers small scores for the D-cal and KM-cal. This is due to the slight distribution-shift caused by the dataset split, even though we have balanced the partition via both $\{t_i\}_{i=1}^N$ and $\{\delta_i\}_{i=1}^N$. Therefore, the calibration metric scores of the KM estimator can serve as the **empirical lower-limits** for the calibration performance.

4.4. Results

We compare 7 baseline models and their post-CSD versions on 11 clinical datasets. Note that for the DLBCL datasets, because the number of features (7399) is significantly greater than the sample size (240), AFT model fails to converge on this dataset. Consequently, we have in

Table 2. Comparative Analysis of CSD performance by counting the number of times none-CSD baselines is better, CSD is better, and ties, over five evaluation metrics. The number in bracket means significantly better ($p < 0.05$ using two-sided t-test).

	C-index	D-cal	KM-cal	IBS	MAE-PO
Non-CSD	3(0)	8(1)	20(7)	12(0)	30(0)
CSD	13(0)	**68(35)**	**56(30)**	**61(14)**	**45(4)**
ties	**60**	0	0	3	1

total $11 \times 7 - 1 = 76$ comparisons. Due to the space limit, the main text only reports the results on two datasets (METABRAC and SURRPORT) in Figure 4. Appendix F presents the complete results.

Discriminative Performance   The first column of Figure 4 shows there are essentially no differences in the C-index between the non-CSD models and their CSD counterparts. Specifically, out of 76 comparisons, 60 show equal performance (ties), and in 13 instances, the CSD models is (insignificantly) better (Table 2). This implies that the CSD framework preserves the discriminative performance in approximately 96% of cases. Moreover, in the remaining 3 cases where non-CSD models perform better, the differences are not statistically significant.

Calibrative Performance   Our results, columns 2 and 3 of Figure 4 and the summary in Table 2, indicate a significant improvement in calibration performance due to the CSD framework. Specifically, the CSD framework demonstrates improvement in 68 cases for D-cal (with 35 significantly better) and 56 cases for KM-cal (with 30 significantly better). However, note that in some instances, baseline models may already reach optimal calibration levels, comparable to the KM estimator (*e.g.*, the AFT model in METABRIC dataset). In such cases, further enhancement of calibration performance is challenging. Nevertheless, the CSD framework consistently performs well by maintaining the same calibration level in these already well-calibrated cases.
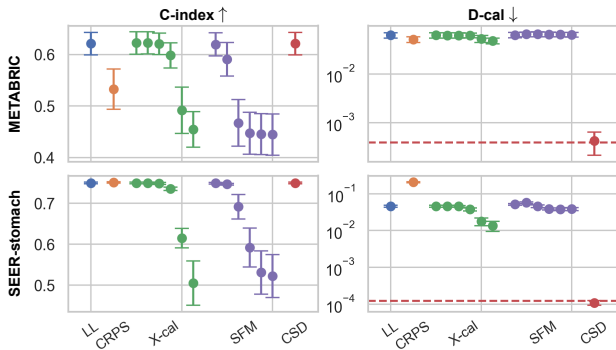
Figure 5. Compare CSD with objective-based methods on a deep log-normal model. The baseline (blue) uses likelihood loss (LL). For X-cal and SFM methods, we gradually increase the weight for the calibration loss. The red dashed lines serve as the empirical lower-limit.

Other Metrics The fourth columns in Figure 4 give the IBS score, which is the expectation of Brier scores (BS) over time. According to DeGroot & Fienberg (1983), BS can be decomposed into a calibration part (1-cal) and a discrimination part (AUC). This implies that IBS, being an integrated form of BS, includes elements of KM-cal (integrated 1-cal). Indeed, the results in Table 2 show that the trends in IBS closely mirror those in KM-cal, with CSD outperforming in 61 (and significantly in 14) of the 76 comparisons (80%).

For MAE-PO, the performance enhancement by CSD is less pronounced compared to other metrics. Nonetheless, CSD still shows improvement in this metric, outperforming 45 (and significantly in 4) out of 76 comparisons (59%).

Compare with Objective-Based Methods In Figure 5, we compare CSD with three objective-based methods previously discussed in Section 2.3: CRPS (Avati et al., 2020), X-cal (Goldstein et al., 2020) and SFM (Chapfuwa et al., 2020b). The experimental procedures and the complete results for all 11 datasets are detailed in Appendix F.2.

Our results on 11 datasets indicate that X-cal can indeed enhance D-cal's performance by increasing the weight assigned to calibration loss. However, X-cal falls short in two ways: (1) it fails to achieve optimal calibration levels comparable to those achieved by KM and CSD (the red dashed line), and (2) it is often accompanied by a notable decrease in the C-index. Specifically, in METABRIC and SEER-stomach, as X-cal improves D-cal, the C-index drops to ≈ 0.5. Regarding CRPS, its beneficial effects on D-cal are less consistent, *e.g.*, it is effective for the METABRIC dataset but not for SEER-stomach. SFM also exhibits inconsistent trends; it is effective for SEER-stomach but not for METABRIC. In cases where CRPS and SFM are effective, they display trends comparable to those of X-cal.

In conclusion, the CSD framework significantly surpasses the performance of these objective-based methods.

Ablation Studies CSD is influenced by three key hyperparameters: (1) how to handle censorship; (2) how to construct a conformal set; and (3) the number of discretized percentiles. While Appendix F.3 provides complete experimental details, results and summaries, the main findings are:

1. For medium to large datasets, the KM sampling method significantly outperforms other approaches.
2. For large datasets, using the validation set alone as the conformal set is typically adequate. However, for smaller or medium datasets, it is beneficial to combine both validation and training sets to form the conformal set.
3. Different numbers of percentiles have minimal impacts on performance.

Computational Complexity The quantile operation in (5) requires storing the conformity scores for all data in the conformal set in advance, followed by computing the quantile amount in a substantial array. This introduces two types of complexity: space and time. In short summary, KM-sampling increases the space complexity by a factor of R and the time complexity by approximately 16 times compared to the simple uncensored method. The complete results can be found in Appendix F.4.

5. Conclusion

There are two crucial aspects in survival prediction: **discrimination** assesses the model's ability to correctly rank subjects, and **calibration** evaluates the alignment of predicted distributions with actual outcomes. Unfortunately, many survival prediction learners encounter a trade-off between discrimination and calibration.

Our study introduces the conformalized survival distribution (CSD), a generic post-processing framework designed to mitigate this trade-off. Inspired by the principle of focusing on one task at a time, CSD disentangles calibration from discrimination during training. Specifically, we first produce a model, by concentrating solely on enhancing discrimination. Post-training, CSD steps in to refine the predicted survival distribution curves through conformal regression, producing new curves that are (in general) more calibrated.

Our empirical evidence shows that CSD successfully maintains discrimination ability on par with baseline models and enhances calibrative performance in terms of both D-cal and KM-cal. Also, CSD's effectiveness extends to two other widely used survival metrics, showing encouraging results.

Acknowledgements

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Andersen, P. K. and Pohar Perme, M. Pseudo-observations in survival analysis. *Statistical methods in medical research*, 19(1):71–99, 2010.

Andersen, P. K., Klein, J. P., and Rosthøj, S. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1): 15–27, 2003.

Angelopoulos, A. N., Bates, S., et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

Avati, A., Duan, T., Zhou, S., Jung, K., Shah, N. H., and Ng, A. Y. Countdown regression: sharp and calibrated survival predictions. In *Uncertainty in Artificial Intelligence*, pp. 145–155. PMLR, 2020.

Binder, N., Gerds, T. A., and Andersen, P. K. Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime data analysis*, 20(2):303–315, 2014.

Breslow, N. E. Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, pp. 45–57, 1975.

Candès, E., Lei, L., and Ren, Z. Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):24–45, 2023.

Chapfuwa, P., Li, C., Mehta, N., Carin, L., and Henao, R. Survival cluster analysis. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 60–68, 2020a.

Chapfuwa, P., Tao, C., Li, C., Khan, I., Chandross, K. J., Pencina, M. J., Carin, L., and Henao, R. Calibration and uncertainty in neural time-to-event modeling. *IEEE transactions on neural networks and learning systems*, 2020b.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125, 2010.

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

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

D'Agostino, R. B. and Nam, B.-H. Evaluation of the performance of survival analysis models: discrimination and calibration measures. *Handbook of statistics*, 23:1–25, 2003.

Davidson-Pilon, C. lifelines, survival analysis in python, January 2024. URL https://doi.org/10.5281/zenodo.10456828.

Dawid, A. P. The well-calibrated bayesian. *Journal of the American statistical Association*, pp. 605–610, 1982.

DeGroot, M. H. and Fienberg, S. E. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Farrokh, M., Qi, S., Kumar, N., and Greiner, R. Effective survival prediction and evaluation for cancer patients. 2024.

Foekens, J. A., Peters, H. A., Look, M. P., Portengen, H., Schmitt, M., Kramer, M. D., Brünner, N., Jänicke, F., Gelder, M. E. M.-v., Henzen-Logmans, S. C., et al. The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients. *Cancer research*, 60 (3):636–643, 2000.

Fotso, S. Deep neural networks for survival analysis based on a multi-task framework. *arXiv preprint arXiv:1801.05512*, 2018.

Goldstein, M., Han, X., Puli, A., Perotte, A., and Ranganath, R. X-cal: Explicit calibration for survival analysis. *Advances in neural information processing systems*, 33:18296–18307, 2020.

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

Graw, F., Gerds, T. A., and Schumacher, M. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis*, 15(2):241–255, 2009.

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

Harrell Jr, F. E., Lee, K. L., and Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

Hosmer, D. W. and Lemesbow, S. Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, 9(10):1043–1069, 1980.

Hosmer, D. W., Lemeshow, S., and May, S. *Applied survival analysis*. John Wiley & Sons, Inc., 2008.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. J. Survival ensembles. *Biostatistics*, 7 (3):355–373, 2006.

Jin, P. Using survival prediction techniques to learn consumer-specific reservation price distributions. Master's thesis, University of Alberta, 2015.

Kalbfleisch, J. D. and Prentice, R. L. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

Kamran, F. and Wiens, J. Estimating calibrated individualized survival curves with deep learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):240–248, May 2021.

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

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18 (1):1–12, 2018.

Kazmierski, M. torchmtlr: flexible and modular implementation of multi-task logistic regression in pytorch., 2020. URL https://github.com/mkazmier/torchmtlr.

Knaus, W. A., Harrell, F. E., Lynn, J., Goldman, L., Phillips, R. S., Connors, A. F., Dawson, N. V., Fulkerson, W. J., Califf, R. M., Desbiens, N., et al. The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*, 122(3):191–203, 1995.

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

Lee, C., Zame, W. R., Yoon, J., and van der Schaar, M. Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Li, Y., Wang, J., Ye, J., and Reddy, C. K. A multi-task learning formulation for survival analysis. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1715–1724, 2016.

National Cancer Institute, DCCPS, Surveillance Research Program. Surveillance, epidemiology, and end results (seer) program seer*stat database: Incidence - seer research data, 18 registries, nov 2020 sub (2000-2018) - linked to county attributes - time dependent (1990-2018) income/rurality, 1969-2019 counties. www.seer.cancer.gov, 04 2015. based on the November 2020 submission.

Pearce, T., Jeong, J.-H., Zhu, J., et al. Censored quantile regression neural networks for distribution-free survival analysis. In *Advances in Neural Information Processing Systems*, 2022.

Pepe, M. and Janes, H. Methods for evaluating prediction performance of biomarkers and tests. In *Risk assessment and evaluation of predictions*, pp. 107–142. Springer, 2013.

Pölsterl, S. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020. URL http://jmlr.org/papers/v21/20-729.html.

Qi, S.-A., Kumar, N., Farrokh, M., Sun, W., Kuan, L.-H., Ranganath, R., Henao, R., and Greiner, R. An effective meaningful way to evaluate survival models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 28244–28276. PMLR, 23–29 Jul 2023a.

Qi, S.-a., Sun, W., and Greiner, R. SurvivalEVAL: A comprehensive open-source python package for evaluating individual survival distributions. In *Proceedings of the AAAI Symposium Series*, volume 2, pp. 453–457, 2023b.

Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Schumacher, M., Bastert, G., Bojar, H., Hübner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, R., and Rauschecker, H. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. *Journal of Clinical Oncology*, 12(10):2086–2093, 1994.

Stute, W. Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis*, 45(1):89–103, 1993.

Tang, H., Liu, J., Zhao, M., and Gong, X. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pp. 269–278, 2020.

Therneau, T. and Grambsch, P. *Modeling Survival Data: Extending The Cox Model*, volume 48. Springer, 01 2000. ISBN 978-1-4419-3161-0. doi: 10.1007/ 978-1-4757-3294-8.

Vauvelle, A., Wild, B., Eils, R., and Denaxas, S. Differentiable sorting for censored time-to-event data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Verma, S. and Rubin, J. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pp. 1–7, 2018.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

Yanagisawa, H. Proper scoring rules for survival analysis. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 39165–39182. PMLR, 23–29 Jul 2023.

Yu, C.-N., Greiner, R., Lin, H.-C., and Baracos, V. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in Neural Information Processing Systems*, 24:1845–1853, 2011.

A. Further Discussion on Discrimination and Calibration Trade-Off

A.1. More about Discrimination

This appendix delves deeper into the Concordance Index (C-index), which gauges the model's ability to accurately rank subjects according to their risk levels. We previously introduced the concept and formal definition of the C-index in Section 2.1. To enhance comprehension, we now present a visual example to illustrate the calculation of the C-index.
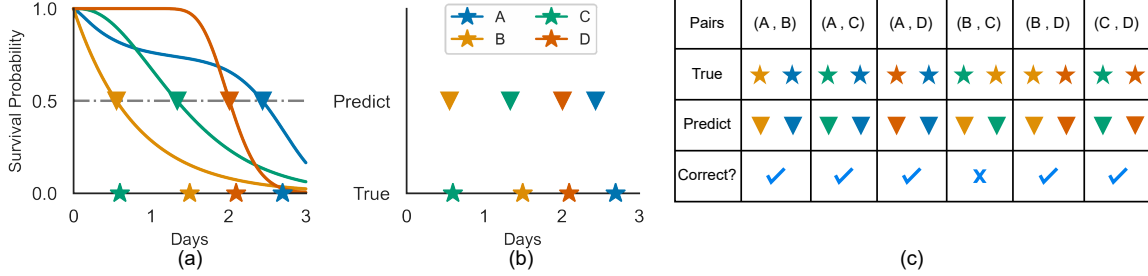

Figure 6. Illustration of concordance index (C-index), using the same patients as Figure 1. The stars denotes the true times, and the triangles denotes predicted median survival times.

Consider Figure 6, which depicts the true outcomes (indicated by stars) and the individual survival distribution (ISD) predictions (represented by curves) for four patients, based on the same example provided in Figure 1. Recall that predicted risk scores are typically negative of the predicted survival times (either median or mean). For simplicity and clearer visualization, we define risk scores as the negative median survival time, marked by the intersection of the curves with the 50% dashed horizontal line (triangles), as illustrated in Figure 6(a).

Following the acquisition of the predicted median survival times, $\hat{t}_i^{(\text{median})}$, we compare their order with that of the actual event times. Consequently, we can reformulate the equation for the C-index as follows:

$$\text{C-index}(\, \{\, \hat{t}_i^{(\text{median})} \,\}_{i=1}^N, \mathcal{D} \,) \quad = \quad \frac{\sum_{i,j} \mathbb{1}[\, t_i < t_j \,] \cdot \mathbb{1}[\, \hat{t}_i^{(\text{median})} < \hat{t}_j^{(\text{median})} \,] \cdot \delta_i}{\sum_{i,j} \mathbb{1}[\, t_i < t_j \,] \cdot \delta_i} \,. \tag{9}$$

This equation can also be used for computing the C-index for the predicted mean survival times, simply by replacing the $\hat{t}_i^{(\text{median})}$ to $\hat{t}_i^{(\text{mean})}$. In Figure 6, out of all $\binom{4}{2} = 6$ possible pairs formed from four patients, only the pair (B, C) is incorrectly predicted, as shown by the yellow and green triangles. Thus, the C-index score for this model prediction is $\frac{5}{6} \approx 83\%$.

To account for censored subjects, we do a modification by identifying *comparable* pairs and examining the accuracy of the predicted order of those pairs. A pair is deemed comparable iff the ordering of their events can be unequivocally determined. Suppose that patient D (indicated in red) is censored, with the red star (at time 2.1) marking the censored time, while patients A, B, and C are fully observed. Among the pairs involving patient D, namely (A, D), (B, D) and (C, D), the pairs (B, D) and (C, D) remain comparable since it is evident that the event of D occurs after the events of B and C. However, the pair (A, D) becomes non-comparable because D's censored time precedes A's event time (at 2.5), leaving the true sequence of events unclear. Consequently, this pair is excluded from both the numerator and denominator in the C-index calculation as per equation (9), resulting in a revised C-index score of $\frac{4}{5} = 80\%$.

A.2. More about Calibration

Distribution Calibration   D-cal examines the calibration ability across the entire range of predictions made by an individual survival distribution ISD model. We can only trust a model's prediction if the ISD predictions are D-calibrated. As we discussed in Section 2.2, for datasets without censorship, the D-calibration of the predictions can be evaluated by examining subsets of data within any probability interval $[a, b] \subset [0, 1]$. These subsets are defined as:

$$\hat{\mathcal{D}}(a, b) \quad = \quad \{\, [\boldsymbol{x}_i, t_i, \delta_i = 1] \in \mathcal{D} \mid \hat{S}(t_i \mid \boldsymbol{x}_i) \in [a, b] \,\},$$

where the proportion of observations within $\hat{\mathcal{D}}(a,b)$ relative to the total dataset size $|\mathcal{D}|$ should reflect the length of the interval, *i.e.*, $\frac{|\hat{\mathcal{D}}(a,b)|}{|\mathcal{D}|} = b - a$. Formally, the size of such subsets is given by:

$$|\hat{\mathcal{D}}(a,b)| \quad = \quad \sum_{i=1}^{N} \delta_i \cdot \mathbb{1}\left[\hat{S}(t_i \mid \boldsymbol{x}_i) \in [a,b] \right].$$

To extend this analysis for censored observations, consider a censored patient with a predicted survival probability at the patient's censoring time $\hat{S}(c_i \mid \boldsymbol{x}_i)$ (where $c_i = t_i$ if $\delta_i = 0$). Given that censoring time provides a lower bound on the actual event time, it follows that $\hat{S}(c_i \mid \boldsymbol{x}_i) > \hat{S}(e_i \mid \boldsymbol{x}_i)$[7]. Therefore, we can estimate the likelihood of the survival probability at the actual event time falling within the interval $[a,b]$, given we know the survival probability at the censor time, such as $\mathbb{P}(\hat{S}(e_i \mid \boldsymbol{x}_i) \in [a,b] \mid \hat{S}(e_i \mid \boldsymbol{x}_i) < \hat{S}(c_i \mid \boldsymbol{x}_i))$.

For this purpose, we consider three scenarios:

- **Case 1:** $a < b \leq \hat{S}(c_i \mid \boldsymbol{x}_i)$

$$\mathbb{P}\left(\hat{S}(e_i \mid \boldsymbol{x}_i) \in [a,b] \,\middle|\, \hat{S}(e_i \mid \boldsymbol{x}_i) < \hat{S}(c_i \mid \boldsymbol{x}_i) \right) \quad = \quad \frac{\mathbb{P}\left(a < \hat{S}(e_i \mid \boldsymbol{x}_i) < b, \; \hat{S}(e_i \mid \boldsymbol{x}_i) < \hat{S}(c_i \mid \boldsymbol{x}_i) \right)}{\mathbb{P}\left(\hat{S}(e_i \mid \boldsymbol{x}_i) < \hat{S}(c_i \mid \boldsymbol{x}_i) \right)}$$

$$= \quad \frac{\mathbb{P}\left(a < \hat{S}(e_i \mid \boldsymbol{x}_i) < b \right)}{\mathbb{P}\left(\hat{S}(e_i \mid \boldsymbol{x}_i) < \hat{S}(c_i \mid \boldsymbol{x}_i) \right)}$$

$$= \quad \frac{b - a}{\hat{S}(c_i \mid \boldsymbol{x}_i)};$$

- **Case 2:** $a \leq \hat{S}(c_i \mid \boldsymbol{x}_i) < b$

$$\mathbb{P}\left(\hat{S}(e_i \mid \boldsymbol{x}_i) \in [a,b] \,\middle|\, \hat{S}(e_i \mid \boldsymbol{x}_i) < \hat{S}(c_i \mid \boldsymbol{x}_i) \right) \quad = \quad \frac{\mathbb{P}\left(a < \hat{S}(e_i \mid \boldsymbol{x}_i) < \hat{S}(c_i \mid \boldsymbol{x}_i) \right)}{\mathbb{P}\left(\hat{S}(e_i \mid \boldsymbol{x}_i) < \hat{S}(c_i \mid \boldsymbol{x}_i) \right)}$$

$$= \quad \frac{\hat{S}(c_i \mid \boldsymbol{x}_i) - a}{\hat{S}(c_i \mid \boldsymbol{x}_i)};$$

- **Case 3:** $\hat{S}(c_i \mid \boldsymbol{x}_i) < a < b$

$$\mathbb{P}\left(\hat{S}(e_i \mid \boldsymbol{x}_i) \in [a,b] \,\middle|\, \hat{S}(e_i \mid \boldsymbol{x}_i) < \hat{S}(c_i \mid \boldsymbol{x}_i) \right) \quad = \quad \frac{\mathbb{P}(\varnothing)}{\mathbb{P}\left(\hat{S}(e_i \mid \boldsymbol{x}_i) < \hat{S}(c_i \mid \boldsymbol{x}_i) \right)}$$

$$= \quad 0.$$

By integrating these scenarios, we can define the adjusted subset size $\hat{\mathcal{D}}(a,b)$, accounting for both uncensored and censored patients, as follows:

$$|\hat{\mathcal{D}}(a,b)| = \sum_{i=1}^{N} \delta_i \cdot \mathbb{1}\left[\hat{S}(t_i \mid \boldsymbol{x}_i) \in [a,b] \right] + (1-\delta_i) \left(\frac{\left(\hat{S}(t_i \mid \boldsymbol{x}_i) - a \right) \mathbb{1}\left[\hat{S}(t_i \mid \boldsymbol{x}_i) \in [a,b] \right]}{\hat{S}(t_i \mid \boldsymbol{x}_i)} + \frac{(b-a)\mathbb{1}\left[\hat{S}(t_i \mid \boldsymbol{x}_i) \geq b \right]}{\hat{S}(t_i \mid \boldsymbol{x}_i)} \right).$$

(10)

---

[7]This is because $c_i < e_i$ for $\delta_i = 0$, and the predicted ISD curve $\hat{S}(t \mid \boldsymbol{x}_i)$ must be monotonically decreasing as the time t increases.

Single-time Calibration 1-cal is a goodness of fit test to evaluate the calibration ability of risk predictions at a specific time. To perform this analysis at a chosen time point, denoted t^*, we begin by sorting the predicted probabilities for all patients at t^* in order. These predictions are then divided into K distinct groups: $G_1, \ldots, G_k, \ldots, G_K$. Within each group, we calculate the expected number of events using the predicted probabilities, $\mathbb{E}_{\boldsymbol{x}_i \sim G_k}[\hat{S}(t^* \mid \boldsymbol{x}_i)]$, and this is compared with the actual observed event rate, which is estimated using the KM method (because KM is an unbiased estimator for the population survival distribution). To assess the statistical similarity between the expected and observed rates of events, we apply the Hosmer-Lemeshow test (Hosmer & Lemesbow, 1980), based on its test statistic:

$$
\begin{aligned}
\hat{\mathrm{HL}}(t^*, \mathcal{D}) &= \sum_{k=1}^{K} \frac{(O_k - n_k \bar{p}_k)^2}{n_k \bar{p}_k (1 - \bar{p}_k)} \\
&= \sum_{k=1}^{K} \frac{\left(|G_k| \cdot S_{\mathrm{KM}(G_k)}(t^*) - \sum_{\boldsymbol{x}_i \in G_k} \hat{S}(t^* \mid \boldsymbol{x}_i) \right)^2}{\sum_{\boldsymbol{x}_i \in G_k} \hat{S}(t^* \mid \boldsymbol{x}_i) \left(1 - \frac{1}{|G_k|} \sum_{\boldsymbol{x}_i \in G_k} \hat{S}(t^* \mid \boldsymbol{x}_i) \right)}.
\end{aligned}
\tag{11}
$$

where, n_k is the total number of patients in the k-th group, O_k is the observed number of patients who experienced the event before t^* in the k-th group, and $\bar{p}_k$ is the average predicted survival probability for the patients in the k-th group.

KM Calibration Chapfuwa et al. (2020b) first proposed KM calibration as an alternation of 1-cal and D-cal. The underlying principle of this approach is to ensure that the average predicted ISD curve for the test set is closely aligned with the KM curve for the testset.

Originally, KM calibration involved a straightforward visual comparison of the two curves. Later, Yanagisawa (2023) refined this method by employing the Kullback-Leibler (KL) divergence to quantify the difference between the two distributions, as expressed in the following equation:

$$
d_{\mathrm{KM\text{-}cal}} = D_{KL} \left(S_{\mathrm{KM}\mathcal{D}_{\mathrm{test}}}(t) \, \big\| \, \mathbb{E}_{\boldsymbol{x}_i \sim \mathcal{D}_{\mathrm{test}}} \left[\hat{S}(t \mid \boldsymbol{x}_i) \right] \right)
$$

However, KL-divergence comes with notable drawbacks: (1) it does not satisfy the properties of metrics, such as symmetric and triangle inequality; (2) it can become undefined if the average predicted ISD, $\mathbb{E}_{\boldsymbol{x}_i \sim \mathcal{D}_{\mathrm{test}}} \left[\hat{S}(t \mid \boldsymbol{x}_i) \right]$, goes to zero for some t; (3) it is sensitive to outliers if the average predicted ISD has a long tail at the end, etc.

To address these issues, our research calculates KM-cal as the integrated squared difference between the two curves:

$$
\mathrm{KM\text{-}cal} = \frac{1}{t_{\max}} \int_{t=0}^{t_{\max}} \left(S_{\mathrm{KM}(\mathcal{D}_{\mathrm{test}})}(t) - \mathbb{E}_{\boldsymbol{x}_i \sim \mathcal{D}_{\mathrm{test}}} \left[\hat{S}(t \mid \boldsymbol{x}_i) \right] \right)^2 dt,
$$

In practice, we first calculate the squared errors at all unique time points within the test set, followed by the application of the trapezoidal rule to compute the integral.

Moreover, this study establishes, for the first time, a link between KM-cal and earlier calibration metrics by demonstrating that KM calibration is equivalent to an unnormalized, integrated single-time calibration across all time points for a single group ($K = 1$), as shown below:

Proposition A.1. *KM calibration is equivalent to unnormalized integrated single-time calibration across all time points, for one group ($K = 1$).*

Proof. By calculating the Hosmer-Lemeshow test statistic for $K = 1$, we find

$$
\begin{aligned}
\hat{\mathrm{HL}}(t^*, \mathcal{D}) &= \frac{\left(|\mathcal{D}| \cdot S_{\mathrm{KM}(\mathcal{D})}(t^*) - \sum_{i=1}^{N} \hat{S}(t^* \mid \boldsymbol{x}_i) \right)^2}{\sum_{i=1}^{N} \hat{S}(t^* \mid \boldsymbol{x}_i) \left(1 - \frac{1}{|\mathcal{D}|} \sum_{i=1}^{N} \hat{S}(t^* \mid \boldsymbol{x}_i) \right)} \\
&= \frac{\left(S_{\mathrm{KM}(\mathcal{D})}(t^*) - \mathbb{E}_{\boldsymbol{x}_i \sim \mathcal{D}} \left[\hat{S}(t \mid \boldsymbol{x}_i) \right] \right)^2}{\frac{1}{|\mathcal{D}|} \mathbb{E}_{\boldsymbol{x}_i \sim \mathcal{D}} \left[\hat{S}(t \mid \boldsymbol{x}_i) \right] \left(1 - \mathbb{E}_{\boldsymbol{x}_i \sim \mathcal{D}} \left[\hat{S}(t \mid \boldsymbol{x}_i) \right] \right)}.
\end{aligned}
\tag{12}
$$

This equation illustrates that KM-cal integrates the numerator across all points, effectively representing an unnormalized statistic by omitting the denominator. $\qquad\square$

A.3. Discrimination versus Calibration

Discrimination and calibration are distinct concepts in model evaluation. A proof of this can be observed that the first and second derivative of (9) and (10) wrt to model's parameter do not match. However, the indicator functions in both equations make thing a bit complex.
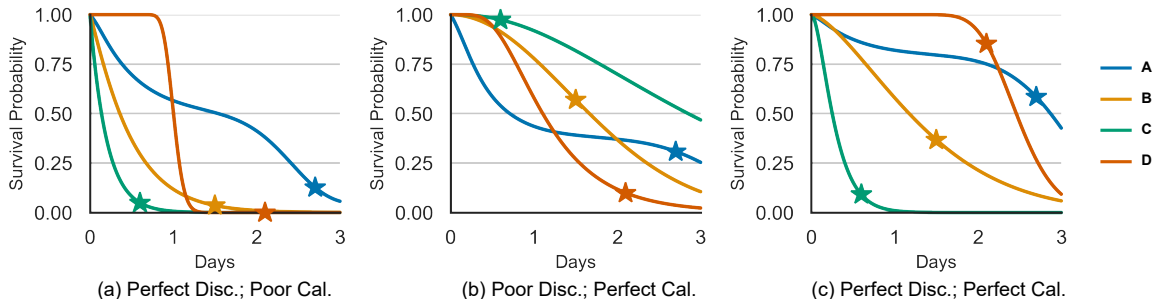

Figure 7. Illustration of discrimination and calibration using three hypothetical examples and the same patients in Figure 1. Here a perfect discrimination means the order of predicted median survival times (where curves intersect with 50%) is aligned with true event times. And a perfect (D-)calibration means the predicted probability at event times is uniformly distributed across $[0, 1]$ (one star in each probability interval). Disc.: discrimination; Cal.: calibration.

In this section, we are using a much easier way to demonstrate their independence through counter-examples:

- **A model $\mathcal{M}_a$ can achieve perfect discrimination and poor calibration;** Figure 7(a) shows the ISD prediction from $\mathcal{M}_a$. By checking the order in which the ISD curves intersect with 50%, we can get the order of predicted median survival times as: green, yellow, red, blue (C, B, D, A), which is perfectly aligned with the order of stars (true event times), indicating perfect discrimination (C-index= 1). However, as to the calibration, we can check the y-coordinate of the stars, which indicates the survival probability at event time. All the y-coordinates are located inside the interval $[0, 0.25]$, indicating the worst possible D-calibration statistics.

- **A model $\mathcal{M}_b$ can achieve poor discrimination and perfect calibration;** Figure 7(b) shows the ISD prediction from $\mathcal{M}_b$. By checking the order in which the ISD curves intersect with 50%, we can get the order of predicted median survival times as: blue, red, yellow, green (A, D, B, C), which is completely reversed from the order of stars (true event times), indicating lowest possible discrimination (C-index= 0). However, as to the calibration, we can find that the y-coordinates of the stars are uniformly distributed among the y-axis (one in each interval), indicating the perfect D-calibration.

- **A model $\mathcal{M}_c$ can achieve perfect discrimination and perfect calibration;** Figure 7(c) shows the ISD prediction from $\mathcal{M}_c$. By checking the order of the ISD curves intersecting with 50%, we can get the order of predicted median survival times as: green, yellow, red, blue (C, B, D, A), which is perfectly aligned with the order of stars (true event times), indicating perfect discrimination (C-index= 1). As to the calibration, we can find that the y-coordinates of the stars are uniformly distributed among the y-axis (one in each interval), indicating the perfect D-calibration.

These examples clarify that a model can exhibit varying levels of discrimination and calibration, underscoring their independence as evaluation metrics.

B. Why Kaplan-Meier models are perfectly calibrated

In this section, we theoretically prove that the Kaplan-Meier (KM) model fitted on the training set is asymptotically calibrated (both KM-cal and D-cal) under some assumptions. Therefore, the calibration scores for the KM model can be used as the *empirical lower bound* for the performance comparison.

Theorem B.1. *The Kaplan-Meier estimation on the training set is asymptotically KM-calibrated, under the* exchangeable *and* conditional independent censoring *assumptions.*

16

Proof. According to the *exchangable* assumption, the samples $\{(\boldsymbol{x}_i, t_i, \delta_i)\}_{i=1}^{N}$ are drawn i.i.d. from an arbitrary joint distribution. Let's assume each subject corresponds to a ground truth condition survival distribution $S(t \mid \boldsymbol{x}_i)$. Then under the law of large numbers, the ground truth marginal survival distribution for the training set should converge to the marginal survival distribution for the testing set, *i.e.*,

$$\mathbb{E}_{\boldsymbol{x}_i \in \mathcal{D}_{\text{train}}} S(t \mid \boldsymbol{x}_i) \quad \rightarrow \quad \mathbb{E}_{\boldsymbol{x}_j \in \mathcal{D}_{\text{test}}} S(t \mid \boldsymbol{x}_j). \tag{13}$$

Furthermore, KM estimator is proven to be unbiased for estimating the marginal survival distribution, under the assumption of *conditional independent censoring* (Hosmer et al., 2008). That means

$$S_{\text{KM}(\mathcal{D}_{\text{train}})}(t \mid \boldsymbol{x}_i) \quad \rightarrow \quad \mathbb{E}_{\boldsymbol{x}_i \in \mathcal{D}_{\text{train}}} S(t \mid \boldsymbol{x}_i) \quad \rightarrow \quad \mathbb{E}_{\boldsymbol{x}_j \in \mathcal{D}_{\text{test}}} S(t \mid \boldsymbol{x}_j) \quad \rightarrow \quad S_{\text{KM}(\mathcal{D}_{\text{test}})}(t \mid \boldsymbol{x}_i), \tag{14}$$

where the final step is the ground truth for KM-cal. This completes the proof. $\qquad\square$

Theorem B.2. *The Kaplan-Meier estimation on the training set is asymptotically D-calibrated, under the (1)* exchangeable *assumption, (2)* independent censoring *assumption, and (3) Kaplan-Meier estimation is* strictly *monotonic decreasing.*

Proof. Firstly, note that here we must use a different censoring assumption than in Theorem B.1. For independent censoring assumption, we think the event and censoring times are independent, regardless of features, *i.e.*, $e_i \perp c_i$. This is a stronger assumption, because $e_i \perp c_i$ can obviously lead to $e_i \perp c_i \mid \boldsymbol{x}_i$, but not the other way around.

Haider et al. (2020) prove that the proportion $\frac{|\hat{\mathcal{D}}(a,b)|}{|\mathcal{D}|}$ is equal to the difference $b - a$ for the true ISD $S(t \mid \boldsymbol{x}_i)$, under the assumption that the true ISD is strictly monotonic decreasing (see Theorem B.3 in Haider et al. (2020)). We can easily extend this proof to the KM estimation by applying the independent censoring assumption. Specifically, independent censoring assumption is required for the probability decomposition step in Haider et al. (2020) proof (while the previous true ISD $S(t \mid \boldsymbol{x}_i)$ only needs the less restrictive conditional independent censoring assumption). Therefore, using these two assumptions, we can get that the KM model fitted on the testing set is D-calibrated.

To get to the last part (the KM model fitted to the training set is D-calibrated), we will need the *exchangable* assumption, just like in Theorem B.1. This completes the proof. $\qquad\square$

C. Proofs

C.1. Proof of Theorem 3.1

Before proving this theorem, we first recall some basic concepts in survival analysis. Following the notations in Section 2, a trained survival model $\mathcal{M}$ predicts the individual survival distribution (ISD) for subject i: $\hat{S}_{\mathcal{M}}(t \mid \boldsymbol{x}_i)$. A predicted event time $\hat{t}_i$ can then be represented using this ISD by either median survival time:

$$\hat{t}_i^{(\text{median})} \quad = \quad \inf\{t : \hat{S}_{\mathcal{M}}(t \mid \boldsymbol{x}_i) \leq 0.5\} \quad = \quad \hat{S}_{\mathcal{M}}^{-1}(\rho = 0.5 \mid \boldsymbol{x}_i), \tag{15}$$

or using mean (expected) survival time:

$$\hat{t}_i^{(\text{mean})} \quad = \quad \mathbb{E}_t\left[\hat{S}_{\mathcal{M}}(t \mid \boldsymbol{x}_i)\right] \quad = \quad \int_0^{t_{\max}} \hat{S}_{\mathcal{M}}(t \mid \boldsymbol{x}_i)\, dt, \tag{16}$$

where we overuse the symbol $t_{\max}$ to represent the time that ISD first reaches zero.

Recall Section 2, we need to discretize the ISD prediction for the conformal step using:

$$\hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i) \quad = \quad \hat{S}_{\mathcal{M}}^{-1}(\rho \mid \boldsymbol{x}_i),$$

then after the conformalization, we can recover the ISD prediction using the conformalized quantile prediction:

$$\hat{S}_{\mathcal{M}}'(\rho \mid \boldsymbol{x}_i) \quad = \quad \hat{q}_{\mathcal{M}}'^{-1}(\rho \mid \boldsymbol{x}_i).$$

Now we have everything we need for the proof. Here we formally restate Theorem 3.1 in Section 3.

Theorem C.1. *Applying the CSD adjustment to the percentile predictions does not impact the C-index of the model. Formally, $\forall\, i, j \in \mathcal{D}$, for median survival times, we can have:*

$$\text{given}\quad \hat{q}_{\mathcal{M}}(\rho = 0.5 \mid \boldsymbol{x}_i) > \hat{q}_{\mathcal{M}}(\rho = 0.5 \mid \boldsymbol{x}_j) \quad \Longrightarrow \quad \hat{q}'_{\mathcal{M}}(\rho = 0.5 \mid \boldsymbol{x}_i) > \hat{q}'_{\mathcal{M}}(\rho = 0.5 \mid \boldsymbol{x}_j), \tag{17}$$

and for mean survival times, we have:

$$\text{given}\quad \int_{t=0}^{t_{max}} \hat{S}_{\mathcal{M}}(t \mid \boldsymbol{x}_i)\, dt > \int_{t=0}^{t_{max}} \hat{S}_{\mathcal{M}}(t \mid \boldsymbol{x}_j)\, dt \quad \Longrightarrow \quad \int_{t=0}^{t_{max}} \hat{S}'_{\mathcal{M}}(t \mid \boldsymbol{x}_i)\, dt > \int_{t=0}^{t_{max}} \hat{S}'_{\mathcal{M}}(t \mid \boldsymbol{x}_j)\, dt, \tag{18}$$

with the assumption that $t_{max} < \infty$ is a finite number.

Proof. For the first part of the proof (median survival time), let's take the partial derivative of (5) with respect to $\boldsymbol{x}_i$ when $\rho = 0.5$:

$$
\begin{aligned}
\frac{\partial \hat{q}'_{\mathcal{M}}(\rho = 0.5 \mid \boldsymbol{x}_i)}{\partial \boldsymbol{x}_i} &= \frac{\partial \hat{q}_{\mathcal{M}}(\rho = 0.5 \mid \boldsymbol{x}_i)}{\partial \boldsymbol{x}_i} - \frac{\partial\, \text{Quantile}\left[\,\rho = 0.5; \mathcal{S}_{\mathcal{M}}(\rho = 0.5)\,\right]}{\partial \boldsymbol{x}_i} \\
&= \frac{\partial \hat{q}_{\mathcal{M}}(\rho = 0.5 \mid \boldsymbol{x}_i)}{\partial \boldsymbol{x}_i}.
\end{aligned}
$$

Because the partial derivative of the second term is 0, we can see that the pre-CSD prediction $\hat{q}_{\mathcal{M}}(0.5 \mid \boldsymbol{x}_i)$ and post-CSD prediction $\hat{q}'_{\mathcal{M}}(0.5 \mid \boldsymbol{x}_i)$ have the same partial derivative. That means that the tangent lines at corresponding points should have the same slopes, or, put another way, their graph should go up and down in the same way when $\boldsymbol{x}_i$ changes. This proves (17).

For the second part of the proof (mean survival time), we first use the predicted percentile times (PCT) to represent the mean survival time. Specifically, based on the integral of inverse function theorem, we can have:

$$
\begin{aligned}
\int_{t=0}^{t_{max}} \hat{S}_{\mathcal{M}}(t \mid \boldsymbol{x}_i)\, dt + \int_{\rho = \hat{S}_{\mathcal{M}}(0 \mid \boldsymbol{x}_i)}^{\hat{S}_{\mathcal{M}}(t_{max} \mid \boldsymbol{x}_i)} \hat{S}_{\mathcal{M}}^{-1}(\rho \mid \boldsymbol{x}_i)\, d\rho &= t_{max} \cdot \hat{S}_{\mathcal{M}}(t_{max} \mid \boldsymbol{x}_i) - 0 \cdot \hat{S}_{\mathcal{M}}(0 \mid \boldsymbol{x}_i) \\
\Longrightarrow \qquad \hat{t}_i^{(\text{mean})} + \int_{\rho=1}^{\rho=0} \hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i)\, d\rho &= t_{max} \cdot 0 - 0 \cdot 1 \\
\Longrightarrow \qquad\qquad\qquad \hat{t}_i^{(\text{mean})} &= \int_{\rho=0}^{1} \hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i)\, d\rho. \tag{19}
\end{aligned}
$$

Correspondingly, we can have the post-CSD mean survival time as

$$
\begin{aligned}
\hat{t}_i^{\prime\,(\text{mean})} &= \int_{\rho=0}^{1} \hat{q}'_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i)\, d\rho \\
&= \int_{\rho=0}^{1} \hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i) - \text{Quantile}[\,\rho; \mathcal{S}_{\mathcal{M}}(\rho)\,]\, d\rho. \tag{20}
\end{aligned}
$$

Therefore, by taking the partial derivative to both (19) and (20) with respect to $\boldsymbol{x}_i$ again (follow the same procedure above), we can easily find that pre- and post-CSD mean survival time have the same partial derivative. Therefore, we can have (18) proved. $\square$

It's important to clarify a potential misunderstanding related to this theorem: there might be confusion that the CSD post-process does not alter the ordering of survival probability predictions. However, this is a misconception. The CSD modifies survival curves along the horizontal axis, specifically the time dimension. Consequently, this adjustment can lead to changes in the vertical sequence of survival probabilities. It is crucial to note that the theorem's claims are specifically concerning median and mean survival times, not the ordering of survival probabilities.

Remark We now discuss about the reasonableness of the assumption that $t_{max} < \infty$ is a finite number. We argue that it is a reasonable assumption in health-care related tasks, as the human lifespan is known to be a finite number (only cancer cells are immortal).

C.2. Proof of Theorem 3.2

Theorem C.2. *Under the exchangeable assumption, the percentile predictions constructed by CSD will exhibit exact distribution calibration, s.t., $\forall \rho \in [0,1]$, and $i \in \mathcal{D}_{test}$, we can have*

$$\rho \quad \leq \quad \mathbb{P}(t_i \in [\,\hat{q}'_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i), \infty\,] \mid \boldsymbol{x}_i) \quad \leq \quad \rho + \frac{1}{|\mathcal{D}_{\mathrm{con}}| + 1}. \tag{21}$$

Proof. This proof is a standard proof in conformal regression (Romano et al., 2019; Angelopoulos et al., 2023), with minor modifications to fit the context of survival analysis. Here for completeness, we will present the complete proof.

Let $s_{i,\mathcal{M}}(\rho)$ be the conformity score (4) for a test point i and percentile level ρ. By the construction of the prediction interval $[\hat{q}'_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i), \infty]$, we have

$$
\begin{aligned}
t_i \in [\,\hat{q}'_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i), \infty\,] \quad &\Longleftrightarrow \quad t_i \geq \hat{q}'_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i) = \hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i) - \mathrm{Quantile}[\,\rho; \mathcal{S}_{\mathcal{M}}(\rho) = \{\,\hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i) - t_j\,\}_{j=1}^{|\mathcal{D}_{\mathrm{con}}|}\,] \\
&\Longleftrightarrow \quad \mathrm{Quantile}[\,\rho; \mathcal{S}_{\mathcal{M}}(\rho)\,] \geq \hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i) - t_i \\
&\Longleftrightarrow \quad \mathrm{Quantile}[\,\rho; \mathcal{S}_{\mathcal{M}}(\rho)\,] \geq s_{i,\mathcal{M}}(\rho),
\end{aligned}
$$

therefore, by taking the expectation over the first and last term, we have

$$\mathbb{P}(t_i \in [\,\hat{q}'_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i), \infty\,] \mid (\boldsymbol{x}_i, t_i, \delta_i) \in \mathcal{D}_{\mathrm{test}}) \quad = \quad \mathbb{P}(s_{i,\mathcal{M}}(\rho) \leq \mathrm{Quantile}[\,\rho; \mathcal{S}_{\mathcal{M}}(\rho)\,] \mid (\boldsymbol{x}_i, t_i, \delta_i) \in \mathcal{D}_{\mathrm{test}}). \tag{22}$$

Without loss of generality, we assume that the conformity scores for the conformal set are sorted in $\mathcal{S}_{\mathcal{M}}(\rho)$, so that $s_1 < s_2 < \cdots < s_{|\mathcal{D}_{\mathrm{con}}|}$. Therefore, for a new subject i in $\mathcal{D}_{\mathrm{test}}$, its conformity score has the same probability of falling into any of the following intervals, $[-\infty, s_1], [s_1, s_2], \cdots, [s_{|\mathcal{D}_{\mathrm{con}}|}, \infty]$. As we have $|\mathcal{D}_{\mathrm{con}}|$ conformity scores in the set, so we have $|\mathcal{D}_{\mathrm{con}}| + 1$ intervals. We conclude that for subject i in the testset, we can have

$$\mathbb{P}(s_{i,\mathcal{M}} \leq \mathrm{Quantile}[\,\rho; \mathcal{S}_{\mathcal{M}}(\rho)\,]) \quad = \quad \frac{\lceil \rho(|\mathcal{D}_{\mathrm{con}}| + 1)\rceil}{|\mathcal{D}_{\mathrm{con}}| + 1} \quad = \quad \frac{\rho(|\mathcal{D}_{\mathrm{con}}| + 1) + (1 - \epsilon)}{|\mathcal{D}_{\mathrm{con}}| + 1}.$$

The last equality is by converting the ceiling function to the original value plus a residual $(1 - \epsilon)$, where ϵ is an infinitesimally small positive number,[8] ensuring that the ceiling function just rounds up to the next integer.[9] By combining this equation with (22), we have proved the theorem. $\qquad\square$

C.3. Proof of Lemma 3.3

Lemma C.3. *Under the exchangeable assumption, the percentile predictions constructed by CSD will asymptotically exhibit exact integrated calibration at all time points, which means that the prediction is KM-calibrated.*

Proof. We derive this by first building a connection between D-cal and KM-cal. We prove that D-cal and KM-cal are asymptotically equal. Then, under Theorem 3.2, CSD will asymptotically be KM-calibrated.

Let's first consider the uncensored cases, and prove the lemma in such circumstances.

For the KM-cal with uncensored dataset, without the loss of generality, we can convert the integral from time 0 to maximum

---

[8]This also requires that $0 \leq (1 - \epsilon) < 1$.

[9]This transformation is only achievable under the assumption of conformity score being continuous, avoiding ties. However, as pointed out by Angelopoulos et al. (2023), this condition can be solved by adding a small perturbation to the score.

time to the summation over the discretized point $\{t_k\}_{k=1}^{k=K}$, where $t_K = t_{\max}$.[10] Then the KM-cal calculation is converted to:

$$
\begin{aligned}
\mathcal{R}_{\text{KM-cal}}(\mathcal{D}) &= \frac{1}{t_{\max}} \int_{t=0}^{t_{\max}} \left(S_{\text{KM}(\mathcal{D})}(t) - \mathbb{E}_{\boldsymbol{x}_i}\left[\hat{S}(t \mid \boldsymbol{x}_i) \right] \right)^2 dt \\
&= \frac{1}{t_K} \sum_{k=1}^{K} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[e_i > t_k] - \frac{1}{N} \sum_{i=1}^{N} \hat{S}(t = t_k \mid x_i) \right)^2 \quad \text{(No censorship)} \\
&= \frac{1}{N^2 t_K} \sum_{k=1}^{K} \left(\sum_{i=1}^{N} \left(\mathbb{1}[e_i > t_k] - \hat{S}(t = t_k \mid x_i) \right) \right)^2 . \quad (23)
\end{aligned}
$$

Same for D-cal, given a set of percentiles $\mathcal{P} = \{\rho_1, \rho_2, \ldots, \rho_L\}$ with $0 < \rho_1 < \cdots < \rho_l < \cdots < \rho_L \leq 1$, D-cal is calculated by

$$
\begin{aligned}
\mathcal{R}_{\text{D-cal}}(\mathcal{D}) &= \frac{1}{L} \sum_{l=1}^{L} \left(\frac{|\hat{\mathcal{D}}(0, \rho_l)|}{N} - \rho_l \right)^2 \\
&= \frac{1}{N^2 L} \sum_{l=1}^{L} \left(|\hat{\mathcal{D}}(0, \rho_l)| - N \cdot \rho_l \right)^2 \\
&= \frac{1}{N^2 L} \sum_{l=1}^{L} \left(\sum_{i=1}^{N} \left(\mathbb{1}[\hat{S}(e_i \mid \boldsymbol{x}_i) < \rho_l] - \rho_l \right) \right)^2 . \quad (24)
\end{aligned}
$$

Now, if we compare (24) with (23), we can find that D-cal calculates over the percentile domain while KM-Cal does the calculation over the time domain. To make them talk using the same language, we first need to relate the percentiles with times.
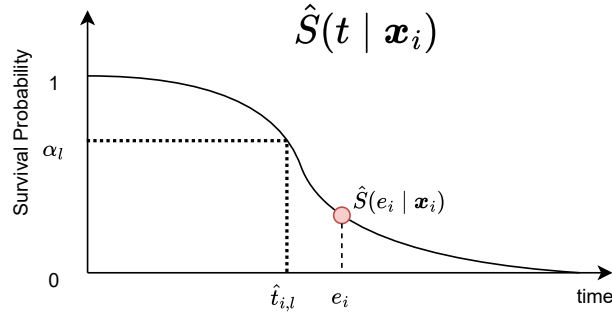

Figure 8. Relationship between $\hat{t}_{i,l}$, ρ_l, and $\hat{S}(t \mid \boldsymbol{x}_i)$.

Figure 8 shows the relationship between a percentile value and a time point. The curve represents the predicted ISD curve, $\hat{S}(t \mid \boldsymbol{x}_i)$. $\rho_l = \hat{S}(t = \hat{t}_{i,l} \mid \boldsymbol{x}_i)$ and its reverse, $\hat{t}_{i,l} = \hat{S}^{-1}(\rho = \rho_l \mid \boldsymbol{x}_i)$. Based on this expression, we can have this proposition.

Proposition C.4. *For a strictly monotonic decreasing (no plateau area) ISD $\hat{S}(t \mid \boldsymbol{x}_i)$, if the predicted probability at the event time, s.t. $\hat{S}(t = e_i \mid \boldsymbol{x}_i)$ is less than ρ_l, then the event time e_i must be greater than the corresponding percentile time $t_{i,l}$, and vice versa. Formally:*

$$
\hat{S}(t = e_i \mid \boldsymbol{x}_i) < \rho_l \iff e_i > \hat{t}_{i,l}.
$$

---

[10]Here we overuse the subscript k to represent the index of the discretized time points, where in Section 2 it represents the index for ordered group for 1-cal calculation.

Therefore, following this proposition, we can reformulate (24) to:

$$\mathcal{R}_{\text{D-cal}}(\mathcal{D}) = \frac{1}{N^2 L} \sum_{l=1}^{L} \left(\sum_{i=1}^{N} \left(\mathbb{1}[\hat{t}_{i,l} < e_i] - \hat{S}(t = \hat{t}_{i,l} \mid \boldsymbol{x}_i) \right) \right)^2, \tag{25}$$

Now, if we compare (23) with (25), the only difference is that KM-Cal uses the discrete time indexed by k, and D-cal uses the time indexed by i and l. However, if the two sets $\{t_k\}_{k=1}^{K}$ and $\{\hat{t}_{i,l}\}_{i=1,l=1}^{N,L}$ are asymptotically equal ($K, N, L \to \infty$), then the integrated KM-cal and D-cal should also be asymptotically equal.

This condition is not always satisfied, for population survival distribution models (*e.g.,* , Kaplan-Meier, Nelson-Aalen, or models that estimated one survival distribution for the entire population), the set $\{\hat{t}_{i,l}\}_{i=1,l=1}^{N,L}$ will just degenerate to $\{\hat{t}_l\}_{l=1}^{L}$, then we can easily align $\{\hat{t}_l\}_{l=1}^{L}$ with $\{t_k\}_{k=1}^{K}$.

However, for the individual survival distribution models, we cannot have $\{\hat{t}_{i,l}\}_{i=1,l=1}^{N,L}$ coincide with $\{t_k\}_{k=1}^{K}$ for almost all cases; this is mostly because the predicted curves can be a lot different across subjects, so we would end up with a very diverse $\hat{t}_{i,l}$. That being said, the more diverse the ISDs are, the greater the divergence between the D-cal and KM-cal. This phenomenon is typical in statistical models, e.g., Cox proportional hazard (and its extensions, GB, DeepSurv) and AFT models are inherently (semi-)parametric, which here means that their ISD curves will have similar shapes. However, other models, *e.g.*, N-MTLR and CQRNN, produce ISD curves that can have very different shapes, therefore, larger divergence. But still, the above discussion provides a helpful prospective to connect the two important calibration methods for survival analysis. □

Remark This proof is not in conflict with the findings presented by Haider et al. (2020), where they demonstrated that a model could achieve optimal D-cal yet exhibit subpar performance in terms of 1-Cal, and vice versa. However, it's important to note the distinction in the context of calibration being discussed: Haider et al. (2020)'s argument pertains to 1-Cal at a specific point in time, whereas our proof addresses the concept of KM-cal, which considers 1-cal performance over a continuum of time points. This differentiation is akin to how a model might excel in terms of the Integrated Brier Score (IBS) across multiple time points but may still underperform when evaluated using the Brier Score (BS) at any given single time point.

Let's now expand this proof to censored cases.

Proof. For censored datasets, we rewrite the KM-cal as

$$\mathcal{R}_{\text{KM-cal}}(\mathcal{D}) = \frac{1}{t_{\max}} \int_{t=0}^{t_{\max}} \left(S_{\text{KM}(\mathcal{D})}(t) - \mathbb{E}_{\boldsymbol{x}_i} \left[\hat{S}(t \mid \boldsymbol{x}_i) \right] \right)^2 dt$$

$$= \frac{1}{N^2 t_K} \sum_{k=1}^{K} \left(\sum_{i=1}^{N} \left(S_{\text{KM}(\mathcal{D})}(t_k) - \hat{S}(t = t_k \mid x_i) \right) \right)^2,$$

and D-cal as:

$$\mathcal{R}_{\text{D-cal}}(\mathcal{D}) = \frac{1}{L} \sum_{l=1}^{L} \left(\frac{|\hat{\mathcal{D}}(0, \rho_l)|}{N} - \rho_l \right)^2$$

$$= \frac{1}{N^2 L} \sum_{l=1}^{L} \left(\sum_{i=1}^{N} \left(\frac{|\hat{\mathcal{D}}(0, \rho_l)|}{N} - \rho_l \right) \right)^2,$$

where the proportion of subset is

$$\frac{|\hat{\mathcal{D}}(0, \rho_l)|}{N} = \frac{1}{N} \sum_{i=1}^{N} \delta_i \cdot \mathbb{1}\left[\hat{S}(t_i \mid \boldsymbol{x}_i) < \rho_l \right] + (1 - \delta_i) \left(\mathbb{1}\left[\hat{S}(t_i \mid \boldsymbol{x}_i) < \rho_l \right] + \frac{\rho_l \cdot \mathbb{1}[\hat{S}(t_i \mid \boldsymbol{x}_i) \geq \rho_l]}{\hat{S}(t_i \mid \boldsymbol{x}_i)} \right). \tag{26}$$

As previously demonstrated in the uncensored cases, the second terms of both equations, $\hat{S}(t = t_k \mid \boldsymbol{x}_i)$ and ρ_i converge asymptotically. Now we only need to establish that the first terms, $S_{\text{KM}(\mathcal{D})}(t_k)$ and $\frac{|\hat{\mathcal{D}}(0,\rho_l)|}{N}$, are also asymptotically equivalent.

21

Using the expression of (10). We can take the expectation of (26),

$$
\begin{aligned}
\mathbb{E}_i\left[\frac{|\hat{\mathcal{D}}(0,\rho_l)|}{N}\right] &= \mathbb{E}_i\left[\mathbb{1}\left[\hat{S}(e_i\mid\boldsymbol{x}_i)<\rho_l\right]\cdot\mathbb{1}\left[e_i\leq c_i\right]\right] + \mathbb{E}_i\left[\mathbb{1}\left[\hat{S}(c_i\mid\boldsymbol{x}_i)<\rho_l\right]\cdot\mathbb{1}\left[c_i<e_i\right]\right] \\
&\quad + \mathbb{E}_i\left[\frac{\rho_l}{\hat{S}(c_i\mid\boldsymbol{x}_i)}\mathbb{1}\left[\hat{S}(c_i\mid\boldsymbol{x}_i)\geq\rho_l\right]\cdot\mathbb{1}\left[c_i<e_i\right]\right] \\
&= \mathbb{P}\left(\hat{S}(e_i\mid\boldsymbol{x}_i)<\rho_l\,,\,e_i\leq c_i\right) + \mathbb{P}\left(\hat{S}(c_i\mid\boldsymbol{x}_i)<\rho_l\,,\,c_i<e_i\right) \\
&\quad + \mathbb{E}_i\left[\frac{\rho_l}{\hat{S}(c_i\mid\boldsymbol{x}_i)}\mathbb{1}\left[\hat{S}(c_i\mid\boldsymbol{x}_i)\geq\rho_l\right]\cdot\mathbb{1}\left[c_i<e_i\right]\right],
\end{aligned}
$$

using the transformation in Proposition C.4, we can have

$$
\mathbb{E}_i\left[\frac{|\hat{\mathcal{D}}(0,\rho_l)|}{N}\right] = \mathbb{P}\left(e_i>\hat{t}_{i,l}\,,\,e_i\leq c_i\right) + \mathbb{P}\left(c_i>\hat{t}_{i,l}\,,\,c_i<e_i\right) + \mathbb{E}_i\left[\frac{\rho_l}{\hat{S}(c_i\mid\boldsymbol{x}_i)}\mathbb{1}[c_i\leq\hat{t}_{i,l}]\cdot\mathbb{1}\left[c_i<e_i\right]\right].
\tag{27}
$$

For the first two terms in the above derivation:

$$
\begin{aligned}
\mathbb{P}\left(e_i>\hat{t}_{i,l}\,,\,e_i\leq c_i\right)+\mathbb{P}\left(c_i>\hat{t}_{i,l}\,,\,c_i<e_i\right) &= \mathbb{P}\left(e_i>\hat{t}_{i,l}\,,\,e_i\leq c_i\right)+\mathbb{P}\left(c_i>\hat{t}_{i,l}\,,\,e_i>\hat{t}_{i,l}\,,\,c_i<e_i\right) \\
&= \mathbb{P}\left(e_i>\hat{t}_{i,l}\,,\,c_i>\hat{t}_{i,l}\right) \\
&= \mathbb{P}(e_i>\hat{t}_{i,l})\cdot\mathbb{P}(c_i>\hat{t}_{i,l}),
\end{aligned}
$$

where the probability decomposition is because of the conditional independent censoring assumption $e_i\perp c_i\mid\boldsymbol{x}_i$. Then, for the third term:

$$
\begin{aligned}
\mathbb{E}_i\left[\frac{\rho_l}{\hat{S}(c_i\mid\boldsymbol{x}_i)}\mathbb{1}[c_i\leq\hat{t}_{i,l}]\cdot\mathbb{1}\left[c_i<e_i\right]\right] &= \mathbb{P}(\hat{S}(e_i\mid\boldsymbol{x}_i)\leq\rho_l)\int_{c_i=0}^{\hat{t}_{i,l}}\int_{e_i=c_i}^{\infty}\frac{1}{\hat{S}(c_i\mid\boldsymbol{x}_i)}f(c_i,e_i\mid\boldsymbol{x}_i)\,de_i\,dc_i \\
&= \mathbb{P}(e_i\geq\hat{t}_{i,l})\int_{c_i=0}^{\hat{t}_{i,l}}\int_{e_i=c_i}^{\infty}\frac{1}{\hat{S}(c_i\mid\boldsymbol{x}_i)}f(c_i,e_i\mid\boldsymbol{x}_i)\,de_i\,dc_i \\
&= \mathbb{P}(e_i\geq\hat{t}_{i,l})\int_{c_i=0}^{\hat{t}_{i,l}}\frac{1}{\hat{S}(c_i\mid\boldsymbol{x}_i)}f(c_i\mid\boldsymbol{x}_i)\int_{e_i=c_i}^{\infty}f(e_i\mid\boldsymbol{x}_i)\,de_i\,dc_i \\
&= \mathbb{P}(e_i\geq\hat{t}_{i,l})\int_{c_i=0}^{\hat{t}_{i,l}}\frac{f(c_i\mid\boldsymbol{x}_i)}{\hat{S}(c_i\mid\boldsymbol{x}_i)}\hat{S}(c_i\mid\boldsymbol{x}_i)\,dc_i \\
&= \mathbb{P}(e_i\geq\hat{t}_{i,l})\int_{c_i=0}^{\hat{t}_{i,l}}f(c_i\mid\boldsymbol{x}_i)\,dc_i \\
&= \mathbb{P}(e_i\geq\hat{t}_{i,l})\cdot\mathbb{P}(c_i\leq\hat{t}_{i,l})
\end{aligned}
$$

Therefore, by combining the above two derivations to (27):

$$
\mathbb{E}_i\left[\frac{|\hat{\mathcal{D}}(0,\rho_l)|}{N}\right] = \mathbb{P}(e_i\geq\hat{t}_{i,l})\cdot\mathbb{P}(c_i>\hat{t}_{i,l}) + \mathbb{P}(e_i\geq\hat{t}_{i,l})\cdot\mathbb{P}(c_i\leq\hat{t}_{i,l}) = \mathbb{P}(e_i\geq\hat{t}_{i,l}).
$$

We know that the KM estimator is simply an empirical estimator of $\mathbb{P}(e_i\geq t_k)$. That implies that if the two sets $\{t_k\}_{k=1}^K$ and $\{\hat{t}_{i,l}\}_{i=1,l=1}^{N,L}$ become asymptotically equivalent as K, N, and L approach infinity, then we can have $S_{\mathrm{KM}(\mathcal{D})}(t_k) = \mathbb{E}\left[\frac{|\hat{\mathcal{D}}(0,\rho_l)|}{N}\right]$. Consequently, the KM-cal and D-cal are asympotically equal. This completes the proof. $\square$

D. Algorithm

D.1. Decensoring method

Margin   Margin (Haider et al., 2020) assigns a "best guess" value (margin time) to each censored subject using the non-parametric population KM (Kaplan & Meier, 1958) estimator. This margin time can be interpreted as a conditional expectation of the event time given the event time is greater than the censoring time (Haider et al., 2020). Given a subject censored at time t_i, we can calculate its margin time by:

$$e_{\text{margin}}(t_i, \mathcal{D}) \quad = \quad \mathbb{E}_t[e_i \mid e_i > t_i] \quad = \quad t_i + \frac{\int_{t_i}^{\infty} S_{\text{KM}(\mathcal{D})}(t)dt}{S_{\text{KM}(\mathcal{D})}(t_i)} \ ,$$

where $S_{\text{KM}(\mathcal{D})}(t)$ is typically derived from the training dataset.

Pseudo-observations   Another way to estimate surrogate event values is using pseudo-observations (PO) (Andersen et al., 2003; Andersen & Pohar Perme, 2010). Let $\{t_i\}_{i=i}^{N}$ be i.i.d. draws of a random variable time, T, and let $\hat{\theta}$ be an unbiased estimator for the event time based on right-censored observations of T. The PO for a censored subject is defined as:

$$e_{\text{PO}}(t_i, \mathcal{D}) \quad = \quad N \times \hat{\theta} - (N - 1) \times \hat{\theta}^{-i} \ , \tag{28}$$

where $\hat{\theta}^{-i}$ is the estimator applied to the $N - 1$ element dataset formed by removing that i-th instance. The PO can be viewed as the contribution of subject i to the unbiased event time estimation $\hat{\theta}$. Here we can use, the mean survival time of the KM estimator, $\hat{\theta} = \mathbb{E}_t[S_{\text{KM}(\mathcal{D})}(t)]$ and $\hat{\theta}^{-i} = \mathbb{E}_t[S_{\text{KM}(\mathcal{D}^{-i})}(t)]$ as unbiased estimators.

Some nice properties of PO values are: (1) PO values can be treated as though they are i.i.d. (Appendix D.1 in Qi et al. (2023a)). (2) Graw et al. (2009) has shown that, as $N \to \infty$, PO can approximate the correct conditional expectation:

$$\mathbb{E}[\, e_{\text{pseudo-obs}}(t_i) \mid \boldsymbol{x}_i \,] \quad \approx \quad \mathbb{E}[\, e_i \mid \boldsymbol{x}_i \,] \ ,$$

in situations where censoring does not depend on the covariates, and it also works well empirically in the situation where censoring is dependent on the covariates Binder et al. (2014).

D.2. CSD with KM-sampling

In this section, we provide the pseudo-algorithm for the conformalized survival distribution with KM sampling.

D.3. Non-monotonicity after CSD

Survival distribution is defined as the complement of CDF function, that means is defined to be always monotonically decreasing. Most survival models will apply a monotonic restriction to the output.[11] That means the discretized percentile prediction should still be monotonic, i.e., $\hat{q}_{\mathcal{M}}(\rho_1 \mid \boldsymbol{x}_i) \leq \hat{q}_{\mathcal{M}}(\rho_2 \mid \boldsymbol{x}_i)$ for arbitrary $\rho_1 > \rho_2$. However, we could encounter some non-monotonic curves after the operation in (5) in some rare cases. This usually occurs when ρ is small (at the tail of the ISD curve). This can be fixed by the bootstrap rearranging method (Chernozhukov et al., 2010).

Remark   However, after fixing the non-monotonicity, there is no such guarantee that the predicted order remains the same, that means Theorem 3.1 will not hold anymore. We anticipate that only few models and predictions will be affected by this, so the difference in C-index will be very slight. Surprisingly, as we show in Section 4.4 and Appendix F, this modification will actually make the C-index slightly better.

E. Experimental Details

E.1. Datasets and Preprocessing

In this section, we will describe how we preprocess the raw survival datasets.

The Veterans' Administration Lung Cancer Trial (VALCT) dataset (Kalbfleisch & Prentice, 2011) is derived from a randomized trial comparing two treatment regimens for lung cancer. It includes data from 137 patients, each described by 6 features. This dataset, recognized as a standard resource for survival analysis, is available in R's `survival` package.

---

[11]But not for all. For example, in the study, the CQRNN algorithm (Pearce et al., 2022) does not have this monotonic constraint.

Algorithm 1 Conformalized Survival Distribution with KM sampling

---

1: **Input:** dataset $\mathcal{D}$, testing set $\mathcal{D}_{\text{test}}$, Survival model $\mathcal{M}$, Predefined percentile levels $\mathcal{P} = \{\rho_1, \rho_2, \ldots\}$, repeat parameter R

2: **Output:** ISD prediction for $\mathcal{D}_{\text{test}}$

3: Randomly split $\mathcal{D}$ into a training set $\mathcal{D}_{\text{train}}$ and a conformal set $\mathcal{D}_{\text{train}}$

4: Train the survival model $\mathcal{M}$ using $\mathcal{D}_{\text{train}}$ as the training set, and $\mathcal{D}_{\text{con}}$ as the validation set.

5: For every subject j in $\mathcal{D}_{\text{con}}$, make ISD predictions $\hat{S}_{\mathcal{M}}(t \mid \boldsymbol{x}_j)$

6: Discretized the ISD predictions to PCTs using the predefined percentiles $\mathcal{P}$, using $\hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i) = \hat{S}_{\mathcal{M}}^{-1}(\rho \mid \boldsymbol{x}_j)$

7: Calculate the KM estimation for the conformal set $S_{KM}(t)$

8: **for** $r = 1$ **to** R **do**

9:   **for** $j \in \mathcal{D}_{\text{con}}$ **do**

10:     **if** $\delta_j == 0$ **then**

11:       Calculate the "best-guess" distribution $S_{\text{KM}}(t \mid t > c_j) = \min\left\{\frac{S_{\text{KM}}(t)}{S_{\text{KM}}(c_j)}, 1\right\}$

12:       Sample $t_j^r =\sim S_{\text{KM}}(t \mid t > c_j)$

13:     **else**

14:       $t_j^r = t_j$

15:     **end if**

16:   **end for**

17: **end for**

18: Calculate the set of conformity scores $\mathcal{S}_{\mathcal{M},R}(\rho)$ for the conformal set using (8), for every $\rho \in \mathcal{P}$

19: For every subject i in $\mathcal{D}_{\text{test}}$, make ISD predictions $\hat{S}_{\mathcal{M}}(t \mid \boldsymbol{x}_i)$

20: Discretized the ISD predictions to PCTs using the predefined percentiles $\mathcal{P}$, $\hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i) = \hat{S}_{\mathcal{M}}^{-1}(\rho \mid \boldsymbol{x}_j)$

21: Get the final prediction. $\hat{q}'_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i) = \hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i) - \text{Quantile}\left[\rho; \mathcal{S}_{\mathcal{M},R}(\rho)\right]$

Diffuse Large B-Cell Lymphoma (DLBCL) dataset (Li et al., 2016) contains Lymphochip DNA microarrays from 240 biopsy samples of DLBCL tumors for studying the survival status of the corresponding patients and the observation lasts 21 years. It is a high-dimensional dataset, containing 7,399 features for each sample.

Mayo clinic Primary Biliary Cholangitis (PBC) data (Therneau & Grambsch, 2000) comprises data from 424 patients diagnosed with PBC, an autoimmune liver disease. This is a standard survival analysis dataset available in the survival package in R. In this data set, the status of each patient at the conclusion of the study is categorized into one of three types: censored, transplanted, or deceased. Our analysis aims to model the survival time following diagnosis. Therefore, we treat both censored and transplant cases as censored events ($\delta_i = 0$), and cases where the patient died as uncensored events ($\delta_i = 1$). For handling missing data, we employ different strategies based on the type of feature. For continuous variables, we impute missing values using their mean. For categorical variables, we fill in missing values with the most common category (mode).

Glioblastoma multiforme (GBM) dataset is retrieved from The Cancer Genome Atlas (TCGA) dataset (Weinstein et al., 2013). We only select patients diagnosed with glioblastoma multiforme cancer to build the GBM data set. The TCGA data can be found in http://firebrowse.org/ or by the instruction in Haider et al. (2020). The median values of radiation therapy, Karnofsky performance score, and ethnicity have been used to fill the missing values.

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset (Curtis et al., 2012) contains survival information for breast cancer patients. This dataset includes a diverse range of feature sets that encompass clinical traits, expression profiles, copy number variation (CNV) profiles, and single nucleotide polymorphism (SNP) genotypes. All these features are derived from breast tumor samples collected during the METABRIC trial. The dataset can be downloaded from (https://www.cbioportal.org/study/summary?id=brca_metabric), and it does not have any missing values.

Rotterdam & German Breast Cancer Study Group (GBSG) dataset (Katzman et al., 2018) represents an amalgamation of data from two sources: the Rotterdam tumor bank (Foekens et al., 2000) and the German Breast Cancer Study Group (Schumacher et al., 1994). The Rotterdam tumor bank dataset contains records for 1,546 patients with node-positive breast cancer the GBSG contains complete data for 686 patients. This combined dataset, which was already post-processed in the DeepSurv study (Katzman et al., 2018), is available for download at https://github.com/jaredleekatzman/

`DeepSurv/`.

The Northern Alberta Cancer Dataset (`NACD`) (Haider et al., 2020) described 2402 patients with various cancers, including lung, colorectal, head and neck cancers, esophagus, stomach, and other cancers. The event of interest in this dataset is failure time. We drop patients with negative or zero survival time and leave with 2,396 patients. There are no missing values for the features in this dataset. The dataset can be downloaded from `http://pssp.srv.ualberta.ca` under "Public Predictors".

The Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (`SUPPORT`) dataset (Knaus et al., 1995) comprises 9,105 participants with the aim of examining survival outcomes and clinical decision-making for seriously ill hospitalized patients. The dataset consists of a proportion of missing values for a large proportion of features. The official website (`https://biostat.app.vumc.org/wiki/Main/SupportDesc`) for the SUPPORT dataset provides a guideline for imputing baseline physiologic features, we followed that procedure. For the rest features with missing values, we will also use the mean value imputation for continuous features and the mode value imputation for categorical features.

Surveillance, Epidemiology, and End Results (`SEER`) Program dataset (National Cancer Institute, DCCPS, Surveillance Research Program, 2015) is a comprehensive collection of data on cancer patients in the United States. This dataset, which encompasses about 49% of the U.S. population, includes vital information on patient diagnoses, survival times, and other relevant details sourced from various registries. Our study focuses specifically on three distinct subsets extracted from the main `SEER` dataset: `SEER-brain`, `SEER-liver`, and `SEER-stomach`. These subsets, respectively, contain data on patients diagnosed with brain, liver, and stomach cancers. The primary objective of the datasets is to model the time elapsed from the diagnosis of these patients until a failure event, such as death or disease progression, following the data preprocessing steps by Farrokh et al. (2024). The feature set chosen for analysis includes age, sex, behavior recode, combined summary stage, grade, RX summary, Summary stage 2000, SEER historic stage A, derived AJCC TNM and stage group for the 6th edition, and derived AJCC stage group for the 7th edition. In preparing each subset for analysis, we first eliminated any features with more than 70% missing or unknown values. We also excluded patients who had no recorded follow-up times, whether due to death or censoring. To ensure the uniqueness of the data, we removed any duplicate records, retaining only a single instance for each patient with matching clinical features and outcomes. In addition, patients with a recorded survival time of zero (indicating death or censoring on the same day as diagnosis) were also excluded from the study. The SEER cohort is available for application and download from `www.seer.cancer.gov`.

E.2. Evaluation Metrics

All five metrics (C-index, D-cal, KM-cal, IBS, MAE-PO) are implemented in the `SurvivalEVAL`[12] package (Qi et al., 2023b). The package can output the C-index, KM-cal, IBS, and MAE-PO scores directly. For D-cal, it generates the histogram and a p-value, we then need to calculate the D-cal statistics from the histogram, as discussed in Section 4.

E.3. Model Implementation Details

In this section, we will describe the implementation of the baseline models utilized in the performance comparison.

AFT (Stute, 1993) with Weibull distribution is a linear parametric model with two estimated coefficients (scale and shape), using a small L2 penalty on parameters during optimization. The model is implemented in `lifelines` packages (Davidson-Pilon, 2024).

GB (Hothorn et al., 2006) is an ensemble method with component-wise least squares as the base learner. We use the 100 boosting stages with partial likelihood loss for optimization and 100% subsampling for fitting each base learner. The model is implemented in `scikit-survival` packages (Pölsterl, 2020).

N-MTLR (Fotso, 2018) is a neural network extension of the multi-task logistic regression model (MTLR) (Yu et al., 2011). MTLR is the first survival model that estimated the probability of survival of patients at each of a vector of discretized time points $[t_1, t_2, \ldots, t_{\max}]$. To achieve that, MTLR set up a series of logistic regression models, for each time point. Here we implement the extended N-MTLR proposed by Fotso (2018), which essentially attached a multilayer perceptron (MLP) before the linear MTLR layer. The number of discrete times is determined by the square root of numbers of uncensored patients, and use quantiles to divide those uncensored instances evenly into each time interval, as suggested in (Jin, 2015; Haider et al., 2020). We reimplement the method based on the code provided in Kazmierski (2020), please see the code in

---

[12]`https://github.com/shi-ang/SurvivalEVAL`

the github repository.

DeepSurv (Katzman et al., 2018) is a neural network adaptation of the Cox proportional hazard model (CoxPH) (Cox, 1972). CoxPH is a semi-parametric model with a proportional hazard assumption. It consists of a population-level baseline hazard function (baseline hazard, which is non-parametric) and a partial hazard function (relative hazard, which is parametric). The original CoxPH model only predicts a relative hazard score (risk score) for each patient using the partial hazard function. To make ISD prediction, we use the Breslow method (Breslow, 1975) to estimate the population-level baseline hazard function. We add the ISD prediction to the method based on the code provided in Katzman et al. (2018), please see the code in the repository.

DeepHit (Lee et al., 2018) is also a discrete model, like N-MTLR. It models the probability density function (PDF) of the event for each individual (and PDF can be used to calculate the survival distribution accordingly), while N-MTLR models the survival distribution directly. Furthermore, apart from the standard likelihood loss, it also contains a ranking loss term which changes the undifferentiated indicator function in (1) with an exponential decay function. Due to this reason, DeepHit might exhibit stronger discrimination power than other models (on the other hand, lower calibration power than others). The number and locations of discrete times are determined the same way as N-MTLR model (square root of numbers of uncensored patients, and quantiles). The model is implemented in `pycox` packages (Kvamme et al., 2019).

CoxTime (Kvamme et al., 2019) is a non-proportional neural network extension of the CoxPH. Besides the MLP layer that DeepSurv added, CoxTime relieves the proportional assumption by allowing the baseline hazard function to model interactions between time t and features x_i (before it was only time). The model is implemented in `pycox` packages (Kvamme et al., 2019).

CQRNN (Pearce et al., 2022) is a quantile regression-based method, which means instead of predicting survival probability given a certain time, it predicts survival time given a certain quantile. It uses the Portnoy's pinball loss and an expectation-maximization step to estimate the model. Because the quantile regression-based methods do not enforce the prediction to be monotonic, we add the bootstrap-rearranging post processing (Chernozhukov et al., 2010) to the model's prediction. We reimplement the method based on the code provided in Pearce et al. (2022), please see the code in the repository.

"Dummy model" *Kaplan Meier (KM) (Kaplan & Meier, 1958)* is a non-parametric estimator to predict the survival distribution for a group of subjects. It is not a personalized prediction tool. As we discussed in Section 2, the KM model will theoretically achieve the perfect calibration. Therefore, we use this model as the **empirical lower-limits** for the calibration performance comparison. See also Appendix B for why the KM model theoretically achieves the perfect calibration. The model is implemented in `lifelines` packages (Davidson-Pilon, 2024).

In our study, we employ a uniform Multilayer Perceptron (MLP) architecture for a range of deep learning models, including *N-MTLR*, *DeepSurv*, *DeepHit*, *CoxTime*, *CQRNN*, and *LogNormalNN*[13]. Specifically, we configure the MLP with two layers, each containing 64 neurons, and build the model with batch normalization layers and rectified linear unit (ReLU) activation layers.

For the training process, we utilize Adam optimizer combined with an L2 penalty for weight decay to fine-tune the models. The learning parameters are set as follows: a learning rate of 0.001, a batch size of 256, and a dropout rate of 0.4. Additionally, we implement an early stopping mechanism across all deep learning models, which is based on performance validation using a separate validation dataset. This approach ensures that the effectiveness of the models is not compromised by variations in the MLP configuration or other optimization parameters.

For further details, please refer to the code in the repository.

F. Complete Experimental Results

F.1. Main Results: Compare CSD with Baseline Models

In Figure 9, we showcase the comprehensive outcomes of our empirical evaluation for the CSD framework. This detailed figure organizes the data by dataset in each row, ordered by sample size. The datasets are segmented into small, medium, and large categories by the two black dashed lines for the sake of clarity. Evaluation metrics are depicted in columns, with the C-index employing a positive scoring rule – where a higher value signifies superior performance, and the remaining

---

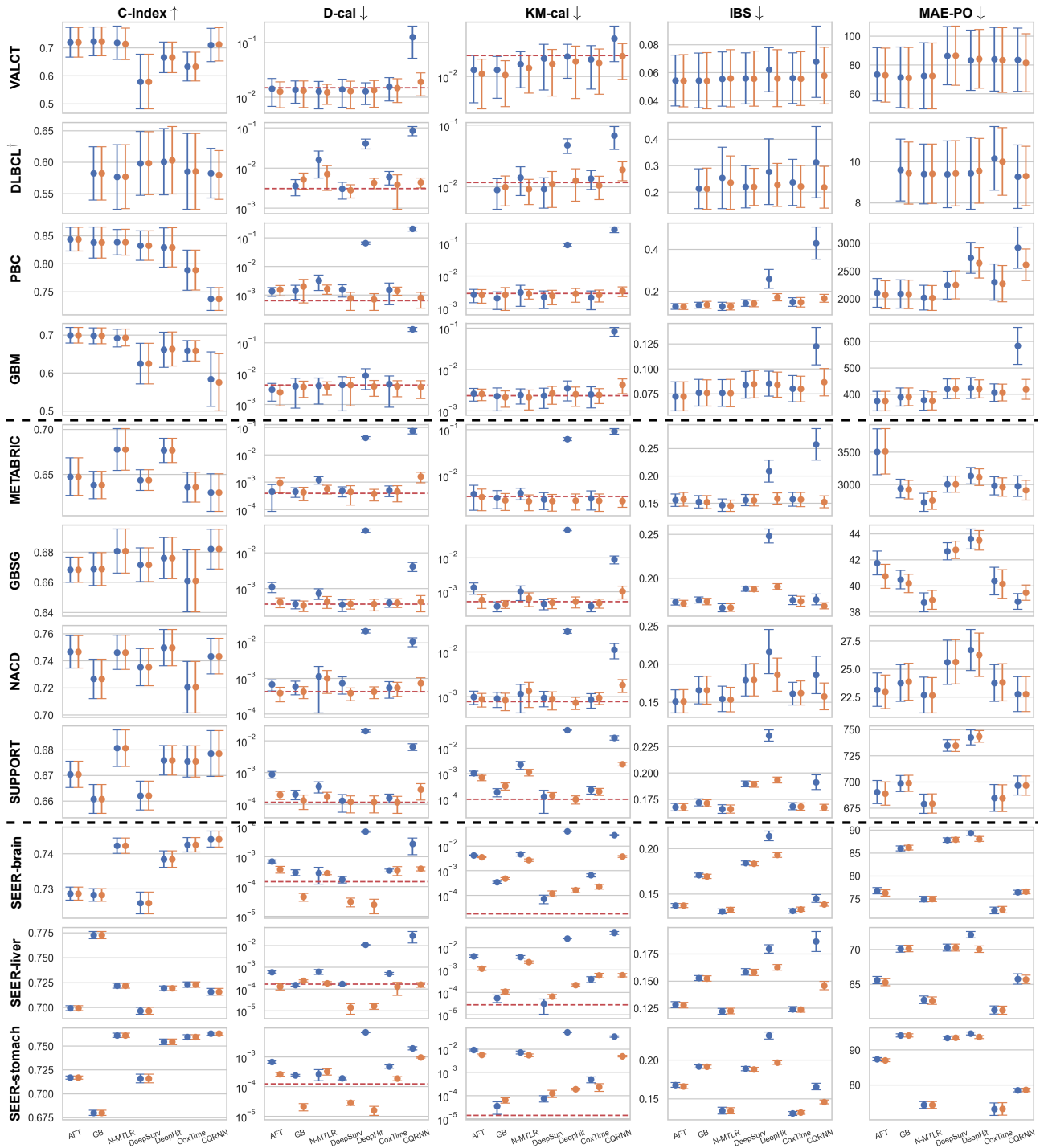[13]The introduction of the LogNormalNN model will be provided in subsequent Appendix F.2.

Figure 9. Complete empirical results. The error bars represent mean and 95% CI over ten runs, with blue denoting the non-CSD baseline and orange for the CSD-version. The red dashed line represents the mean calibration performance for KM estimator, serving as an empirical lower-limit. A higher C-index score indicates better performance, whereas lower scores are preferable for other metrics. †AFT is fail to converge on DLBCL dataset due to the large feature-sample-ratio.

metrics using negative scoring rules. Within the D-cal and KM-cal diagrams, red dashed lines highlight the empirical lower limits, which represent the mean calibration performance of a "dummy" KM model. In each subfigure, the performance of the standard non-CSD baseline is illustrated with blue bars, while the CSD-enhanced versions are depicted with orange bars, facilitating a clear comparison between the two.

Figure 10 and 11 present some qualitative calibration results for the CSD methods. Specifically, Figure 10 shows the P-P plots for the non-CSD baselines and post-CSD models, and Figure 11 shows the predictions of the average survival curves for the non-CSD baselines and the post-CSD models. Here, we use the DeepHit model as the baseline. We can see that for small and medium datasets, DeepHit tends to overestimate the risks, and our CSD process can calibrate this overestimation to almost perfectly calibrative estimates. For large datasets, DeepHit tends to underestimate the lower risks and overestimate the higher risks, and our CSD can also calibrate this estimation to almost perfect calibration.
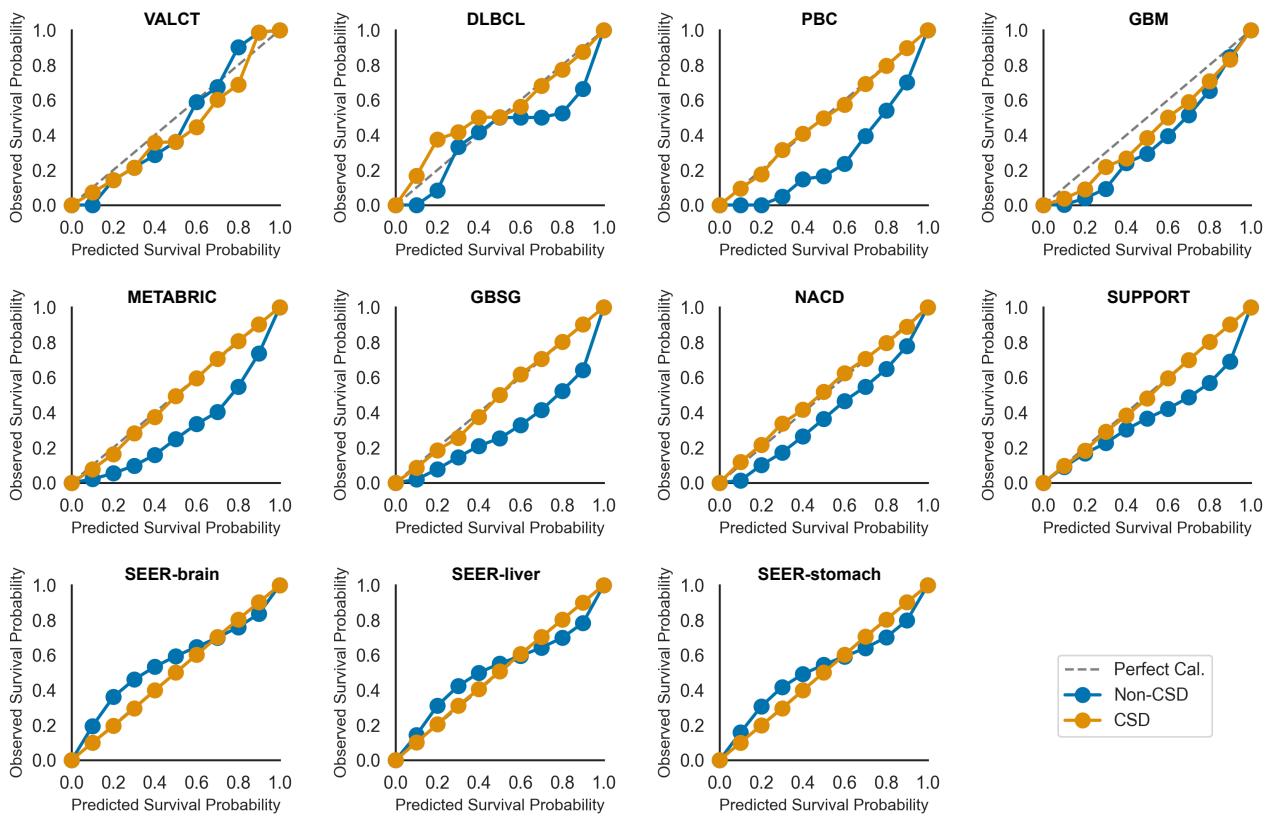
Figure 10. P-P plots assessing the calibration performance of the CSD and non-CSD methods. Here we use DeepHit model as the baseline (non-CSD) method, and the CSD method uses KM-sampling with 19 percentile levels. The dashed line represents perfect calibration.

F.2. Compare with Objective-Based Methods

To compare the proposed CSD with the objective-function based methods, we test the performance on a parametric log-normal model with neural networks (*LogNormalNN*). *LogNormalNN* uses two independent neural networks to predict the personalized mean and variance of the log-normal distribution, respectively. This is the model used in both Avati et al. (2020) and Goldstein et al. (2020) as the baseline, which best shows the effectiveness of their methods.
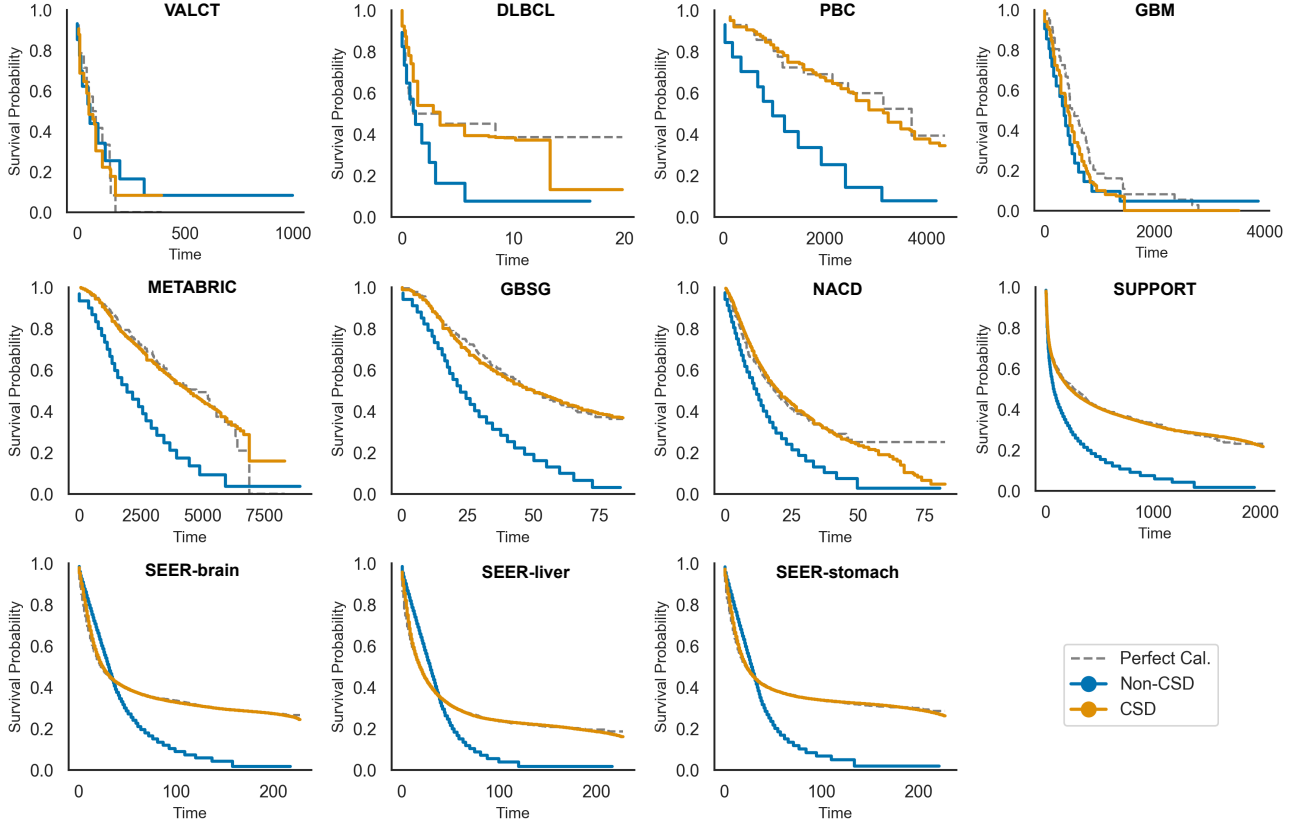
Figure 11. Compare the average predicted curve from the CSD and non-CSD methods with the true survival curve. Here we use DeepHit model as the baseline (non-CSD) method, and the CSD method uses KM-sampling with 19 percentile levels. The dashed line represents the Kaplan-Meier curve of the test set.

Under *conditional independent censoring* assumption, the baseline *LogNormalNN* optimizes the log-likelihood function:

$$\mathcal{L}_{\text{likelihood}}(\mathcal{M}) = \log \prod_{i=1}^{N} \hat{f}_{\mathcal{M}}(t = t_i \mid \boldsymbol{x}_i)^{\delta_i} \hat{S}_{\mathcal{M}}(t = t_i \mid \boldsymbol{x}_i)^{1-\delta_i}$$

$$= \sum_{i=1}^{N} \delta_i \times \log \hat{f}_{\mathcal{M}}(t = t_i \mid \boldsymbol{x}_i) + (1 - \delta_i) \times \log \hat{S}_{\mathcal{M}}(t = t_i \mid \boldsymbol{x}_i), \tag{29}$$

where $\hat{f}_{\mathcal{M}}(t = t_i \mid \boldsymbol{x}_i)$ is the predicted probability density function (PDF) which can be derived using ISD: $\hat{f}_{\mathcal{M}}(t = t_i \mid \boldsymbol{x}_i) = \frac{-\partial \hat{S}_{\mathcal{M}}(t=t_i\mid\boldsymbol{x}_i)}{\partial t}$.

Continuous Ranked Probability Score (CRPS) for survival is proposed by Avati et al. (2020) to make sharp and calibrated ISD prediction. Specifically, they proposed to directly use the CRPS as the objective function:

$$\mathcal{L}_{\text{CRPS}}(\mathcal{M}) = \sum_{i=1}^{N} \left(\int_{t=0}^{t_i} \left(1 - \hat{S}_{\mathcal{M}}(t \mid \boldsymbol{x}_i) \right)^2 dt + \delta_i \int_{t=t_i}^{\infty} \hat{S}_{\mathcal{M}}(t \mid \boldsymbol{x}_i)^2 dt \right). \tag{30}$$

As many other research points out (Qi et al., 2023a), the CRPS loss is essentially the IBS score without the IPCW weighting. Therefore, because IBS can be decomposed into an integrated 1-cal component, optimizing CRPS should be able to lead to a calibrated prediction.

A follow-up study by Goldstein et al. (2020) proposed a differentiable D-cal (X-cal) as part of the objective function with (29). For the mutually exclusive percentile intervals $\mathcal{I} = \{[0, 0.1], [0.1, 0.2], \dots, [0.9, 1]\}$, the X-cal can be calculated
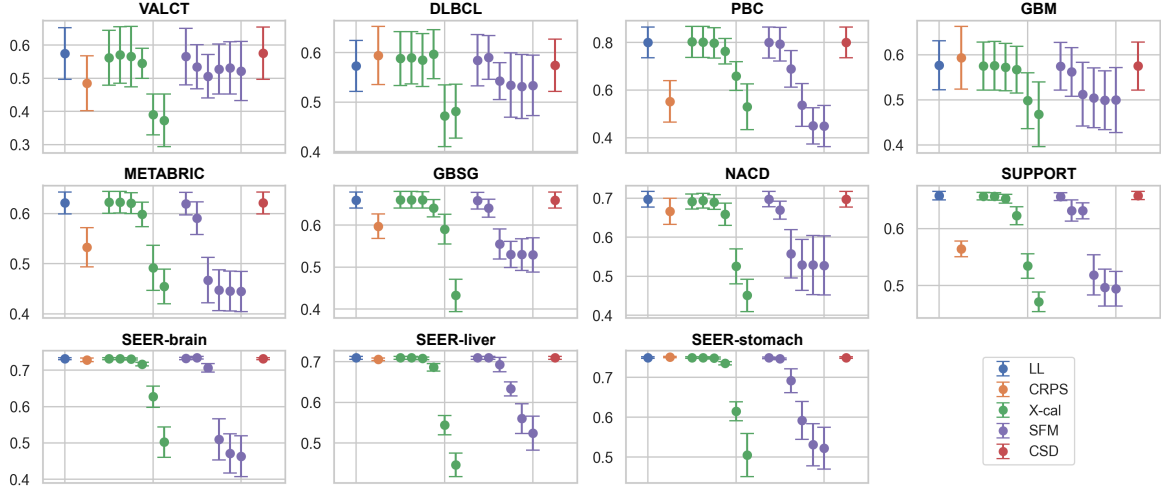
Figure 12. C-index comparison of CSD and objective-based methods. Higher values indicate superior performance. The baseline (blue bars) uses likelihood loss (LL), the yellow bars denote CRPS, the green bars denote X-cal, purple denotes SFM, and red bars denote CSD. For the X-cal and SFM methods, we gradually increase the weight for the calibration loss.

using:

$$\mathcal{L}_{\text{X-cal}}(\mathcal{M}) = \sum_{I \in \mathcal{I}} \left(\sum_{i=1}^{N} \zeta_\gamma \left(\hat{S}_\mathcal{M}(t_i \mid \boldsymbol{x}_i); I \right) - |I| \right)^2,$$

where

$$\zeta_\gamma(u; I = [a, b]) = \frac{1}{1 + \exp(-\gamma(u - a)(b - u))},$$

is an approximation of the indicator function, which checks whether the predicted survival probability at the event time $\hat{S}_\mathcal{M}(t_i \mid \boldsymbol{x}_i)$ is inside the percentile interval I.

Chapfuwa et al. (2020a) proposed the survival function matching (SFM) loss as the penalty term for their survival models. The core component of the SFM loss is similar to the KM-cal (defined in Section 4.2), with a slight modification from the squared error to absolute error. Note that our implementation of SFM loss differs from the original by Chapfuwa et al. (2020a) in two aspects: (1) The SFM loss was proposed in conjunction with a model that makes time prediction. In our baseline models, we need to predict survival distributions, which requires us to make substantial modifications to the SFM calculation. (2) The original was based on TensorFlow-based code, while our implementation is in PyTorch.

We reimplemented the baseline *LogNormalNN* models with log-likelihood loss, CRPS loss, and X-cal loss, based on the code provided in (Goldstein et al., 2020). We then test how our proposed CSD performs compared to these two objective function-based methods.

Figure 12, 13 and 14 provide a comparative analysis of the performance between the CSD and the objective-based methods, and various objective-based methods, focusing on three key metrics: C-index, D-cal, and KM-cal.

The performance of different variations of the LogNormalNN model is depicted through color-coded bars:

- The blue bars represent the *LogNormalNN* model optimized using a likelihood function, illustrating the base performance without additional calibration or optimization criteria.

- The yellow bars indicate the *LogNormalNN* model enhanced by the Continuous Ranked Probability Score (CRPS) loss function, showcasing the impact of optimizing for probabilistic forecasting accuracy.

- The green bars denote the *LogNormalNN* model that integrates optimization for both the likelihood and an additional D-calibration penalty term (X-cal). Here, the emphasis is on the dynamic adjustment of the weight for the X-cal term, which is systematically increased from 10^0 to 10^5, illustrating the effect of calibration emphasis in model training.
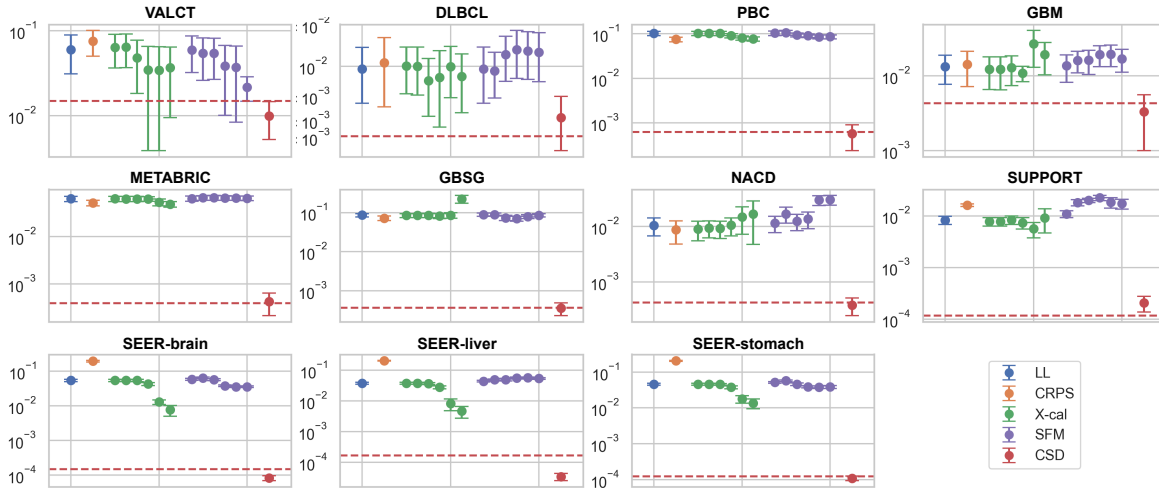
Figure 13. D-calibration comparison of CSD and objective-based methods. Lower values indicate superior performance. The baseline (blue bars) uses likelihood loss (LL), the yellow bars denote CRPS, the green bars denote X-cal, the purple bars denote SFM, and the red bars denote CSD. For the X-cal and SFM method, we gradually increase the weight for the calibration loss. The red dashed line represents the mean calibration performance of the KM estimator, serving as an empirical lower-limit.

- The purple bars denote the *LogNormalNN* model that integrates optimization for both the likelihood and an additional KM-calibration penalty term (SFM). Here, the emphasis is on the dynamic adjustment of the weight for the SFM term, which is systematically increased from 10^0 to 10^5, illustrating the effect of calibration emphasis in model training.

- Lastly, the red bars illustrate the LogNormalNN model initially optimized using the likelihood function and subsequently adjusted using a post-processing CSD method, demonstrating the potential benefits of incorporating CSD adjustments after the initial model training.

F.3. Ablation Studies

F.3.1. ABLATION #1: HANDLING RIGHT-CENSORING

We present an ablation study in Figure 15 that investigates various approaches to managing censorship within the Conformalized Survival Distribution (CSD) framework. These approaches include uncensored, margin, pseudo-observation (PO), and Kaplan-Meier (KM) sampling. Similar to Figure 9, each row in this figure corresponds to a different dataset, arranged in ascending order of sample size, while the columns represent various evaluation metrics. The performance of each method is color-coded in the subfigures: uncensored (blue), margin (yellow), PO (green), and KM sampling (red).

TL;DR KM sampling (red bars) demonstrates superior calibration performance over the other methods (uncensored, margin, and PO), particularly in datasets of medium to large size.

The first columns of Figure 15 show negligible differences in the C-index between the non-CSD models and their CSD counterparts. As to the calibration, the censoring method has difference impact for different types of datasets.

- In datasets with very low censorship rates, such as VALCT, which contains only 9 censored subjects, the uncensored approach (blue bars) performs optimally. Indicating that omitting these few censored subjects from the CSD analysis has minimal effect on the overall data distribution.

- For other small datasets (DLBCL, PBC, and GBM), the performances of margin (yellow), PO (green), and KM sampling (red) are similar, likely due to inaccurate KM curve estimations from small sample sizes, affecting the KM sampling's "best guess" distribution.

- For medium and large datasets, KM sampling method (red bars) consistently (and significantly) works better than the others.
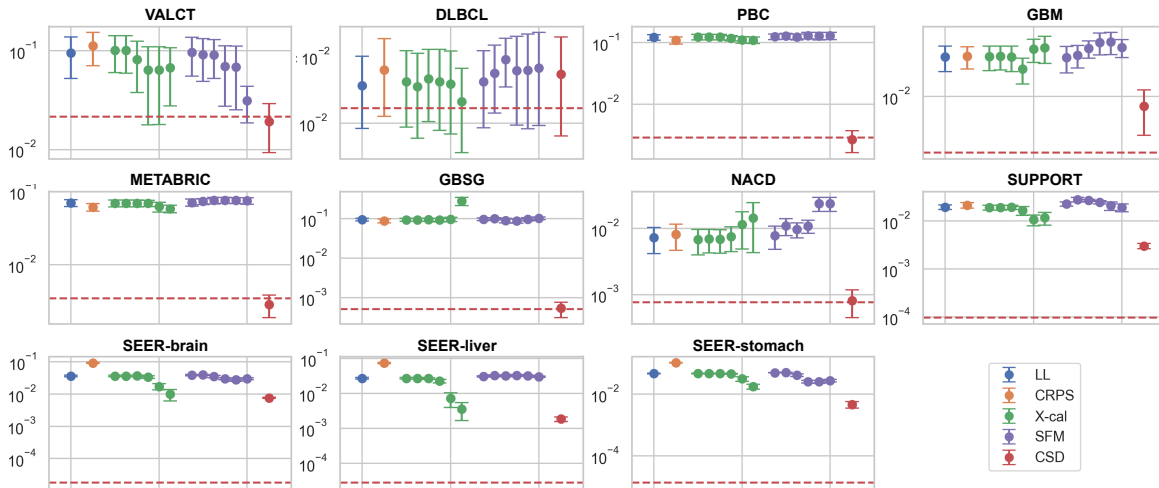
Figure 14. KM-calibration comparison of CSD and objective-based methods. Lower values indicate superior performance. The baseline (blue bars) uses likelihood loss (LL), the yellow bars denote CRPS, the green bars denote X-cal, the purple bars denote SFM, and the red bars denote CSD. For the X-cal and SFM method, we gradually increase the weight for the calibration loss. The red dashed line represents the mean calibration performance of the KM estimator, serving as an empirical lower-limit.

Regarding IBS performance, margin, PO, and KM sampling methods exhibit comparable outcomes within the 95% confidence interval, yet all surpass the uncensored method. The MAE-PO shows no significant performance differences among the four methods.

F.3.2. ABLATION #2: CONFORMAL SET SPLITTING

Figure 16 explores the performance of different strategies for dividing the conformal set. We compare two policies (see also Section 3.3): (1) allocating a separate 10% of the training set as the conformal set and (2) merging this separate conformal set with the training set. Each row in the figure corresponds to a dataset, arranged by sample size, with evaluation metrics displayed in columns. The performance of policy 1 is represented by blue bars, and policy 2 by yellow bars.

TL;DR   For smaller or medium-sized datasets, policy 2 is more effective, while for larger datasets, policy 1 is preferred due to comparable performance and enhanced computational efficiency.

The figure indicates that the choice of data splitting policy has minimal effect on the C-index, IBS, and MAE-PO. However, for calibration performance in large datasets, using a separate validation set as the conformal set proves adequate. In such cases, a 10% validation set yields at least 1000 conformal samples, which aligns with previous research findings (Angelopoulos et al., 2023). For smaller or medium-sized datasets, where a 10% validation set might result in only a few hundred samples, merging the validation and training sets for the conformal set (policy 2) is beneficial, significantly enhancing calibration performance. For large datasets, the calibration outcomes of both policies are similar, thus we recommend policy 1 (a separate conformal set) to save computational resources(faster and less memory usage).

F.3.3. ABLATION #3: NUMBER OF PERCENTILES

In Figure 17, we examine the effect of varying the number of discretized percentiles on CSD performance. We consider several options for $\mathcal{P}$: 9 percentiles (10% to 90% at 10% intervals), 19 percentiles (5% to 95% at 5% intervals), 39 percentiles (2.5% to 97.5% at 2.5% intervals), 49 percentiles (2% to 98% at 2% intervals), and 99 percentiles (1% to 99% at 1% intervals). This figure also presents one dataset in each row, ordered by sample size, and evaluation metrics are depicted in columns. The performance of different numbers of percentile is color-coded in the subfigures: 9 (blue), 19 (yellow), 39 (green), 49 (red), and 99 (purple).

TL;DR   The number of percentiles has minimal impact on the performance.

The choice of percentile number has little to no effect on survival metric performance. However, for large datasets
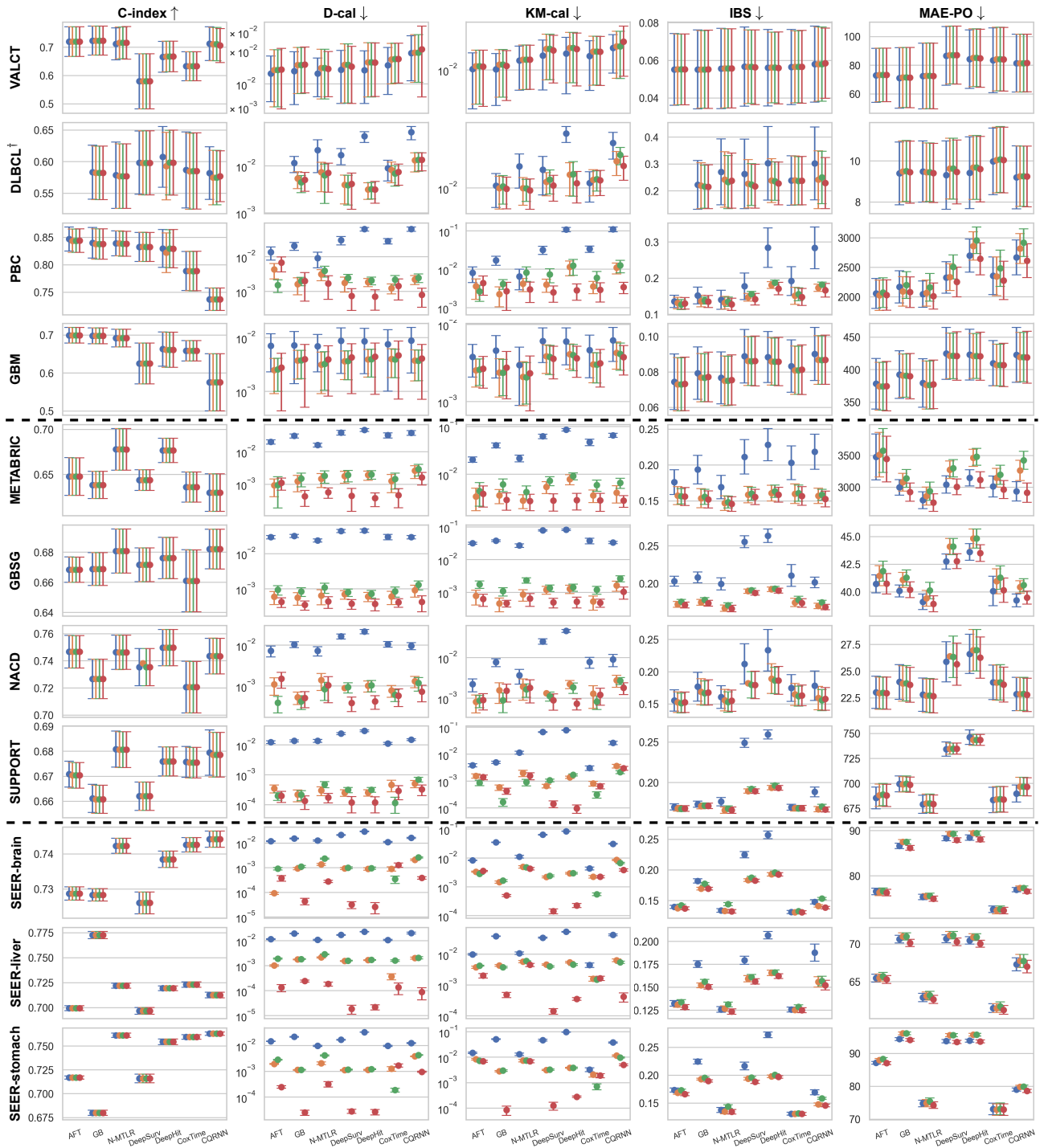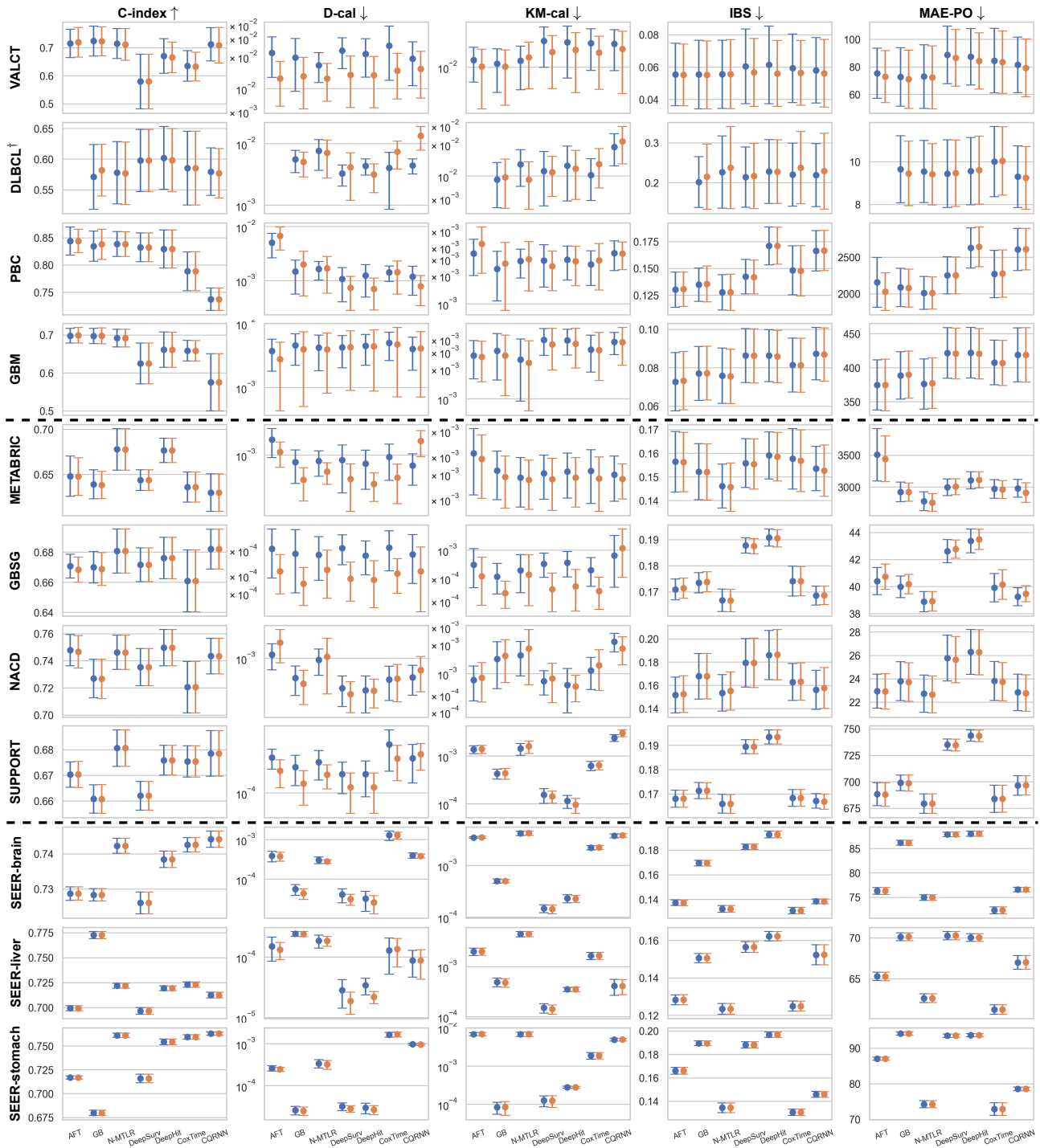
Figure 15. Comparison results for **ablation #1: way of handling right-censoring**. The error bars represent mean and 95% CI over ten runs, with blue denoting the uncensored method, yellow for margin, green for PO, and red for KM-sampling. A higher C-index score indicates better performance, whereas lower scores are preferable for other metrics. [†]AFT is fail to converge on DLBCL dataset due to the large feature-sample-ratio.
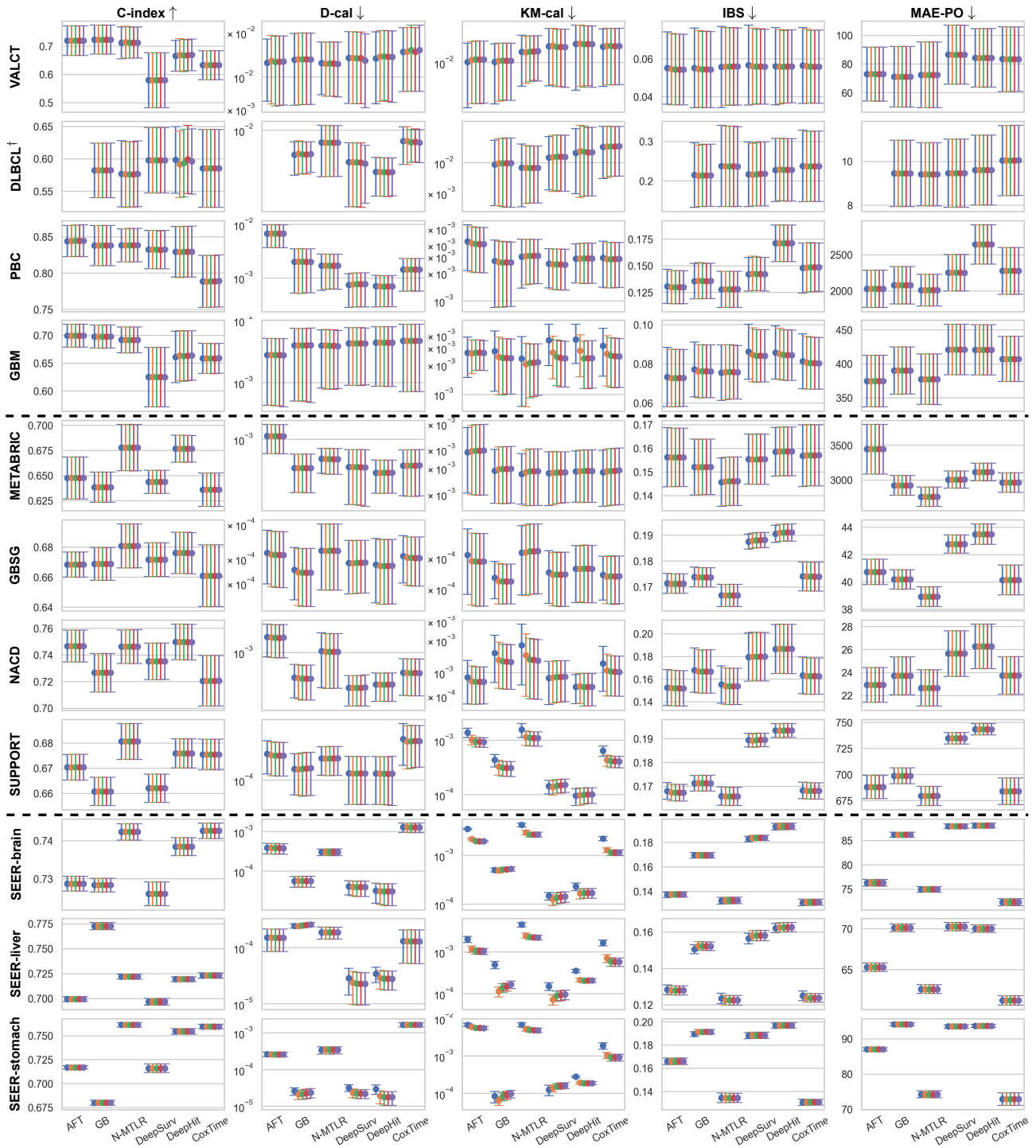
Figure 16. Comparison results for **ablation #2: way of splitting conformal set**. The error bars represent mean and 95% CI over ten runs, with blue denoting using solely the separate conformal set and yellow for combining the separate conformal set with the training set together for comformalization. A higher C-index score indicates better performance, whereas lower scores are preferable for other metrics. [†]AFT is fail to converge on DLBCL dataset due to the large feature-sample-ratio.

Figure 17. Comparison results for **ablation #3: number of discretized percentiles**. The error bars represent mean and 95% CI over ten runs, with blue denoting 9 percentiles, yellow for 19 percentiles, green for 39 percentiles, red for 49 percentiles, and purple for 99 percentiles. A higher C-index score indicates better performance, whereas lower scores are preferable for other metrics. [†]AFT is fail to converge on DLBCL dataset due to the large feature-sample-ratio.

(`SEER-brain`, `SEER-liver`, and `SEER-stomach`), KM-cal results (third column) indicate that using only 9 percentiles leads to significantly poorer performance, likely due to information loss with such coarse discretization. For practical applications, selecting either 9 or 19 percentiles is advisable to maintain a balance between calibration accuracy and computational demand.

F.4. Computational Analysis

One limitation of the CSD method is that it cannot perform the conformalization step in batch mode, which is computationally inefficient.

Specifically, the quantile operation in (5) requires storing the conformity score for all data in the conformal set in advance, followed by computing the quantile amount on a substantial array. This will introduce two complexities: space complexity and time complexity.

Although there are ways to perform quantile calculations in batch modes, this will only approximate the true quantile values. To achieve precise quantile values, it is necessary to store the conformity scores beforehand, which incurs complexity overhead.

F.4.1. SPACE COMPLEXITY

To simplify notation, let $|\mathcal{D}_{con}|$ be the number of subjects in the conformal datasets, $\mathcal{P} = \{\rho_1, \rho_2, \ldots\}$ be the predefined discretized percentiles, so that $|\mathcal{P}|$ be the number of predefined percentiles, and R be the repeat parameter for KM-sampling. Table 3 presents the space complexity for storing the conformity scores. Compared with the uncensored method, the KM-sampling method will asymptotically cause around R times memory overhead.

Table 3. The space complexity comparison between Conformal Quantile Regression (CQR), CSD-uncensored, CSD-margin, CSD-PO, and CSD-KM-sampling. [†]This assumes the number of uncensored subjects in the datasets is $O(|\mathcal{D}_{con}|)$.

	CQR	CSD-uncensored	CSD-margin	CSD-PO	CSD-KM-sampling
Space Complexity	$O(|\mathcal{D}_{con}|)$	$O(|\mathcal{D}_{con}| \cdot |\mathcal{P}|)^{\dagger}$	$O(|\mathcal{D}_{con}| \cdot |\mathcal{P}|)$	$O(|\mathcal{D}_{con}| \cdot |\mathcal{P}|)$	$O(|\mathcal{D}_{con}| \cdot |\mathcal{P}| \cdot R)$

F.4.2. TIME COMPLEXITY

Figure 18 and 19 empirically compare the running times of the CSD method versus various baseline methods. Each value (dot) in Figure 18 represents an average extra running time of the method, over the associated non-CSD baselines, over 10 random splits. Each value (dot) in Figure 19 represents an average relative running time of the method, compared with the associated non-CSD baselines, over 10 random splits. We see that the running time of the CSD framework is independent of the baseline algorithms. For survival algorithms that require less training time (*e.g.*, AFT), the CSD post-process with KM-sampling may significantly amplify the running time (> 100 times compared with non-CSD baselines for SEER datasets, and ~ 16 times compared with CSD-uncensored for SEER datasets). For survival algorithms that require long training times (*e.g.*, GB, CQRNN, etc.), the relative time for the CSD post-process can sometimes be ignored.
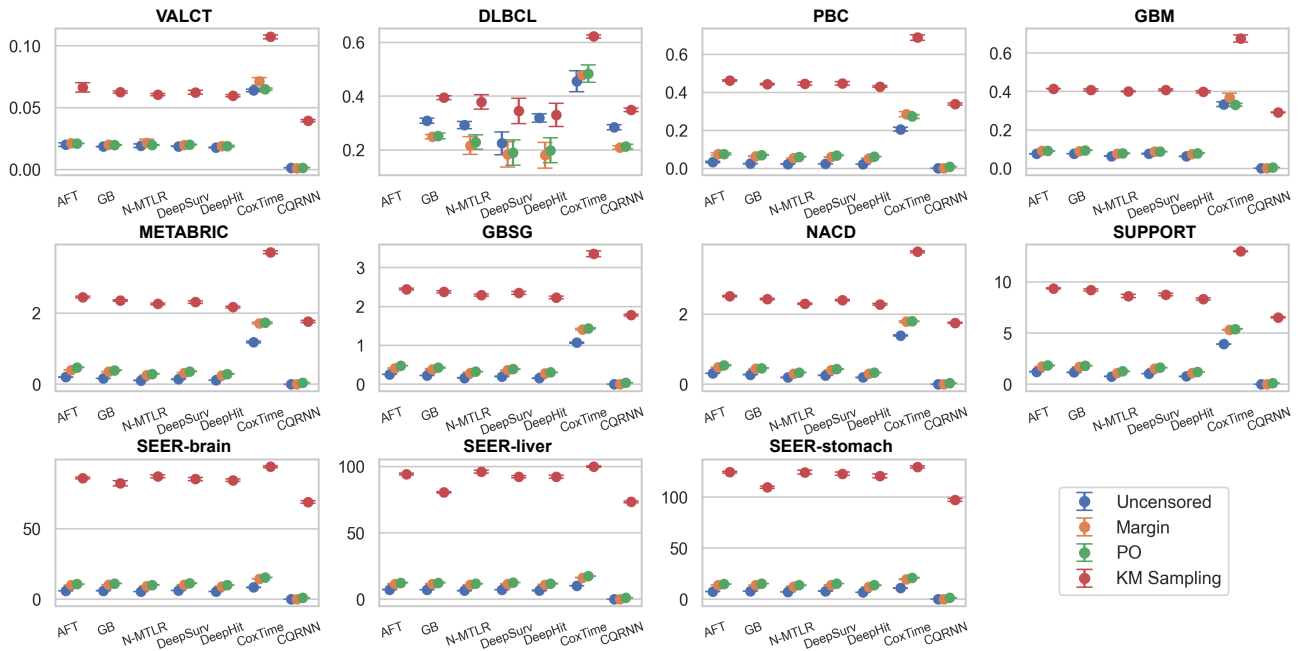
Figure 18. Extra running time (mean with 95% CI) of the CSD method (with four different methods for handling censorship) on top of the baseline methods.
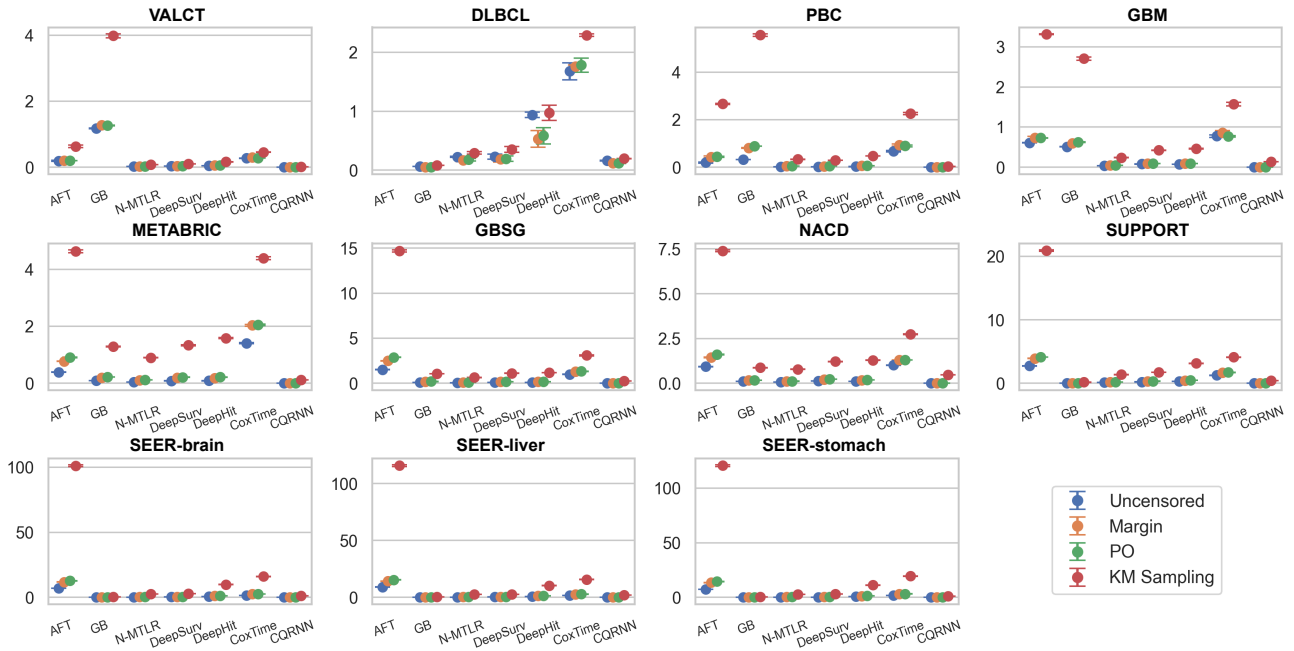

Figure 19. Relative running time (mean with 95% CI) of the CSD method (with four different methods for handling censorship) on top of the baseline methods.