An Effective Meaningful Way to Evaluate Survival Models

Shi-ang Qi [1] **Neeraj Kumar** [2] **Mahtab Farrokh** [1] **Weijie Sun** [1] **Li-Hao Kuan** [1]
Rajesh Ranganath [3] **Ricardo Henao** [4] **Russell Greiner** [1 2]

Abstract

One straightforward metric to evaluate a survival prediction model is based on the Mean Absolute Error (MAE) – the average of the absolute difference between the time predicted by the model and the true event time, over all subjects. Unfortunately, this is challenging because, in practice, the test set includes (right) censored individuals, meaning we do not know when a censored individual actually experienced the event. In this paper, we explore various metrics to estimate MAE for survival datasets that include (many) censored individuals. Moreover, we introduce a novel and effective approach for generating realistic semi-synthetic survival datasets to facilitate the evaluation of metrics. Our findings, based on the analysis of the semi-synthetic datasets, reveal that our proposed metric (MAE using pseudo-observations) is able to rank models accurately based on their performance, and often closely matches the true MAE – in particular, is better than several alternative methods.

1. Introduction

Survival prediction models are often used to predict how long an individual will survive – or in general, the time until an individual experiences a specific event. These have many applications in medicine (time to death, relapse, or recovery), business (time to service cancellation), and social sciences (war or peace duration). Unlike typical regression problems, one challenge of training and evaluating a survival prediction model is that survival datasets often contain censored observations (Klein & Moeschberger, 2003). This paper focuses on the most prevalent type of censorship, right-censoring, which provides only a lower bound on event time. For example, consider a patient who entered a 5-year study at its beginning and was still alive at the end of the study. We only know this patient survived for at least 5 years, but do not know whether that patient lived a day, a month, or 20 years after the study ended.

Numerous statistical and machine learning models have been developed to estimate survival outcomes from input features. In this paper, we focus on a class of survival models that learns to compute an individual's survival distribution (ISD) (Haider et al., 2020): a probability curve for all future time points for a specific patient. Note that one can use an individual's ISD to (i) compute that individual's expected time-to-event, (ii) provide single-time estimations (*e.g.*, 5-year cancer onset probability, like the Gail model (Costantino et al., 1999)), or (iii) estimate a risk score (like Cox Proportional Hazard model (Cox, 1972)). Obviously, the computation of aforementioned quantities from ISDs will only be reliable if a model's predicted ISD is "accurate". One question that plagues the survival prediction community is: *What is an appropriate scoring rule for evaluating survival models?* The answer may vary from task to task – *e.g.*, the concordance index (C-index) (Harrell Jr et al., 1996) is useful in several clinical problems that require comparing patients, such as prioritizing patients for liver transplants (a patient at the highest risk of death should be treated first). Figure 1 shows a visualization of six typical evaluation metrics (see discussion in Section 3).

However, the Mean Absolute Error (MAE) seems to be the most intuitive metric for evaluating survival prediction models, as it measures the expected difference between predicted and actual event times. While this difference is trivial to compute for uncensored individuals, it is problematic for censored individuals. Thus, researchers have proposed several versions of MAE to handle censored individuals, such as MAE-uncensored and MAE-hinge (Haider et al., 2020). However, these approaches often produce biased MAEs for high censoring.

Here, we propose an MAE-inspired evaluation metric using the pseudo-observation de-censoring technique (Andersen et al., 2003), MAE-PO (defined in Section 3.6). Our metric

[1]Computing Science, University of Alberta, Edmonton, Canada [2]Alberta Machine Intelligence Institute, Edmonton, Canada [3]Computer Science & Center for Data Science, New York University, New York City, USA [4]Biostatistics & Bioinformatics, Duke University, Durham, USA. Correspondence to: Shi-ang Qi <shiang@ualberta.ca>, Russell Greiner <rgreiner@ualberta.ca>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

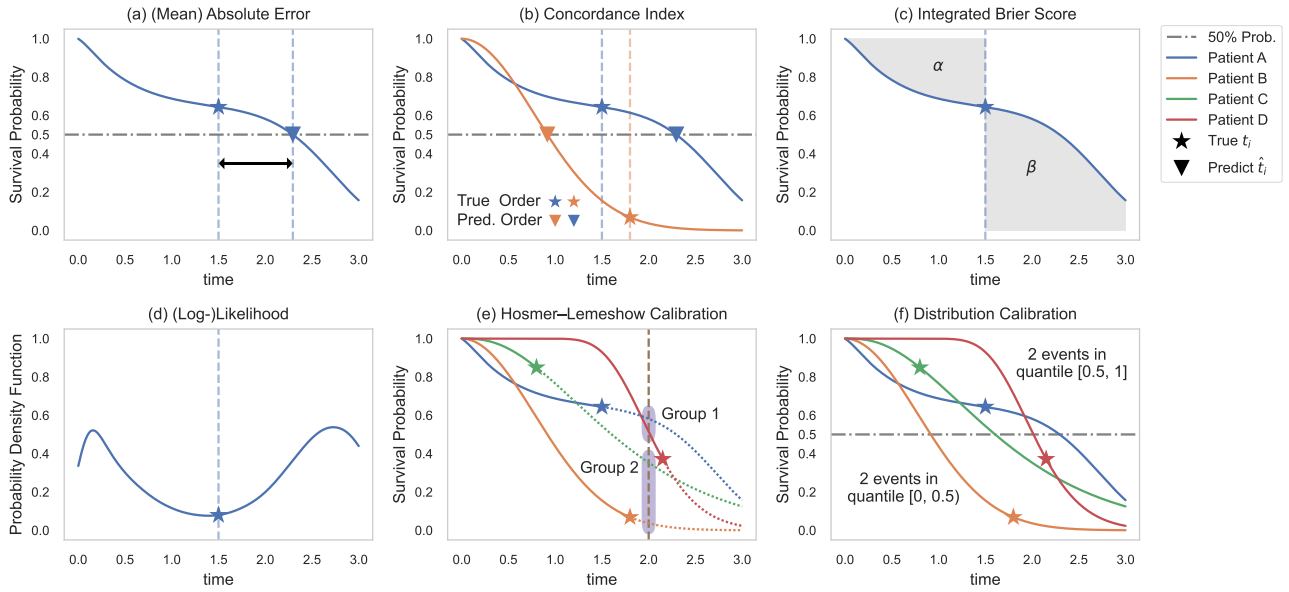

Figure 1. Illustration of six common evaluation metrics using four uncensored subjects. Note that the y-axis of (d), Log-Likelihood, is the probability density function, while the others are survival functions. Here we use the median survival time of the survival function as the predicted time for a more intuitive visualization. Please refer to Appendix C for the detailed discussion of these metrics.

effectively deals with a censored subject by (a) estimating its pseudo-observation value for the censored subject by calculating how much it counts toward the group-level Kaplan-Meier (KM) estimator (Kaplan & Meier, 1958), and (b) employing a weighted scheme to represent the confidence of pseudo-observation estimation. Our main contributions are:

1. We show that MAE is fundamentally different from the other standard evaluation measures (Section 3 and Appendix C);
2. We theoretically prove some relevant properties of pseudo-observation, which helps justify the MAE pseudo-observation approach for evaluating the time-to-event prediction (Section 3.6 and Appendix D);
3. We provide a way to produce semi-synthetic (but realistic) survival datasets, where we know the true survival time of all instances (Section 4), which is essential for evaluating evaluation methods;
4. We compare six versions of MAE-inspired metrics (on various realistic semi-synthesized datasets) to determine which metric (i) ranks the ISD models in a way that is close to the true MAE, or (ii) reports the error closest to the true MAE (Section 4);
5. We show that, in many situations, MAE-PO is the most suitable estimator. We also provide a code base for these MAE approaches, for this and other variants.

We believe this is the first methodology for identifying the appropriate evaluation metrics using meaningful semi-synthetic datasets.

2. Preliminaries

In general, a survival dataset contains N time-to-event tuples, $\mathcal{D} = \{(\boldsymbol{x}_i, t_i, \delta_i)\}_{i=1}^{N}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ represents the observed d-dimensional features for the i-th subject, $t_i \in \mathbb{R}_+$ denotes the event or censor time, and $\delta_i \in \{0, 1\}$ is a censor/event indicator where $\delta_i = 0$ means the subject is right-censored (subject has not experienced an event at time t_i) and $\delta_i = 1$ means subject died at time t_i. Conceptually, we assume subject i has an event time e_i and a censoring time c_i, and assign $t_i \triangleq \min\{e_i, c_i\}$ and $\delta_i \triangleq \mathbb{1}[e_i \leq c_i]$. We assume *independent censoring*: e_i and c_i are assumed independent, conditional on the covariates $\boldsymbol{x}_i$.

ISD models target the individual survival distribution $S(\, t \mid \boldsymbol{x}_i\,) = P(T > t \mid \mathbf{X} = \boldsymbol{x}_i)$. Further, $F(t \mid \boldsymbol{x}_i) = 1 - S(\, t \mid \boldsymbol{x}_i\,)$ denotes the (conditional) cumulative density function and $f(t \mid \boldsymbol{x}_i) = \frac{-\partial S(\, t \mid \boldsymbol{x}_i\,)}{\partial t}$ denotes the (conditional) probability density function (PDF) of the event time T. A predicted event time $\hat{t}_i$ can then be represented by either mean (expected) or median survival time, respectively:

$$\hat{t}_{i,\text{mean}} = \mathbb{E}_t[\, S(\, t \mid \boldsymbol{x}_i\,)\,] = \int_0^\infty S(\, t \mid \boldsymbol{x}_i\,)\, dt\,, \quad (1)$$

$$\hat{t}_{i,\text{median}} = \text{median}(S(\, t \mid \boldsymbol{x}_i\,)) = S^{-1}(\, \tau = 0.5 \mid \boldsymbol{x}_i\,)\,, \quad (2)$$

where $\tau \in [0, 1]$ represents the quantile probability level. If necessary, a linear extrapolation (extrapolate from the initial to the last time point of the ISD curve) might be applied to ISDs that do not reach 0% or 50% quantile probability.

Suppose we have two models, each producing a predicted ISD curve for every individual $\{S_{\text{M1}}(\, t \mid \boldsymbol{x}_i\,)\}_{i=1}^N$ and $\{S_{\text{M2}}(\, t \mid \boldsymbol{x}_i\,)\}_{i=1}^N$. Our goal is to find an MAE-inspired metric $\mathcal{R}(\cdot)$ that, given an ISD distribution $S_{\text{m}}(\, t \mid \boldsymbol{x}_i\,)$, the observed time t_i and event indicator δ_i of a subject, returns a good approximation to the true MAE[1]. We can then consider the average score, over the dataset, $\mathbb{E}[\,\mathcal{R}(\cdot)\,]$. Our goal is a measure whose average is as close as possible to the true MAE. To be specific, we hope that $\mathbb{E}[\,\mathcal{R}(\,S_{\text{m}}(\, t \mid \boldsymbol{x}_i\,),\, t_i,\, \delta_i\,)\,]$, is close to the true MAE score for each model, or at least, can correctly rank the two models.

3. Handling Right-Censoring in MAE

Survival prediction is like regression as it predicts a real number (the subject's time of the event) from a description of that subject, $\boldsymbol{x}_i$. Given this commonality, we want to evaluate survival prediction models using measures for evaluating regression tasks, such as mean absolute error (MAE). As suggested in Figure 1(a), the absolute error for an uncensored subject is the absolute difference between the true event time and the predicted time:

$$\mathcal{R}_{\text{MAE}}(\,\hat{t}_i,\, t_i,\, \delta_i = 1\,) \;=\; |t_i - \hat{t}_i|, \tag{3}$$

where $\hat{t}_i$ is the median survival time[2] from Equation 2. MAE is formally a *negative scoring rule*, as more precise models have smaller values.

It is vital to have a thorough grasp of the limitations of all types of evaluation metrics when selecting metrics for model optimization, and separately for model evaluation. In this section, we briefly motivate that the MAE score is the most appropriate metric if the objective is to quantify the time-to-event accuracy. Figure 1 shows that, C-index measures the ranking accuracy by assessing if the order of true event times (stars) is concordant with the order of predicted event times (triangles) (b); integrated Brier score (Graf et al., 1999) measures the accuracy of the predicted probabilities over all times via the weighted squared error of the shaded regions (c); log-likelihood measures the magnitude of the predicted probability at event times (d); Hosmer-Lemeshow calibration (Hosmer & Lemesbow, 1980) assesses if the expected and observed event rates are statistically similar (e); and Distribution calibration (D-calibration) (Haider et al., 2020) examines if the proportion of subjects who dies in each quantile interval is uniformly distributed (f). Appendix C shows that the time-to-event precision that MAE captures cannot be covered by other metrics. Moreover, the model preference between MAE and each other metrics can be completely different – *i.e.*, a model can have perfect C-index but terrible MAE, while another model is good at MAE but has poor C-index score.

However, evaluating (and also learning) survival prediction models is challenging when the dataset includes right-censored subjects, meaning it is critical to define learning (and evaluation) algorithms that can appropriately incorporate the censored subjects. This section begins by reviewing six MAE variants that claim to handle censored subjects, including three novel MAE variants. Section 4 then presents an empirical comparison of all these versions.

Unless otherwise specified, we will use the median survival time (Equation 2) as the default method to calculate the predicted time of the ISD curves. This is because (i) the combination of MAE and median survival time is a proper scoring rule (Theorem C.1); and (ii) linear extrapolation is only required for curves that do not reach 50% for median survival time, whereas it is required for every curve that does not reach 0% for mean survival time – which is extremely common.

3.1. MAE-Uncensored

The simplest solution is to exclude all censored subjects from the evaluation, then use Equation 3 to calculate the absolute error for each uncensored patient and take the average over the uncensored instances. The (marginal) distribution of censored and event subjects can vary substantially, making this strategy susceptible to bias. Moreover, when the censoring rate is high, a sizeable portion of the data will be completely ignored by the performance metric.

3.2. MAE-Hinge

Another way to incorporate censoring is to use the hinge loss – a one-sided metric that considers only if the predicted time is earlier than the censored time. For a censored subject, the MAE-hinge score is:

$$\mathcal{R}_{\text{MAE-hinge}}(\,\hat{t}_i,\, t_i,\, \delta_i = 0\,) \;=\; \max\{\, t_i - \hat{t}_i,\, 0\,\}\,.$$

The MAE-hinge is an optimistic evaluation of the true MAE for two reasons: (i) it assigns a score of 0 if the censoring happens before the predicted survival time; and (ii) it assigns a loss of $c_i - \hat{t}_i$ if the censoring time occurs after the prediction. Both are lower or equal to the true prediction error. Therefore, for a dataset with an extremely high censoring rate, a model can actually obtain an extremely low MAE-hinge by overestimating the event time for all subjects, as MAE-hinge will give zero scores for censored subjects, resulting in an optimistic overall score[3].

[1] Here, based on the true event time e_i, which of course is not given for censored instances.

[2] Alternatively, we could use the mean survival time for $\hat{t}_i$ from Equation 1.

[3] D-calibration can be used in conjunction with MAE-hinge to prevent this type of situation (Qi et al., 2022), as overestimating the event time will result in a skewed proportion for large probability intervals (which means not D-calibrated).

3.3. MAE-Margin

MAE-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. Given a subject censored at time t_i, we can calculate its margin time by:

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

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

Based on the censored time, the margin time can be more trustworthy for some circumstances than for others. For instance, we know effectively nothing about a patient censored at time 0 – hence we should have very little confidence that the margin time matches its actual event time. In contrast, assume that no patients in the training data have ever lived longer than 130 years old. If a patient was censored at 100 years old (*i.e.*, close to the longest known lifespan), we are quite certain that his/her margin time is close to the observed event time. Therefore, for each censored subject, Haider et al. (2020) suggested we use a confidence weight $\omega_i = 1 - S_{\text{KM}(\mathcal{D})}(t_i)$ for the error calculated based on the margin value. This weight ω_i yields lower confidence for early censoring subjects and higher confidence for late censoring data. Of course, we set the weights for uncensored subjects i to $w_i = 1$ as we have full confidence in those error calculations. The overall MAE-margin after a re-weighting scheme is:

$$\mathbb{E}_{i\sim\mathcal{D}}[\mathcal{R}_{\text{MAE-margin}}(\hat{t}_i, t_i, \delta_i)] = \frac{1}{\sum_{i=1}^{N}\omega_i}\sum_{i=1}^{N}\omega_i \left|[(1-\delta_i)\cdot e_{\text{margin}}(t_i) + \delta_i \cdot t_i] - \hat{t}_i\right|. \quad (4)$$

3.4. MAE-IPCW-D

Inverse Probability Censoring Weight (IPCW) was originally designed for handling censored subjects in the calculation of Brier score (BS) (Graf et al., 1999). The method uniformly transfers a censored subject's weights to subjects with known status at that time (Vock et al., 2016). In its simplest form, IPCW requires completely independent censoring; IPCW can also be extended to independent censoring conditional on covariates to even estimate survival models (Han et al., 2021).

Inspired by the simple form of IPCW, we can design an MAE-based evaluation method by uniformly transferring the weight of a censored subject to the uncensored subjects with later event times (Graf et al., 1999). Similar to how the prediction error of a censored subject can be approximated using the deterministic errors of subsequent uncensored subjects in the IPCW Brier Score, we can formulate this new MAE-based evaluation as:

$$\mathbb{E}_{i\sim\mathcal{D}}[\mathcal{R}_{\text{MAE-IPCW-D}}(\hat{t}_i, t_i, \delta_i)] = \frac{1}{N}\sum_{i=1}^{N}\frac{|t_i - \hat{t}_i|\cdot \mathbb{1}_{\delta_i=1}}{G(t_i)}, \quad (5)$$

where $G(t_i)$ is the probability of not being censored at the event time. We call this method "MAE-IPCW-D" (where D stands for difference) since the IPCW reweighing is essentially an approximation of the difference between the expected and observed times.

3.5. MAE-IPCW-T

One problem with MAE-IPCW-D is that Equation 5 considers only the predicted time for the uncensored subjects, but does not consider the predictions for the censored subjects. For example, imagine two models, M1 and M2, have identical predictions for every subject, except for one censored subject who is censored at time 100. M1 predicts this subject will die at 5, but M2 predicts it at 90. Notice M1 is off by at least 95 here, and M2 by at least 10; indeed, we know that M1's error for this patient is 85 worse than M2's. However, MAE-IPCW-D gives both models the same score.

To avoid this, the MAE-IPCW-T (where T stands for time) metric instead produces an estimated surrogate time of the event for each censored subject, as the average over the times of all subsequent uncensored subjects:

$$e_{\text{IPCW}}(t_i, \mathcal{D}) = \frac{\sum_{j=1}^{N}\mathbb{1}_{t_i<t_j}\cdot\mathbb{1}_{\delta_j=1}\cdot t_j}{\sum_{j=1}^{N}\mathbb{1}_{t_i<t_j}\cdot\mathbb{1}_{\delta_j=1}}. \quad (6)$$

After calculating this IPCW-weighted time using Equation 6, the MAE-IPCW-T method then uses Equation 4 (but with e_{IPCW} rather than e_{margin}) to compute MAE scores.

Importantly, the IPCW-T approach has downsides as well. IPCW weighted time is incapable of approximating the value for censored subjects with no subsequent event times. The same problem applies to IPCW-D and IPCW Brier score as well (Graf et al., 1999), where the denominator $G(t_i)$ in Equation 5 will equal to zero for those censored-at-last subjects. These subjects must be excluded from the evaluation. This is consistent with Administrative Brier Score (Kvamme & Borgan, 2019), in which individuals are removed from evaluation after their administrative censoring time.

3.6. MAE-PO

Both MAE-margin and MAE-IPCW-T use surrogate event values for the censored subjects; of course, this is only useful if those surrogate values are accurate.

Another way to estimate surrogate event values is using the pseudo-observations (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 pseudo-observation for a censored subject is defined as:

$$e_{\text{pseudo-obs}}(t_i, \mathcal{D}) = N \times \hat{\theta} - (N-1) \times \hat{\theta}^{-i}\,, \qquad (7)$$

where $\hat{\theta}^{-i}$ is the estimator applied to the $N-1$ element dataset formed by removing that i-th instance. The pseudo-observation 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. After calculating the pseudo-observation values using Equation 7, MAE-pseudo-observation (MAE-PO) then uses the re-weighting scheme in Equation 4 to produce the overall score.

Pseudo-observation values can be treated as though they are i.i.d. (Appendix D.1). Graw et al. (2009) has shown that, as $N \to \infty$, the pseudo-observation 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. However, when comparing the empirical performance of MAE-PO, we find it also works well in the situation where censoring is dependent on the covariates, which is consistent with Binder et al. (2014).

To be a meaningful estimate, we need to consider that these pseudo-observation values have some important properties – in particular, the pseudo-observation value for a censored subject is always greater than the censored time (see Theorem D.3). Appendix D provides more details about MAE-PO and its properties.

4. Experiments and Results

We conducted extensive experiments[4] to evaluate the effectiveness of the proposed evaluation methods – comparing the effectiveness of these 6 evaluation metrics for estimating the actual MAE of various survival models on a wide range of survival datasets. The two primary research objectives we care about are:

- Can the metric accurately **rank** the performance of models, *i.e.*, can it identify the better-performing models?
- Does the performance score generated by the MAE variants closely **approximate** the true MAE?

[4]Code to replicate all experiments can be found at https://github.com/shi-ang/CensoredMAE

These two questions encompass the discrimination and calibration aspects of the metric performance, respectively. Since it is not possible to obtain a true MAE evaluation for any existing survival dataset, we will construct semi-synthetic datasets using real datasets with synthetic censorship. Of course, the algorithms for learning the survival models will only see the censored time for those synthetic-censored subjects; we will only use their true event time when we calculate the true MAE.

4.1. Semi-Synthetic Datasets

To evaluate the MAE-inspired evaluation metrics, we need to know the true MAE, which means explicitly knowing when each subject will experience the event. As this information is not available in a real-world survival dataset, we need to produce a synthetic one. While it is easy to make up arbitrary covariates $\boldsymbol{x}_i$, and arbitrary event time e_i and censor time c_i, to be useful, we instead want synthetic data to be realistic, and matching the covariates, event distribution, and censoring distribution of some real-world dataset.

This motivated our approach for generating the semi-synthetic datasets (see also the flowchart in Figure 2):

1. Start with a real-world survival dataset $\mathcal{D}$;
2. Calculate some useful statistics and the censoring distributions for generating the censor times;
3. Produce $\mathcal{D}'$ by removing the censored instances from $\mathcal{D}$, leaving just the uncensored ones;
4. Form $\mathcal{D}''$ by applying some reasonable but synthetic censoring types to $\mathcal{D}'$. The censoring types are based on the statistics calculated in step 2.

We apply synthetic censoring to $\mathcal{D}'$ based on the independent censorship assumption, by computing a synthetic censoring time $\tilde{c}_i$ for each subject, and then censoring that subject if the synthetic censoring time is earlier than the event time ($\tilde{c}_i < t_i$), and otherwise leaving it uncensored. We consider six different kinds of censoring distributions for generating the synthetic censor times:

- Uniform distribution, $\tilde{c}_i \sim U[0, t_{\text{max}}]$ where t_{max} represents the maximum event time in $\mathcal{D}'$.
- Uniform distribution, augmented with administrative censoring at the median event time, formally: $\tilde{c}_i = \min\{c_i', t_{\text{median}}\}$, where $c_i' \sim U[0, t_{\text{max}}]$, and t_{median} is the median time in $\mathcal{D}'$.
- Exponential distribution, $\tilde{c}_i \sim \text{Exp}(\sigma_t)$, where σ_t is the standard deviation of the event times in $\mathcal{D}'$.
- Original censoring distribution *independent of the features*, $\tilde{c}_i \sim G_{\text{KM}(\tilde{\mathcal{D}})}(t)$, where $G_{\text{KM}(\tilde{\mathcal{D}})}$ is the censoring

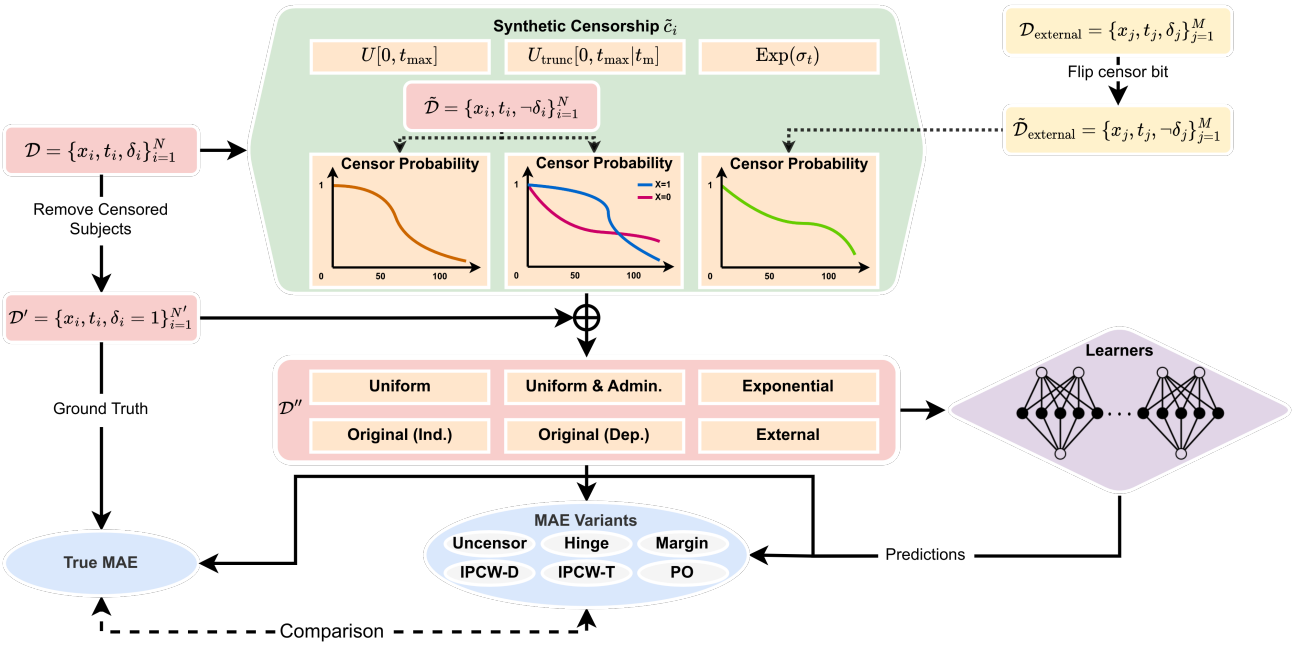

Figure 2. Flowchart illustrating the generation of realistic semi-synthetic survival datasets and the subsequent evaluation of MAE metrics.

distribution estimated by the KM algorithm, with the censor-bit-flipped datasets ($\tilde{D}$ in Figure 2).

- Original censoring distribution *dependent on the features*, $\tilde{c}_i \sim G_{\text{CoxPH}(\tilde{\mathcal{D}})}(t \mid \boldsymbol{x}_i)$, where $G_{\text{CoxPH}(\tilde{\mathcal{D}})}$ is the feature-dependent censoring distribution estimated by a Cox Proportional Hazard model (CoxPH) (Cox, 1972) with Breslow estimator (Breslow, 1975).

- Censoring distribution from an external (GBM) dataset, $\tilde{c}_i \sim G_{\text{KM}(\tilde{\mathcal{D}}_{\text{external}})}(t)$, due to its large percentage of early censoring. We rescaled the sampled censoring time, so the range matches the event times in $\mathcal{D}'$.

While this is not perfect, at least we know that these resulting semi-synthetic datasets, $\mathcal{D}''$, will have many properties of a real-world survival dataset: the real-world covariates domain, close-to-reality event distribution, and (a version of) close-to-reality censoring distribution[5].

We apply this synthetic censoring to 5 real-world datasets: GBM, SUPPORT, METABRIC, MIMIC-IV (Johnson et al., 2022) all-cause mortality datasets (MIMIC-A) and MIMIC-IV hospital mortality datasets (MIMIC-H). Table 1 summarizes the characteristics of these five datasets, and Appendix E.1 contains information on the dataset preprocessing and MIMIC-IV datasets construction.

[5]We are aware these steps do introduce some bias to the data, but this seems unavoidable.

4.2. Models

We compare the time-to-event prediction results using 10 survival prediction models: a naive linear regressor, KM (Kaplan & Meier, 1958), CoxPH (using an extension to produce an ISD) (Cox, 1972), Accelerate Failure time (AFT) (Stute, 1993) with the Weibull distribution, Gradient Boosting Machine with component-wise least squares (GBM-C) (Hothorn et al., 2006), Random Survival Forest (RSF) (Ishwaran et al., 2008), Multi-Task Logistic Regression (MTLR) (Yu et al., 2011; Jin, 2015), DeepHit (Lee et al., 2018), Survival Cluster Analysis (SCA) (Chapfuwa et al., 2020), and Survival Mixture Density Network (S-MDN) (Han et al., 2022). Appendix E.2 describes these 10 models, including the implementation details and hyper-parameter settings – and how the non-survival prediction models dealt with censoring training instances.

All models are trained on the semi-synthetic datasets, $\mathcal{D}''$. We split the data into a training set (80%) and a test set (20%) using a stratified 5-fold cross-validation (5CV) procedure (stratified wrt both time t and censor indicator δ). If the model requires a validation set for hyper-parameter tuning or early stopping, we will split 20% of the training set as the validation set. We then compute the mean on all the evaluation metrics across the 5CV folds.

4.3. Evaluation Metrics

We will use all six MAE metrics described in Section 3 to measure the prediction error between the synthetic censoring

Table 1. Summary of five datasets used in the empirical comparison.

Dataset	%Censored	#Instances	#Event	Max Event $t_{\max}$	#Features[†]
GBM	17.65%	595	490	3,881	8 (10)
SUPPORT	31.89%	9,105	6,201	1,944	26 (31)
METABRIC	42.07%	1,904	1,103	355	9 (9)
MIMIC-IV (all-cause mortality)	66.65%	38,520	12,845	4,404	91 (91)
MIMIC-IV (hospital mortality)	97.69%	293,907	6,780	248	10 (10)

[†] The number of features before performing one-hot encoding, with brackets (the number of features after one-hot encoding).

times and estimated times. We will also compute the true MAE score as the ground truth, using the hidden-to-learner true event times.

4.4. Experimental Results

We have 5 clinical datasets with 6 types of synthetic censoring, therefore there are in total $5 \times 6 - 1$[6] semi-synthetic datasets in our experiments. Due to the space limit, the main text only reports the results on three datasets (GBM, METABRIC, and MIMIC-A) and three censoring distributions (uniform with administrative censoring, feature-independent original censorship, and feature-dependent original censorship); see Figure 3. Appendix F presents the results of all semi-synthetic datasets – all combinations of datasets and censoring distributions.

Each subplot in Figure 3 (and also the figures in Appendix F) corresponds to a semi-synthetic dataset. Within each subplot, the first column is the ranking performance evaluation using the 5CV mean of true MAE (ground truth). Survival prediction models with larger circles had better true MAE. The gray bar plots on columns other than the first show how close that MAE-inspired metric is to the true MAE. If the error is large, we may still want to know which of the metrics is best at identifying which of the survival prediction models is best, by having a ranking of the models that agrees with the true ranking (for instance by identifying the three best models; see the red rounded-box at the top).

In the top left plot (GBM dataset with uniform and administrative censorship), we see that the GBM-C model was the most accurate, as it is represented by the largest blue solid circle, followed by DeepHit (second largest orange solid circle), then MTLR (solid green circle), then the 7 models with smaller open circles. They appear, in descending order, in the far left column labeled "True MAE". The other columns show how well the various variants of MAE do, in terms of approximating this true MAE.

The red "rounded box" at the top of the plot shows each metric's preference over the models. Here, a concordant preference to true MAE's indicates a great performance. We can see that PO (pseudo-observation) top three choices were GBM-C, DeepHit, and MTLR – which exactly matched the truth (first column by true MAE). By contrast, Margin had DeepHit first (not second), GBM-C second (not first), then MTLR in position 5 (not third). Other models did even worse at matching the order. Now consider the gray vertical bars, which show how close that MAE-inspired metric was to the true MAE, over these 10 different learning models. Here, smaller values mean that measure did well. We see that the far right PO was the smallest and with fairly tight error bars. Margin was second best, then Uncensored, IPCW-D, and IPCW-T, essentially tied, though the latter showed smaller variation, and with Hinge coming in last.

To demonstrate the effectiveness of the surrogate time method in MAE-margin, MAE-IPCW-T, and MAE-PO values, we also perform an ablation study that used the mean survival time of the KM estimation on the whole group as the proxy event time; this is the MAE population pseudo-observation (MAE-Pop-PO), see Appendix F.

4.4.1. Uniform with Administrative Censorship

The three subplots in the first row in Figure 3 demonstrate the metrics performance for uniform censoring distributions with administrative censoring. The MAE-PO is the best here, in both ranking performance (as it is the only one that correctly identifies the top-three models in GBM and METABRIC, and the only one that identifies two of the top-three models for MIMIC-A) and closeness to the true MAE (significantly better in GBM and METABRIC with p-value < 0.05 via t-test, and one of the best in MIMIC-A).

MAE-margin is the runner-up as it can identify parts of the best-performing models and has the second closest difference to true MAE. Between IPCW-D and IPCW-T, the performance does not have a significant difference in both ranking and proximity to true MAE. However, IPCW-D is associated with quite large error bars, which may be because the accuracy of later uncensored subjects will dominate the score (as we discussed in Section 3.4).

[6]The GBM dataset with external (GBM dataset) censoring will just be the same as feature-independent original censorship.

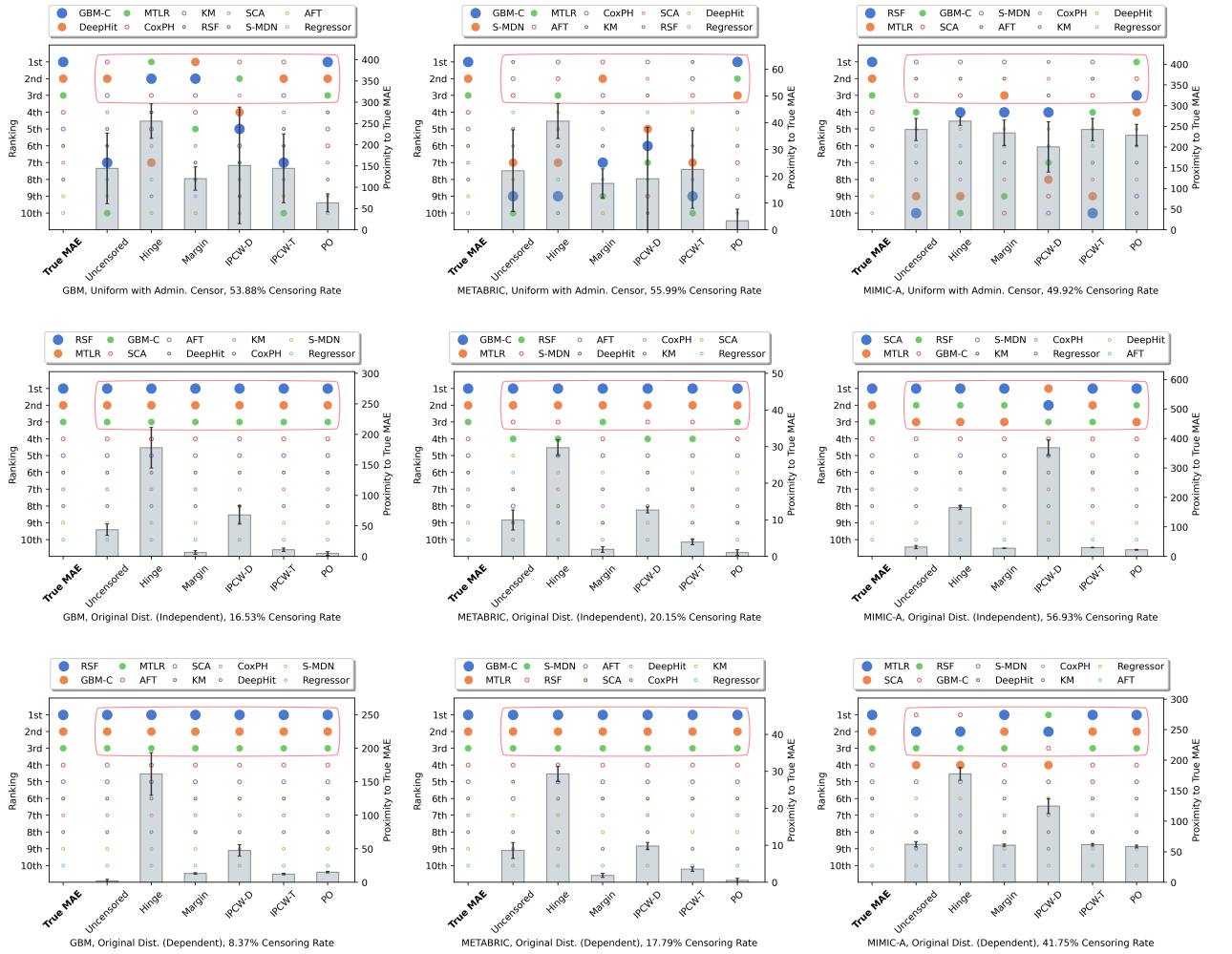

Figure 3. Evaluation metrics comparison in terms of ranking accuracy (left axis) and proximity to true MAE (right axis). Each row refers to a specific censoring type (uniform with administrative censoring, feature-independent original censoring, and feature-dependent original censoring), and each column to a specific dataset (GBM, METABRIC, MIMIC-A).

4.4.2. Feature-Independent Original Censorship

The three subplots in the second row in Figure 3 demonstrate the metrics performance for feature-independent original censoring distribution. Among all the evaluation metrics, margin and pseudo-observation perform equally the best for identifying the top-three performing models. MAE-pseudo-observation has a slight advantage in the proximity to true MAE, as its value is closer to the true MAE on GBM and significantly closer (p-value < 0.05) on METABRIC and MIMIC-A datasets. In addition, we also observed that IPCW-D is always associated with a large variance (reason explained above).

4.4.3. Feature-Dependent Original Censorship

The three subplots in the third row in Figure 3 demonstrate the metrics performance for feature-independent original censoring distribution. MAE-uncensored performs the best on GBM datasets, which may be due to the low synthetic censoring rate of this dataset (8.37% censoring rate), meaning the whole dataset could be approximately represented by the uncensored population. Among the MAE metrics that can handle the censored subjects, MAE-margin, IPCW-T, and MAE-PO perform equally well on GBM and MIMIC-A datasets. For the METABRIC dataset, pseudo-observation is the best metric as it has the significantly lowest error to true MAE among all the metrics that can identify the top-three performing models.

Table 2. Summarization of MAE-based metric performance by counting the number of times each metric is best. *Note this includes ties.

	Uni.	Uni.&Admin.	Exp.	Orig.(Ind.)	Orig.(Dep.)	GBM	Total
Uncensor	0	0	0	0	1	0	1
Hinge	0	0	0	0	0	0	0
Margin	2*	0	3	2*	1	0	8*
IPCW-D	0	0	0	0	0	0	0
IPCW-T	0	0	0	0	0	0	0
PO	4*	5	2	4*	3	4	22*

4.4.4. Other Types of Censoring Distributions

Figures 8, 10, and 13 in Appendix F demonstrate the performance metrics for uniform censoring, exponential censoring, and GBM censoring distribution. In all 14 cases (5 for uniform, 5 for exponential, and 4 for GBM censoring), MAE-PO is the best 10 times (71%) while MAE-margin is the best for the other four. The MAE-margin metric prevails in 3 out of 5 datasets with exponential censoring, indicating that its performance is either superior or comparable to MAE-PO for this specific censorship type.

Table 2 presents a comprehensive summary of the performance of all MAE-based metrics across 29 semi-synthetic datasets. Each column in the table corresponds to a specific censorship type, as outlined in Section 4.1. The values within the table indicate the number of times that each metric outperforms the others for a given censorship distribution. The final columns provide an overview of the cumulative results from all experiments. Notably, the MAE-PO metric demonstrates its robustness by emerging as the superior choice in 22 out of 29 experiments, accounting for a 76% success rate. MAE-margin is best for most of the other ones, especially when the censoring is exponentially distributed. As a result, we recommend that researchers consider using MAE-margin for datasets that seem to exhibit an exponential censoring distribution, and using MAE-PO for datasets with other types of censoring.

5. Discussion and Conclusion

Here, we first argue that MAE should be used as an evaluation metric for evaluating survival prediction models (*e.g.*, ISDs), especially for the standard such task, which requires predicting time-to-event. Appendix C shows that MAE is a *proper* scoring rule (for an uncensored dataset), and that no other metric quantifies the time-to-event accuracy like MAE does. To handle the right censorship, we introduced three novel MAE variations: MAE-IPCW-D, MAE-IPCW-T, and MAE-PO, and empirically compared them to the three existing variants: MAE-uncensored, MAE-hinge, and MAE-margin, over several realistic semi-synthetic datasets. These empirical results demonstrate that MAE-PO (i) can often correctly identify the top-performing models and (ii) often has error closest to true MAE.

We recognize that MAE-PO method has certain limitations, particularly subject to the limitation of KM, *e.g.*, (i) not accounting for the effects of covariates, and (ii) the requirement for the independent censoring assumption to be valid.

This research focuses on only the MAE score. There are, however, many other ways to measure the errors between the predicted and observed time. For instance, one can use the Mean Squared Error (MSE) to penalize large prediction errors, relative MAE or (relative MSE) score to compare models where errors are measured in different scales, or normalized versions to set an upper constraint on the MAE or MSE score. Importantly, we can easily apply all of the techniques mentioned above to handle censored subjects to these variations. Note also that we can use the realistic semi-synthetic datasets defined above along with the proposed methodology to evaluate other evaluation metrics.

Following the famous remark "All models are wrong, but some are useful" (Box, 1976), it is important to precisely define "usefulness". For many clinical tasks, where this corresponds to time-to-event, it is important to have measures that can determine if a model can accurately estimate event time values. This paper has provided such a measure (MAE-PO) for survival prediction, and also demonstrated that it works effectively. Note this also required finding ways to produce realistic semi-synthetic datasets. We anticipate others will be able to use this methodology for evaluating evaluation measures, and more importantly, for the result of this analysis, suggesting that MAE-PO often approximates the true MAE.

Acknowledgements

This research received support from the Natural Science and Engineering Research Council of Canada (NSERC), the Alberta Machine Intelligence Institute (Amii), NIH/NIDDK R01-DK123062, NIH/NINDS R61-NS120246, NIH/NHLBI Award R01HL148248, NSF Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science, and NSF CAREER Award 2145542. The authors extend their gratitude to the anonymous reviewers for their insightful feedback and valuable suggestions.

References

Aalen, O. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pp. 701–726, 1978.

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.

Antolini, L., Boracchi, P., and Biganzoli, E. A time-dependent discrimination index for survival data. *Statistics in medicine*, 24(24):3927–3944, 2005.

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.

Box, G. E. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

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

Buja, A., Stuetzle, W., and Shen, Y. Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft, November*, 3:13, 2005.

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, 2020.

Costantino, J. P., Gail, M. H., Pee, D., Anderson, S., Redmond, C. K., Benichou, J., and Wieand, H. S. Validation studies for models projecting the risk of invasive and total breast cancer incidence. *Journal of the National Cancer Institute*, 91(18):1541–1548, 1999.

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

Cox, D. R. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

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.

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.

Fleming, T. R. and Harrington, D. P. *Counting processes and survival analysis*. John Wiley & Sons, 2011.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

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.

Han, X., Goldstein, M., Puli, A., Wies, T., Perotte, A., and Ranganath, R. Inverse-weighted survival games. *Advances in Neural Information Processing Systems*, 34: 2160–2172, 2021.

Han, X., Goldstein, M., and Ranganath, R. Survival mixture density networks. *ArXiv*, abs/2208.10759, 2022.

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.

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

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S., et al. Random survival forests. *Annals of Applied Statistics*, 2(3):841–860, 2008.

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

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L., Anthony, and Mark, R. MIMIC-IV (version 2.0). *PhysioNet (2022). https://doi.org/10.13026/7vcr-e114.*, 2022.

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

Klein, J. P. and Moeschberger, M. L. *Survival analysis: techniques for censored and truncated data*, volume 1230. Springer, 2003.

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.

Kumar, N., Qi, S.-a., Kuan, L.-H., Sun, W., Zhang, J., and Greiner, R. Learning accurate personalized survival models for predicting hospital discharge and mortality of covid-19 patients. *Scientific reports*, 12(1):1–11, 2022.

Kvamme, H. and Borgan, Ø. The brier score under administrative censoring: Problems and solutions. *arXiv preprint arXiv:1912.08581*, 2019.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

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.

Miscouridou, X., Perotte, A., Elhadad, N., and Ranganath, R. Deep survival analysis: Nonparametrics and missingness. In *Machine Learning for Healthcare Conference*, pp. 244–256. PMLR, 2018.

Nelson, W. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972.

Qi, S.-a., Kumar, N., Xu, J.-Y., Patel, J., Damaraju, S., Shen-Tu, G., and Greiner, R. Personalized breast cancer onset prediction from lifestyle and health history information. *Plos one*, 17(12):e0279174, 2022.

Rindt, D., Hu, R., Steinsaltz, D., and Sejdinovic, D. Survival regression with proper scoring rules and monotonic neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 1190–1205. PMLR, 2022.

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

Vock, D. M., Wolfson, J., Bandyopadhyay, S., Adomavicius, G., Johnson, P. E., Vazquez-Benitez, G., and O'Connor, P. J. Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. *Journal of biomedical informatics*, 61:119–131, 2016.

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.

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. Overview of the Appendix

Appendix B provides a summary of the notation and assumptions used throughout the study. Appendix C compares the MAE metric with five other commonly used evaluation metrics. Appendix D theoretically describes the property of pseudo-observation and proves its authenticity. Appendix E describes the implementation details, including the datasets and the ten models used to estimate the timing of events. Appendix F includes the complete results for MAE-inspired evaluation metrics comparison on the 29 semi-synthetic datasets.

B. Notation

In general, a survival dataset contains N time-to-event tuples, $\mathcal{D} = \{(\boldsymbol{x}_i, t_i, \delta_i)\}_{i=1}^{N}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ represents the observed d-dimensional features for the i-th instance, $t_i \in \mathbb{R}_+$ denotes the event or censor time, and $\delta_i \in \{0, 1\}$ is a censor/event indicator where $\delta_i = 0$ means the subject is right-censored (the subject has not experienced an event) and $\delta_i = 1$ means observed event times. We assume each patient has a event time e_i, and a censoring time c_i, and assign $t_i \leftarrow \min\{e_i, c_i\}$ and $\delta_i \leftarrow \mathbb{1}[e_i \leq c_i]$.

Assumption *(Independent censoring)* In this study, we follow the standard convention that event time and censor time are assumed independent and conditional on the covariates. Formally, for event time $e_i \sim E$, censoring time $c_i \sim C$ and covariates $\boldsymbol{x}_i \sim \mathbf{X}$, we assume $E \perp C \mid \mathbf{X}$.

Random censoring is another commonly used assumption with the definition $T \perp C$ without conditioning on $\mathbf{X}$. However, since random censoring implies independent censoring, all the nature and properties proved under the independent assumption will also hold for the random assumption.

The predicted ISD curves can also serve as a surrogate for the risk scores as well. For instance, the time-independent risk scores can be defined as the negative value of the predicted survival times of ISDs. Alternatively, we can define the time-dependent risk scores, at time $t > 0$, as the negative of the survival probability, $-S(\, t \mid \boldsymbol{x}_i \,)$.

We include a notation table, Table 3, to summarize the symbols and abbreviations we used in the paper.

C. Comparing MAE with Other Evaluation Metrics

Survival prediction is like regression as it predicts a real number (the subject's event time) from a description of a patient $\boldsymbol{x}_i$. Given this commonality, we want to evaluate survival prediction models using measures for evaluating regression tasks, such as mean absolute error (MAE) or mean squared error (MSE).

MAE measures the mean of the absolute difference between the true event time and the predicted time. As suggested in Figure 1 (a) the MAE score for an uncensored subject is (reformulated Equation 3)

$$\mathcal{R}_{\text{MAE}}(\, S_{\text{m}}(\, \cdot \mid \boldsymbol{x}_i \,),\, t_i,\, \delta_i = 1 \,) \;=\; |\, t_i - \hat{t}_i \,|, \tag{8}$$

where $\hat{t}_i$ is the median survival time from Equation 2[7]. MAE is a negative scoring rule which means the smaller the loss, the better the model performs.

Theorem C.1. *The MAE score for uncensored subjects, $\mathcal{R}_{MAE}(\, S_m(\, t \mid \boldsymbol{x}_i \,),\, t_i,\, \delta_i = 1 \,)$, is a proper scoring rule if we use median survival time of the predicted ISD as the predicted time.*

Proof. By the proper scoring rule definition proposed by (Gneiting & Raftery, 2007) and (Rindt et al., 2022), a negative scoring rule (a lower score indicates better performance) is proper if for any model m we have

$$\mathbb{E}_{i\sim\mathcal{D}}\, \mathcal{R}(\, S_{\text{true}}(\, t \mid \boldsymbol{x}_i \,),\, t_i,\, \delta_i \,) \;\leq\; \mathbb{E}_{i\sim\mathcal{D}}\, \mathcal{R}(\, S_{\text{m}}(\, t \mid \boldsymbol{x}_i \,),\, t_i,\, \delta_i \,)\,,$$

where $S_{\text{true}}(\, t \mid \boldsymbol{x}_i \,)$ is the true ISD distribution for each subject. Similarly, a positive scoring rule will change the above inequality from $\leq$ to $\geq$.

For an uncensored dataset, every subject has the observed time equal to the event time ($t_i = e_i$). We need to prove that, for any $e_i \sim S_{\text{true}}(\, \cdot \mid \boldsymbol{x}_i \,)$, for any $\boldsymbol{x}_i \sim \mathbf{X}$, using the median survival time of the true distribution will minimize the MAE. We

[7] Alternatively, we could use the mean survival time for $\hat{t}_i$ from Equation 1.

Table 3. Table of notation. Ordered alphabetically.

Symbol/Abbreviation	Definition
c_i	Censor time of subject i
$\tilde{c}_i$	Synthetic censor time of subject i
$\mathcal{D}$	Raw Dataset
$\mathcal{D}'$	Raw Dataset with only uncensored subjects
$\mathcal{D}''$	Semi-synthetic Dataset with synthetic censoring on $\mathcal{D}'$
$\tilde{\mathcal{D}}$	Raw Dataset with flipped censor bit
$\text{Exp}(\lambda)$	Exponential distribution with λ as the parameter
$\mathbb{E}[\cdot]$	Expectation
e_i	Event time of subject i
$F(t \mid \boldsymbol{x}_i)$	Cumulative density function given the covariates $\boldsymbol{x}_i$
$f(t \mid \boldsymbol{x}_i)$	Probability density function given the covariates $\boldsymbol{x}_i$
$G(t)$	Censor distribution
$G_{\text{KM}}(t)$	Feature-independent censor distribution, estimated using KM model
$G_{\text{CoxPH}}(t \mid \boldsymbol{x}_i)$	Feature-dependent censor distribution, estimated using CoxPH model
N	The size of the dataset, number of subjects
$\mathcal{R}$	Scoring rule
$S_m(t \mid \boldsymbol{x}_i)$	Survival distribution given the covariates, estimated by model m
$S_{\text{KM}(\mathcal{D})}(\,t\,)$	Group-level survival distribution, estimated using KM model on the dataset $\mathcal{D}$
t_i	Observed time of subject i, $t_i = e_i$ if $\delta_i = 1$ and $t_i = c_i$ if $\delta_i = 0$
$\hat{t}_i$	Predicted event time for subject i
$U[a, b]$	Uniform distribution starting at a and ending at b
$\boldsymbol{x}_i$	Covariates of subject i
δ_i	Censor bit / censor indicator, $\delta_i = \mathbb{1}_{e_i > c_i}$
$\hat{\theta}$	Unbiased estimator for the event time
ω_i	Confidence weights for subject i
1-calibration	Hosmer-Lemeshow Calibration
BS	Brier Score
C-index	Concordance Index
CV	Cross-Validation
D-calibration	Distribution Calibration
IBS	Integrated Brier Score
IPCW	Inverse Probability Censoring Weight
IPCW-D	Inverse Probability Censoring Weight on Difference
IPCW-T	Inverse Probability Censoring Weight on Time
ISD	Individual Survival Distribution
KM	Kaplan-Meier estimator
LL	Log-Likelihood
MAE	Mean Absolute Error
PDF	Probability Density Function
PO	Pseudo-Observation

can formulate the MAE score by:

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_i, t_i \sim \mathcal{D}}\left[\mathcal{R}_{\text{MAE}}(\, S_{\text{m}}(\,\cdot \mid \boldsymbol{x}_i\,),\, t_i,\, \delta_i = 1\,) \mid \delta_i = 1\right] \\
&\qquad = \mathbb{E}_{\boldsymbol{x}_i \sim \mathbf{X}, e_i \sim S_{\text{true}}(\, e_i \mid \boldsymbol{x}_i\,)}\left[\left|e_i - \hat{t}_i\right|\right] \\
&\qquad = \int_{\boldsymbol{x}_i \sim \mathbf{X}} p(\boldsymbol{x}_i) \int_0^{\infty} f_{\text{true}}(e_i \mid \boldsymbol{x}_i)\left|e_i - \hat{t}_i\right| \, de_i \, d\boldsymbol{x}_i \\
&\qquad = \int_{\boldsymbol{x}_i \sim \mathbf{X}} p(\boldsymbol{x}_i) \left(\int_0^{\hat{t}_i} f_{\text{true}}(e_i \mid \boldsymbol{x}_i)(\hat{t}_i - e_i)\, de_i + \int_{\hat{t}_i}^{\infty} f_{\text{true}}(e_i \mid \boldsymbol{x}_i)(e_i - \hat{t}_i)\, de_i \right) d\boldsymbol{x}_i \,,
\end{aligned}
$$

where $f_{\text{true}}(e_i \mid \boldsymbol{x}_i)$ represents the true PDF function, and $p(\boldsymbol{x}_i)$ represents the marginal over covariates. Computing the derivative of the above equation and setting it to zero will allow us to identify the minimum of this function with respect to $\hat{t}_i$. By taking the derivative:

$$\begin{aligned}
&\frac{d}{d\hat{t}_i}\int_{\boldsymbol{x}_i} p(\boldsymbol{x}_i)\left(\int_0^{\hat{t}_i} f_{\text{true}}(e_i \mid \boldsymbol{x}_i)(\hat{t}_i - e_i)\, de_i + \int_{\hat{t}_i}^{\infty} f_{\text{true}}(e_i \mid \boldsymbol{x}_i)(e_i - \hat{t}_i)\, de_i\right) d\boldsymbol{x}_i \\
&= \int_{\boldsymbol{x}_i} p(\boldsymbol{x}_i)\left(f_{\text{true}}(\hat{t}_i \mid \boldsymbol{x}_i)(\hat{t}_i - \hat{t}_i) + \int_0^{\hat{t}_i} f_{\text{true}}(e_i \mid \boldsymbol{x}_i)\, de_i - f_{\text{true}}(\hat{t}_i \mid \boldsymbol{x}_i)(\hat{t}_i - \hat{t}_i) - \int_{\hat{t}_i}^{\infty} f_{\text{true}}(e_i \mid \boldsymbol{x}_i)\, de_i\right) d\boldsymbol{x}_i \\
&= \int_{\boldsymbol{x}_i} p(\boldsymbol{x}_i)\left(\int_0^{\hat{t}_i} f_{\text{true}}(e_i \mid \boldsymbol{x}_i)\, de_i - \int_{\hat{t}_i}^{\infty} f_{\text{true}}(e_i \mid \boldsymbol{x}_i)\, de_i\right) d\boldsymbol{x}_i\ ,
\end{aligned}$$

where the first equality holds by applying the Leibniz integral rule. Setting the above derivation to zero leads to $\hat{t}_i$ such that

$$\int_0^{\hat{t}_i} f_{\text{true}}(e_i \mid \boldsymbol{x}_i)\, de_i = \int_{\hat{t}_i}^{\infty} f_{\text{true}}(e_i \mid \boldsymbol{x}_i)\, de_i \implies 1 - S_{\text{true}}(\,\hat{t}_i \mid \boldsymbol{x}_i\,) = S_{\text{true}}(\,\hat{t}_i \mid \boldsymbol{x}_i\,)\ .$$

Therefore, the MAE is minimized if $S_{\text{true}}(\,\hat{t}_i \mid \boldsymbol{x}_i\,) = \frac{1}{2}$, that is, when $\hat{t}_i$ is the median time of the true ISD distribution. This completes the proof. □

The preceding derivation uses the median survival time and MAE to demonstrate the properness of this combination. We can prove that mean survival time with MSE is also proper following the same line of reasoning.

Theorem C.2. *The MSE score for uncensored subjects, $\mathcal{R}_{MSE}(\,S_m(\,t \mid \boldsymbol{x}_i\,),\, t_i,\, \delta_i = 1\,)$, is a proper scoring rule if we use mean survival time as the predicted time.*

Proof. Follow the logic in Theorem C.1, we can formulate the uncensored MSE score as:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}_i, t_i \sim \mathcal{D}}\ [\mathcal{R}_{\text{MSE}}(\,S_{\text{m}}(\,t \mid \boldsymbol{x}_i\,),\, t_i,\, \delta_i = 1\,) \mid \delta_i = 1] &= \mathbb{E}_{\boldsymbol{x}_i \sim \mathbf{X}, e_i \sim S_{\text{true}}(\,e_i \mid \boldsymbol{x}_i\,)}\left[(\hat{t}_i - e_i)^2\right] \\
&= \int_{\boldsymbol{x}_i \sim \mathbf{X}} p(\boldsymbol{x}_i) \int_0^{\infty} f_{\text{true}}(e_i \mid \boldsymbol{x}_i)(\hat{t}_i - e_i)^2\, de_i\, d\boldsymbol{x}_i\ .
\end{aligned}$$

We also compute the derivative of the above equation with respect to $\hat{t}_i$ and set the derivation to zero:

$$\begin{aligned}
\frac{d\,\mathbb{E}_{\boldsymbol{x}_i, e_i \sim \mathcal{D}}\ [\mathcal{R}_{\text{MSE}}(\,S_{\text{m}}(\,t \mid \boldsymbol{x}_i\,),\, e_i,\, \delta_i = 1\,) \mid \delta_i = 1]}{d\hat{t}_i} &= \frac{d}{d\hat{t}_i}\int_{\boldsymbol{x}_i} p(\boldsymbol{x}_i) \int_0^{\infty} f_{\text{true}}(e_i \mid \boldsymbol{x}_i)(\hat{t}_i - e_i)^2\, de_i\, d\boldsymbol{x}_i \\
&= \int_{\boldsymbol{x}_i} p(\boldsymbol{x}_i) \int_0^{\infty} 2 f_{\text{true}}(e_i \mid \boldsymbol{x}_i)(\hat{t}_i - e_i)\, de_i\, d\boldsymbol{x}_i\ .
\end{aligned}$$

Setting the above derivation to zero leads to $\hat{t}_i$ such that

$$\hat{t}_i \cdot \int_0^{\infty} f_{\text{true}}(e_i \mid \boldsymbol{x}_i)\, de_i = \int_0^{\infty} e_i \cdot f_{\text{true}}(e_i \mid \boldsymbol{x}_i)\, de_i \implies \hat{t}_i = \int_0^{\infty} S_{\text{true}}(\,e_i \mid \boldsymbol{x}_i\,)\, de_i\ .$$

Therefore, the MSE is minimized if $\hat{t}_i$ is the mean time of the true ISD distribution. Then the proof is complete. □

The main text compared the MAE to five other frequently used metrics for survival prediction (from both discriminative and calibration perspectives). It is necessary to have a thorough grasp of the limitations of all six metrics when selecting metrics for model optimization, and separately for model evaluation. We suggest that a task-oriented strategy is needed for selecting the right evaluation metrics for the application. However, the following section argues that MAE loss is the best option if the objective is to quantify the time-to-event accuracy or to make tailored clinical decisions.

Note that the MAE methods (and also C-index) require an ISD model to use a single value as the predicted time for an instance. In this context, the obvious candidates are either median or mean time (Equations 1 and 2) of the survival curves.

However, in practice, the predicted survival curves from many ISD models often terminate at a specific time (typically the largest observed time in the training dataset) with non-zero probabilities. This limitation poses a challenge as the subsequent ISD distribution is unknown/censored, rendering the accurate computation of mean or median time impossible. In this study, we adopt a linear extrapolation method to extend survival curves, following the approach outlined by Haider et al. (2020). The extrapolation of survival curves remains an active area of research, and we aim to explore the impact of various extrapolation techniques on MAE evaluation in future investigations.

C.1. Concordance Index

The C-index gauges the model's ability to rank the subject's risk. It is described as the proportion of all comparable pairs where the predictions and outcomes are concordant. A pair is comparable if we can determine who has the event first. The C-index is defined by Harrell Jr et al. (1996):

$$\text{C-index} = \frac{\sum_{i,j} \mathbb{1}_{t_i<t_j} \cdot \mathbb{1}_{\eta_i>\eta_j} \cdot \delta_i}{\sum_{i,j} \mathbb{1}_{t_i<t_j} \cdot \delta_i} ,$$

where η_i represents the risk score of subject i. As we discussed in Section 2, the risks can be either defined as the negative of expected time or of survival probability at some specified time. If we use the negative of predicted time, it is called time-independent C-index (C^{ti}). A visual illustration of C^{ti} using a discordant pair can be found in Figure 1 (b), which assesses if the order of true event times (stars) is concordant with the order of predicted event times (triangles).

As many researchers pointed out (Antolini et al., 2005), C^{ti} is known to be biased upwards if the amount of censoring in the data is high (also proved in Proposition C.3). Therefore, Antolini et al. (2005) proposes a time-dependent C-index (C^{td}) that claims to solve the issue. It is essentially a weighted average of the time-dependent area under the curve (AUC)[8] scores over time. Unfortunately, this does not solve the issue, as C^{td} is not a proper scoring rule, proved by Rindt et al. (2022). Note that C^{ti} and AUC scores are also not proper, using the same reasoning.

Proposition C.3. *Given a dataset with a censoring rate of $1-a$. To evaluate the concordance index, the ratio of comparable pairs to total pairs is bounded by:*

$$a^2 \;\le\; r \;\le\; 2a - a^2 , \tag{9}$$

where r is the comparable-to-total pairs ratio.

Proof. For a dataset with n instances, the total number of pairs is $C(n,2) = \frac{n(n-1)}{2}$ and the number of comparable pairs is varied based on the temporal position of censored and event instances. In the two extreme cases, the minimum number of comparable pairs happens when all censored instances happened before the earliest event instance (none of the censor-event pairs is comparable), while the maximum number of comparable pairs happens when all censored instances happened after the last event instance (arbitrary censor-event pair is comparable). Therefore, the number of comparable pairs is bounded by:

$$C(n \times a, 2) \le \text{number of comparable pairs} \le C(n \times a, 2) + na \times (n - na) .$$

Then the ratio of comparable pairs of total pairs (by dividing the above equation by the total number of pairs $\frac{n(n-1)}{2}$) is bounded by:

$$\frac{a(na-1)}{n-1} \;\le\; r \;\le\; \frac{a(na-1)}{n-1} + \frac{2na(1-a)}{n-1} .$$

As the dataset size grows ($n \to \infty$), the bound becomes:

$$\lim_{n\to\infty} \frac{a(na-1)}{n-1} \;\le\; r \le \lim_{n\to\infty} \frac{2na - a - na^2}{n-1} \quad\Longrightarrow\quad a^2 \le\; r \le 2a - a^2 ,$$

then the proof is complete. □

It would be tempting to assume that MAE and C-index favor the models monotonically, *i.e.*, if Model 1 has a lower MAE than Model 2, then Model 1 must have a higher C-index and vice versa. This is not always the case. Figure 4 shows three

[8] In datasets with binary outcomes and no censoring, the AUC and C-index are equivalent. In survival settings, C^{ti} can be considered as a generalization of the AUC, as it uses the observed times to construct pairs and predicted events to compare the concordance, whereas the AUC only uses binary statuses and predicted probabilities at a specific time point, respectively.

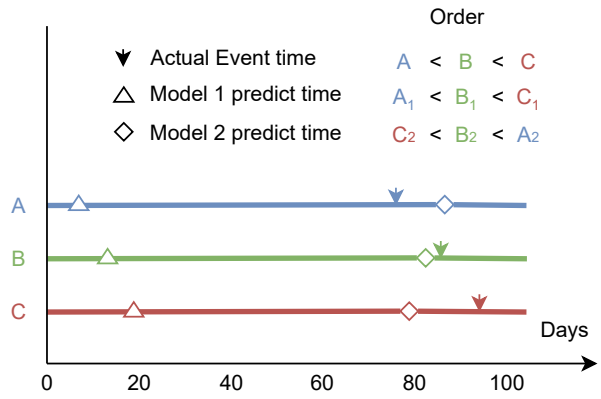

Figure 4. Comparison between MAE and Concordance Index. The arrows indicate the actual event times, and the triangles and diamonds indicate the predicted event times using Model 1 and Model 2, respectively.

patients who die at 77 days (A), 85 days (B), and 93 days (C). We can see from the model's predicted event times that Model 1 perfectly ranks the events (C-index = 1), but its predicted event times are far from the actual event times. While Model 2 predicts the order of the events in the wrong order (C-index = 0), it has a more accurate time estimation. This illustration shows that MAE preferences do not coincide with C-index preferences over the model. This is because MAE estimation focuses on measuring the discriminative accuracy at the individual level, while C-index measures the discriminative accuracy at the pairwise level.

C.2. Integrated Brier Score

The integrated Brier score (IBS) is the expectation of single-time Brier scores (BS) (Graf et al., 1999) over time. For each time t, BS calculates the squared difference between the model's predicted probability and the status (0 or 1) of an uncensored subject at that time. For subjects censored before time t, BS will use the inverse probability censoring weight (IPCW) technique to uniformly transfer their weights to subjects with known status at that time. IBS for survival prediction is typically defined as:

$$\mathcal{R}_{\text{IBS}}(S_{\text{m}}(\cdot \mid \boldsymbol{x}_i), t_i, \delta_i) = \frac{1}{t_{\max}} \cdot \int_0^{t_{\max}} \frac{S_{\text{m}}(t \mid \boldsymbol{x}_i)^2 \cdot \mathbb{1}_{t_i \le t, \delta_i = 1}}{G(t_i)} + \frac{(1 - S_{\text{m}}(t \mid \boldsymbol{x}_i))^2 \cdot \mathbb{1}_{t_i > t}}{G(t)} \, dt \,,$$

where $t_{\max}$ is normally the maximum event time of the combined training and validation datasets. $G(t)$ is the non-censoring probability at time t, which is typically estimated with KM, and its reciprocal is referred to as the IPCW weights (refers to Equation 5). The principle of IPCW for IBS is to evenly and repetitively distribute the weight of a censored subject to subjects who will experience the event after its censored time (Graf et al., 1999; Vock et al., 2016). Figure 1 (c) provides a graphic depiction of IBS for an uncensored subject. IBS score is represented by the weighted squared error of the shaded regions.

There are some variants within the IBS family. For instance, survival continuous ranked probability score (survival-CRPS) (Avati et al., 2020), also called integrated mean absolute error, simply omits the IPCW weights from the calculation, and integrated binomial log-likelihood uses the log-absolute error instead of squared error. Due to their close resemblance, this paper will solely discuss IBS, as the same reasoning applies to other variants.

IBS is a negative scoring rule which means smaller score implies better performance. It is also known to be a proper scoring rule (Buja et al., 2005) if the censoring distribution is estimated correctly (Kvamme & Borgan, 2019; Rindt et al., 2022). The relationship between IBS and MAE is subtle. Intuitively, IBS tries to minimize the summation of squared difference in the shaded areas ($\alpha + \beta$) as shown in Figure 1 (c), while MAE tries to minimize the absolute difference between the two areas ($|\alpha - \beta|$). Motivate by this, we can conclude that:

- minimizing IBS can lead to minimizing MAE;
- minimizing MAE does not necessarily lead to minimizing IBS.

One of the disadvantages of BS and IBS is that they are difficult to interpret, *i.e.*, they do not correspond to the expected time error nor to the ranking of the patient's time to event. It might be useful when clinicians need to make decisions concerning time-specific probabilities, *i.e.*, conservative treatment if a 5-year survival probability is greater than 80%.

A further issue with the IBS with IPCW weights is that it is dominated by the few uncensored subjects (especially those who experience events at a late period) if the censoring rate is high, which implies a high variance. In other words, if only accurate prediction for those uncensored subjects were made at a late time (and an imprecise prediction for others), the overall IBS score will still tend to be good, and *vice versa*. An extreme case happened when the dataset had some administrative censoring[9] subjects, the weights for those subjects cannot transfer to later subjects because they are the last observed time in the datasets; see also examples in (Kvamme & Borgan, 2019).

C.3. Log-likelihood

The log-likelihood (LL) score measures the logarithmic values of the PDF function at the moment of the event (see Figure 1 (d)). For censored patients, LL assesses the survival probability at the censoring time, and it is a positive scoring rule. Therefore, the greater the PDF intensity or survival probability, the higher the performance. Given the assumption of independent and noninformative censorship:

$$\mathcal{R}_{LL}(S_{\text{m}}(\,\cdot \mid \boldsymbol{x}_i\,), t_i, \delta_i) = \delta_i \log f_{\text{m}}(t_i \mid \boldsymbol{x}_i) + (1-\delta_i) \log S_{\text{m}}(\,t \mid \boldsymbol{x}_i\,)\,.$$

LL has been widely used as the loss function to train an ISD model (see for example Lee et al. (2018); Miscouridou et al. (2018)). Rindt et al. (2022) proved that LL is a proper scoring rule. The following subsections prove that maximizing the LL will lead to minimizing the MAE-hinge.

Despite all LL's benefits, we discourage its usage as an evaluation metric as it does not have a boundary – so it is not clear what value means we have a good model. Furthermore, considering the nature of the PDF function and probability mass function (PMF), we cannot use LL to compare (1) continuous and discrete models; (2) two discrete models with different bin boundaries; nor (3) models of the same type trained with different datasets/time ranges.

C.3.1. Minimize uncensored LL will lead to Minimize MAE-uncensored

The expectation of the predicted survival probability is presented in Equation 1 (here for simplicity, we omit the condition on $\boldsymbol{x}_i$). For uncensored data, with maximizing the likelihood at the event time $f(e_i) \to \infty$, the expectation becomes:

$$\begin{aligned}
\mu_i &= \lim_{f(e_i)\to\infty} \int_0^\infty S(t)\, dt = \lim_{f(e_i)\to\infty} \int_0^\infty t\, f(t)\, dt \\
&= \lim_{f(e_i)\to\infty} \left(\int_0^{e_i-\Delta t} t\, f(t)\, dt + \int_{e_i-\Delta t}^{e_i+\Delta t} t\, f(t)\, dt + \int_{e_i+\Delta t}^{\infty} t\, f(t)\, dt \right) \\
&= e_i + \lim_{f(e_i)\to\infty} \left(\int_0^{e_i-\Delta t} t\, f(t)\, dt + \int_{e_i+\Delta t}^{\infty} t\, f(t)\, dt \right)\,,
\end{aligned}$$

where the first term is the event time e_i when $\Delta t \to \infty$, and the second term is close to zero ($f(t) \to 0$ for $t \neq e_i$). So the L1-loss $|\mu_i - e_i|$ will be close to 0.

C.3.2. Minimize censored LL will lead to Minimize MAE-hinge

For censored data, the MLE would be $S(c_i) \to 1$ (c_i is the censored time), the expectation becomes:

$$\begin{aligned}
\mu_i &= \lim_{f(e_i)\to\infty} \int_0^\infty S(t)\, dt = \lim_{f(e_i)\to\infty} \left(\int_0^{c_i} S(t)\, dt + \int_{c_i}^{\infty} S(t)\, dt \right) \\
&= c_i + \lim_{f(e_i)\to\infty} \int_{c_i}^{\infty} S(t)\, dt\,,
\end{aligned}$$

where the L1-hinge loss max equals to $\max\{(c_i - \mu_i), 0\} = \max\{-\lim_{f(e_i)\to\infty} \int_{c_i}^{\infty} S(t)\, dt, 0\} = 0$.

[9] Administrative censoring refers to the censorship that occurs when the study observation period ends.

C.4. Hosmer-Lemeshow Calibration

Hosmer-Lemeshow calibration (1-calibration) (Hosmer & Lemesbow, 1980) is a statistical test to evaluate the calibration ability of the risk predictions at a specific time. Similar to C-index, we can substitute risk prediction with survival probability prediction to analyze the performance of ISD models.

To calculate 1-calibration at the specific time t^*, we first sort the predicted probabilities at this time for all patients and group them into K bins, $B_1, \ldots, B_K$. Within each bin, we calculate the expected number of events using the predicted probabilities, $\mathbb{E}_{\boldsymbol{x}_i \sim B_k}[S_{\mathrm{m}}(\, t^* \mid \boldsymbol{x}_i\,)]$, and compare it to the observed event rate. Finally, we use the Hosmer-Lemeshow test to assess if the expected and observed event rates are statistically similar. Figure 1 (e) demonstrates this process with four uncensored patients and two groups. In Group 1, the observed event rate is 0.5, as patient A was alive at t=2, while patient B died at the same time point. The expected number of events in this group is calculated as the average of 0.58 and 0.51, based on the probabilities observed at the intersections. Using the same reasoning, the observed event rate is 1 and the expected number of events is $\frac{0.35+0.04}{2}$ in Group 2. To handle censored subjects, we can use the KM estimator to approximate the observed statistics (D'Agostino & Nam, 2003).

1-calibration can aid clinicians in making group-level decisions, such as arranging medical resources in response to COVID-19 lockdown restrictions being lifted. For example, if a 1-calibrated (on 100-th day) model forecasts that the expected number of patients with severe symptoms in 100 days is 10,000, then we should arrange the among of ICU beds correspondingly because the expected number and observed numbers are statistically similar.

BS can be decomposed into a 1-calibration part (DeGroot & Fienberg, 1983). Let's first discretize the probability $S(t \mid \boldsymbol{x}_i)$ by assuming there are K distinct values for the probability. Therefore, everyone's predicted probability at time t^* can be rounded to one of $\{p_k\}_{k=1}^K$. Let's $n_k(t^*)$ be the number of people that has the same probability p_k at time t^*, and $\lambda_k(t^*)$ be the proportion of people in n_k who have event happened before t^*, the BS for a set of uncensored instances can be represented as

$$\mathcal{R}_{\mathrm{BS}}(t^*, (S_{\mathrm{m}}(\,\cdot \mid \boldsymbol{x}_i\,), t_i, \delta_i)) = \frac{1}{N}\sum_{k=1}^{10} n_k(t^*)(\lambda_k(t^*) - p_k)^2 + \frac{1}{N}\sum_{k=1}^{10} n_k(t^*)\lambda_k(t^*)(1-\lambda_k(t^*))\,.$$

where the first term is equivalent to 1-calibration at the target time, and the second term is called refinement or sharpness at the target time. Some people may argue that there is no compelling demand to use 1-calibration if BS or IBS are used as the evaluation metrics. However, as we can see from the decomposition, it is trivial to conclude that BS can be large while the 1-calibration term can be small. Therefore, if a model has a bad (large) BS or IBS score, it doesn't necessarily mean that it doesn't 1-calibrated.

C.5. Distribution Calibration

Distribution calibration (D-calibration) (Haider et al., 2020) is also a statistical test to evaluate the calibration ability of the entire ISD prediction. As to the notation, for any probability interval $[a, b] \subset [0, 1]$, let

$$\mathcal{D}_{\mathrm{m}}(a, b) = \{[\boldsymbol{x}_i, t_i, \delta_i = 1] \in \mathcal{D} \mid S_{\mathrm{m}}(\, t_i \mid \boldsymbol{x}_i\,) \in [a, b]\}\,,$$

be the subset of the subjects in the dataset $\mathcal{D}$ whose predicted probability at its event time, $S_{\mathrm{m}}(\, t_i \mid \boldsymbol{x}_i\,)$, is in the interval $[a, b]$. The model is D-calibrated if the proportion of patients $\frac{|D_m(a,b)|}{|D|}$ is statistically similar to the proportion $b - a$. Models can be trained to be D-calibrated using a differentiable relaxation (Goldstein et al., 2020). Haider et al. (2020) advises using equal-sized, mutually exclusive intervals with Pearson's χ^2 test to examine if the proportion of patients in each bin is uniformly distributed. Figure 1 (f) provides a visual illustration of a D-calibrated model, where the predicted time-to-event probability is equally distributed across two intervals. To incorporate censored patients, we can "split" each censored patient uniformly among the subsequent probability intervals after the time-to-censor-probability (Haider et al., 2020).

As D-calibration is a goodness-of-fit test that involves the entire distribution rather than a single time point, clinicians can use it to make group-level decisions with more flexibility (e.g., estimate the number of ICU beds available in 10 days for a group of patients recruited at various periods), see motivation in Kumar et al. (2022).

D-calibration and MAE measure different aspects of the performance, meaning they can give different orderings of models. For example, a KM model, despite being a perfectly D-calibrated model, will predict the same time for everyone, which can produce a terrible MAE score. Contrarily, a model with zero MAE score can have $S_{\mathrm{m}}(\, t_i \mid \boldsymbol{x}_i\,) = 0.5$ for all the subjects,

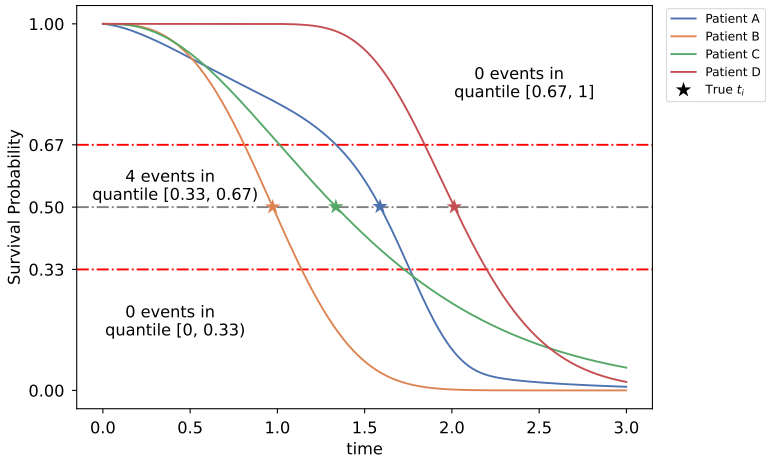

Figure 5. A toy example with four uncensored subjects, with best-possible MAE (MAE = 0) while not D-calibrated. Note that the true event times (position of stars) are overlapped with predicted median survival times.

resulting in one of the probability intervals containing all the patients for D-calibration (p-value = 0). This example is illustrated in Figure 5. Below we provide a theoretical comparison between these two metrics.

Proposition C.4. *To demonstrate that MAE is fundamentally distinct from D-calibration, we show that it is possible for*

- *a model to have small MAE (in fact, 0), but not be D-calibrated; and*
- *a model to be perfectly D-calibrated, but have arbitrarily large MAE (relative to the entire range of times).*

Proof. In the following section, we prove the above proposition using uncensored datasets $\mathcal{D} = \{(\boldsymbol{x}_i, t_i, \delta_i = 1)\}_{i=1}^N$ for the sake of simplicity.

C.5.1. Small MAE but Poor D-calibration

Assume survival curve $S_{\mathrm{m}}(\cdot \mid \boldsymbol{x}_i)$ for subject $\boldsymbol{x}_i$ is a logistic function, centered at the correct time-of-death t_i. Here, the median survival time of the estimated ISD is t_i, and the MAE error is zero (best MAE score possible) for all the subjects. However, as $S_{\mathrm{m}}(t_i \mid \boldsymbol{x}_i) = 0.5$ for all the subjects (because the logistic function is censored at t_i), therefore, one of the probability intervals must contain all the patients, leading to non-D-calibration.

C.5.2. D-calibrated but Large MAE

As the MAE has a physical interpretation, which is obviously related to the time "units" (*i.e.*, minutes, days, or years), we will normalize the times by dividing it by the largest value of the event times, so the largest value is 1 ($t_i \leq 1$ for all i). For any constant $R > 1$, we will provide a model that is D-calibrated (using 2 discretized bins), but the MAE is $\geq$ R. This hypothetical model produces linear survival curves. Let assume that, for each subject x_i (with time-of-death e_i), the curve starts from $(0, 1)$ and ends at $(\frac{t_i}{1-\kappa_i}, 0)$, and it descends linearly through (t_i, κ_i) and crosses 0.5 at $(\frac{t_i}{2(1-\kappa_i)}, 0.5)$ (see Figure 6).

For the first half subjects, define κ_i (the probability associated with the time of death t_i) to be $\kappa_i = 1 - \frac{t_i}{4R+2t_i}$ and for the other half subjects, set $\kappa_i = 0.25$.

To show this model is D-calibrated with 2 bins ($|B| = 2$), observe that $t_i \leq 1 \leq R$ implies $\frac{t_i}{4R+2t_i} \leq \frac{t_i}{4R} \leq \frac{1}{4}$, which means $\kappa_i = 1 - \frac{t_i}{4R+2t_i} \geq 1 - \frac{1}{4} = \frac{3}{4} > \frac{1}{2}$. Hence, $S(t_i \mid \boldsymbol{x}_i) > \frac{1}{2}$ for each individual of the first half of the subjects, and $1 - \frac{3}{4} = \frac{1}{4} < \frac{1}{2}$ for the second half.

Now consider the MAE: For the first half of the subjects, the median survival time for each individual is

$$\text{median}(S(\cdot \mid \boldsymbol{x}_i)) = \frac{t_i}{2 \cdot (1 - \kappa_i)} = \frac{t_i}{2} \frac{1}{1 - \kappa_i} = \frac{t_i}{2} \frac{4R + 2t_i}{t_i} = 2R + t_i \,.$$

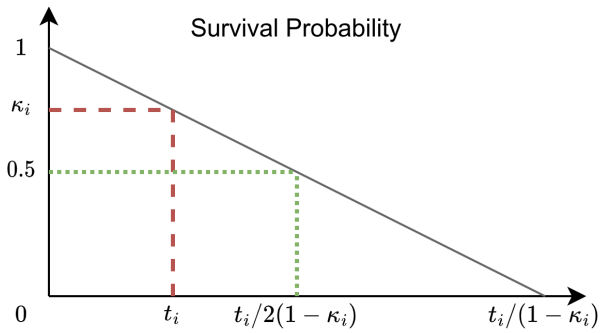

Figure 6. An toy example of a linear survival curve which starts at $(0, 1)$ and crosses (t_i, κ_i).

This means the MAE for each of the patients is $|2R + t_i - t_i| = 2R$.

Hence, the overall MAE, over the entire set of N patients, is

$$\frac{1}{N}\sum_{i=1}^{N}|\text{median}(S(\cdot \mid \boldsymbol{x}_i)) - t_i| \geq \frac{1}{N}\sum_{i=1}^{N}|\text{median}(S(\cdot \mid \boldsymbol{x}_i)) - t_i| = \frac{1}{n}(\frac{n}{2}\cdot 2R) = R\,.$$

Then we complete the proof. □

NOTES:

(1) It is trivial to extend this proof to deal with $|B| = 5$ or 10 different probability intervals (rather than the $|B| = 2$ covered above): For 1 bin, use $\kappa_i = 1 - \frac{t_i}{2|B|\cdot R + 2t_i}$. For the other bins, we can use κ_i in the 2nd, 3rd, . . . , up to B-th interval. So for 10 bins, these would be κ_i = 0.15, then = 0.25, 0.35, . . . , 0.85, 0.95 – and everything will still follow the proof.

(2) If there are censored individuals, it is reasonable to have models that predict event times that extend beyond the final recorded time. But if only uncensored instances, one could argue we should only consider times within the range of the training instances – in our case, with the time-of-events $\in [0, 1]$. Of course, if this is the case, it is easy to have an MAE score bounded by 0.5 by choosing the mid-point time for each person. With this in mind, given any uncensored dataset, we can produce a model that (1) is D-calibrated, and (2) has MAE score $\geq \frac{1}{2}$.

D. Properties of Pseudo-observation

In this section, we prove some relevant conjectures about pseudo-observation (Andersen et al., 2003; Andersen & Pohar Perme, 2010) to help justify the MAE-PO as an evaluation metric.

D.1. Pseudo-observation Values Can be Treated as i.i.d. Random Variable.

Graw et al. (2009) proved that the jackknife pseudo-observation values of competing risks could be expressed as a first Gateaux derivative of the Aalen-Johansen estimator (Aalen, 1978) and the cumulative incidence function, if $N \to \infty$. Therefore, the pseudo-observation probabilities of competing risks can be approximated by independent and identically distributed variables. It is trivial to conclude that, in the absence of a competing risk, the estimated cumulative incidence function will reduce to the cumulative density function (and the Aalen-Johansen estimator will reduce to KM estimator); hence, the property of independent and identically distributed pseudo-observations also holds for the survival function and the mean survival times.

D.2. Authenticity of Pseudo-observation

In this section, we prove that the pseudo-observation value (Andersen et al., 2003) of a censored instance using the Kaplan Meier (KM) estimator (Kaplan & Meier, 1958) is always larger or equal to its censoring time. This property has been empirically demonstrated using synthetic examples in Andersen & Pohar Perme (2010) but we will give a theoretical

proof for the first time. In the following, we represent the KM estimator as a stepwise function (see Equation 11 and 12). Readers can use the linear interpolation version of KM without loss of generality. Furthermore, the Nelson-Aalen (NA) estimator (Nelson, 1972; Aalen, 1978) has already been proven to be asymptotically equivalent to KM estimator (Fleming & Harrington, 2011). Therefore, the following theoretical conclusions also hold if we use the NA estimator to predict survival curves.

First, let's recall the definition of pseudo-observation. For a censored instance $\delta_i = 0$, its pseudo-observation can be expressed as:

$$e_{\text{pseudo-obs}}(i) = N \times \mathbb{E}_t[S_{\text{KM}(\mathcal{D})}(\,t\,)] - (N-1) \times \mathbb{E}_t[S_{\text{KM}(\mathcal{D}^{-i})}(\,t\,)]\,, \tag{10}$$

where $S_{\text{KM}(\mathcal{D})}(\,t\,)$ is called the unbiased survival distribution estimator over the whole population using KM, while the $S_{\text{KM}(\mathcal{D}^{-i})}(\,t\,))$ is called the biased estimator over all the data except *i*-th censored instance. Given these two estimations, we have:

Lemma D.1. *Given an instance censored at c_i with $\delta_i = 0$, the unbiased population-based survival distribution estimator is always greater than or equal to the biased population-based estimators. Formally,*

$$S_{KM(\mathcal{D})}(\,t\,) \geq S_{KM(\mathcal{D}^{-i})}(\,t\,), \quad \forall c_i, \forall t\,.$$

Proof. The unbiased KM estimation for the whole dataset can be represented as:

$$S_{\text{KM}(\mathcal{D})}(\,t\,) \;=\; \prod_{k:\,0 \leq t_k < t} \frac{n_k - d_k}{n_k}\,, \tag{11}$$

with $t_k \in \mathbb{R}^+$ denotes a time when at least one event happened, $d_k \in \mathbb{Z}^+$ denotes the number of events that happened at time t_k, and $n_k \in \mathbb{Z}^+$ is the number of individuals known to be at-risk (have not yet had an event or been censored) up to time t_k. Then, with the same notation, the biased KM estimation for the leave-*i*-out population can be represented as:

$$S_{\text{KM}(\mathcal{D}^{-i})}(\,t\,) \;=\; \prod_{k:\,0 \leq t_k < t_j} \frac{n_k - 1 - d_k}{n_k - 1} \times \prod_{k:\,t_j \leq t_k < t} \frac{n_k - d_k}{n_k}\,, \tag{12}$$

with $t_j \in \mathbb{R}^+$ denotes the next time after the censored time c_m when at least one event happened. By definition, t_j and c_i will have the relationship $t_{j-1} < c_i \leq t_j$. Both the nominator and denominator decrease by 1 before t_j due to this censored instance being eliminated from the at-risk population before t_j. Therefore, for any natural number d_k given t_k, we have

$$\frac{n_k - d_k}{n_k} \geq \frac{n_k - 1 - d_k}{n_k - 1}, \quad \forall t_k \implies \prod_{k:\,0 \leq t_k < t_j} \frac{n_k - d_k}{n_k} \geq \prod_{k:\,0 \leq t_k < t_j} \frac{n_k - 1 - d_k}{n_k - 1} \implies S_{\text{KM}(\mathcal{D})}(\,t\,) \geq S_{\text{KM}(\mathcal{D}^{-i})}(\,t\,)\,.$$

Then the proof is complete. We also present a toy example with visual illustration in Figure 7 to demonstrate this Lemma. As we can see, the leave-C_3-out biased estimator is always lower or equal to the unbiased estimator at all times. □

Lemma D.2. *For a survival dataset with only one censored instance $\{\boldsymbol{x}_m, t_m = c_m, \delta_m = 0\}$ (also $\delta_i = 1$ unless $i = m$), the pseudo-observation value (calculated using KM estimators) for this censored instance is lower bound by its censoring time.*

$$e_{pseudo\text{-}obs}(m) \;=\; N \times \mathbb{E}_t[S_{KM(\mathcal{D})}(\,t\,)] - (N-1) \times \mathbb{E}_t[S_{KM(\mathcal{D}^{-m})}(\,t\,)] \;\geq\; c_m\,.$$

Proof. We start by rearranging the expression in Equation 10.

$$\begin{aligned}
e_{\text{pseudo-obs}}(m) &= \mathbb{E}_t[S_{\text{KM}(\mathcal{D})}(\,t\,)] + (N-1)\left[\mathbb{E}_t[S_{\text{KM}(\mathcal{D})}(\,t\,)] - \mathbb{E}_t[S_{\text{KM}(\mathcal{D}^{-m})}(\,t\,)]\right] \\
&= \int_0^{t_j} S_{\text{KM}(\mathcal{D})}(\,t\,)\,dt + (N-1)\int_0^{t_j} \left(S_{\text{KM}(\mathcal{D})}(\,t\,) - S_{\text{KM}(\mathcal{D}^{-m})}(\,t\,)\right)\,dt \\
&\quad + \int_{t_j}^{\infty} S_{\text{KM}(\mathcal{D})}(\,t\,)\,dt + (N-1)\int_{t_j}^{\infty} \left(S_{\text{KM}(\mathcal{D})}(\,t\,) - S_{\text{KM}(\mathcal{D}^{-m})}(\,t\,)\right)\,dt \\
&\geq \int_0^{t_j} S_{\text{KM}(\mathcal{D})}(\,t\,)\,dt + (N-1)\int_0^{t_j} \left(S_{\text{KM}(\mathcal{D})}(\,t\,) - S_{\text{KM}(\mathcal{D}^{-m})}(\,t\,)\right)\,dt\,.
\end{aligned}$$

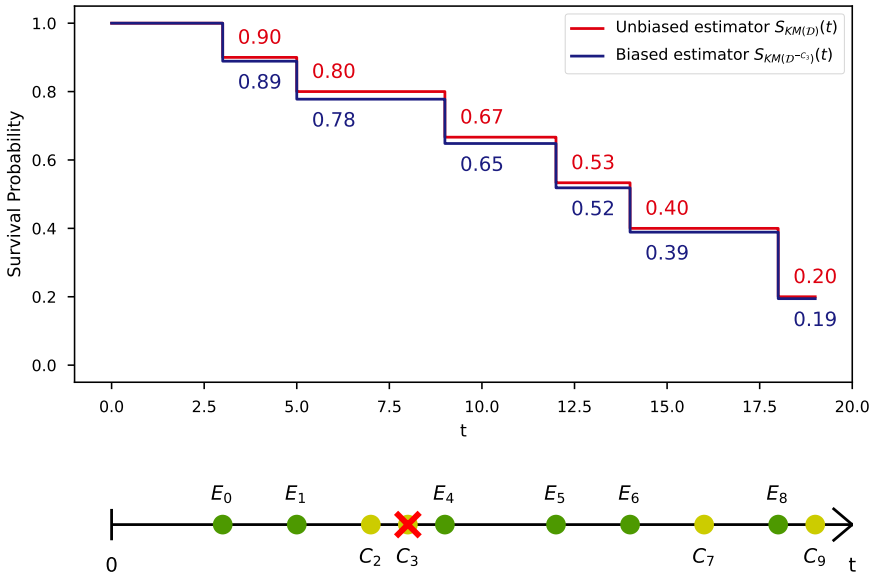

Figure 7. A visual comparison between the unbiased estimator and biased estimator using the Kaplan-Meier method on a toy example. The bottom dots in the x-axis shows the order of the event/censor where E (green dots) and C (yellow dots) represent the event and censor respectively. The biased estimator is a KM estimation for the leave-C_3-out population.

The first equality is due to rearranging the factors. The second equality holds by substituting the KM estimators by Equation 11 and 12. The third equality separates the range for the integral. And the inequality is because of Lemma D.1. Therefore, we can complete the proof as long as we can prove the following inequality:

$$\int_0^{t_j} S_{\text{KM}(\mathcal{D})}(t)\,dt + (N-1)\int_0^{t_j}\left(S_{\text{KM}(\mathcal{D})}(t) - S_{\text{KM}(\mathcal{D}^{-m})}(t)\right)\,dt \;\geq\; c_m\,. \tag{13}$$

Or alternatively, by substituting the KM estimators by Equation 11 and 12,

$$\int_0^{t_j} \prod_{k:\,0\leq t_k<t} \frac{n_k-d_k}{n_k}\,dt + (N-1)\int_0^{t_j}\left(\prod_{k:\,0\leq t_k<t}\frac{n_k-d_k}{n_k} - \prod_{k:\,0\leq t_k<t_j}\frac{n_k-1-d_k}{n_k-1}\right)\,dt \;\geq\; c_m\,.$$

For a survival dataset that contains *only one censored instance*, and $t_{j-1} < c_m \leq t_j$. We have

$$n_{k-1} - d_{k-1} = n_k, \quad \forall t_k \text{ except } k = j\,.$$

Therefore, using the above equation, the first term in Equation 13 can be expressed as:

$$\begin{aligned}
\int_0^{t_j} S_{\text{KM}(\mathcal{D})}(t)\,dt &= \int_0^{t_j}\prod_{k:\,0\leq t_k<t}\frac{n_k-d_k}{n_k}\,dt \\
&= \int_0^{t_1}\frac{n_0-d_0}{n_0}\,dt + \int_{t_1}^{t_2}\frac{n_0-d_0}{n_0}\frac{n_1-d_1}{n_1}\,dt + \cdots + \int_{t_{j-1}}^{t_j}\frac{n_0-d_0}{n_0}\cdots\frac{n_{j-2}-d_{j-2}}{n_{j-2}}\frac{n_{j-1}-d_{j-1}}{n_{j-1}}\,dt \\
&= \int_0^{t_1}\frac{n_0-d_0}{n_0}\,dt + \int_{t_1}^{t_2}\frac{n_1-d_1}{n_0}\,dt + \cdots + \int_{t_{j-1}}^{t_j}\frac{n_{j-1}-d_{j-1}}{n_0}\,dt\,.
\end{aligned} \tag{14}$$

Similarly, the second term can be expressed as:

$$
\begin{aligned}
&(N-1)\int_0^{t_j}\left(S_{\text{KM}(\mathcal{D})}(\,t\,)-S_{\text{KM}(\mathcal{D}^{-m})}(\,t\,)\right)\,dt \\
&\quad=(n-1)\int_0^{t_j}\Big(\prod_{k:\,0\le t_k<t}\frac{n_k-d_k}{n_k}-\prod_{k:\,0\le t_k<t_j}\frac{n_k-1-d_k}{n_k-1}\Big)\,dt \\
&\quad=(n-1)\left(\int_0^{t_1}(\frac{n_0-d_0}{n_0}-\frac{n_0-1-d_0}{n_0-1})\,dt+\int_{t_1}^{t_2}(\frac{n_1-d_1}{n_0}-\frac{n_1-1-d_1}{n_0-1})\,dt\right. \\
&\qquad\left.+\cdots+\int_{t_{j-1}}^{t_j}(\frac{n_{j-1}-d_{j-1}}{n_0}-\frac{n_{j-1}-1-d_{j-1}}{n_0-1})\,dt\right) \\
&\quad=(n-1)\left(\int_0^{t_1}\frac{d_0}{n_0(n_0-1)}\,dt+\int_{t_1}^{t_2}\frac{n_0-n_1+d_1}{n_0(n_0-1)}\,dt+\cdots+\int_{t_{j-1}}^{t_j}\frac{n_0-n_{j-1}+d_{j-1}}{n_0(n_0-1)}\,dt\right) \\
&\quad=\int_0^{t_1}\frac{d_0}{n_0}\,dt+\int_{t_1}^{t_2}\frac{n_0-n_1+d_1}{n_0}\,dt+\cdots+\int_{t_{j-1}}^{t_j}\frac{n_0-n_{j-1}+d_{j-1}}{n_0}\,dt \\
&\quad=\int_0^{t_1}(1-\frac{n_0-d_0}{n_0})\,dt+\int_{t_1}^{t_2}(1-\frac{n_1-d_1}{n_0})\,dt+\cdots+\int_{t_{j-1}}^{t_j}(1-\frac{n_{j-1}-d_{j-1}}{n_0})\,dt\,.
\end{aligned}
\tag{15}
$$

The third equality holds because $n_0 = N$ for the dataset with only one censored instance. We can easily observe that each term in Equation 14 just complements the corresponding term in Equation 15. Therefore, we can have

$$
\int_0^{t_j}S_{\text{KM}(\mathcal{D})}(\,t\,)\,dt+(N-1)\int_0^{t_j}\left(S_{\text{KM}(\mathcal{D})}(\,t\,)-S_{\text{KM}(\mathcal{D}^{-m})}(\,t\,)\right)\,dt=\int_0^{t_j}1\,dt=t_j\;\ge\;c_m\,,
$$

then we complete the proof. □

Theorem D.3. *For a survival dataset with arbitrary numbers of censored and event instances, the pseudo-observation value (calculated using KM estimators) for any censored instance is lower bound by its censoring time.*

$$
e_{\textit{pseudo-obs}}(i)=N\times\mathbb{E}_t[S_{\textit{KM}(\mathcal{D})}(\,t\,)]-(N-1)\times\mathbb{E}_t[S_{\textit{KM}(\mathcal{D}^{-i})}(\,t\,)]\;\ge\;c_i\,.
$$

Proof. We can still follow the intuition and steps in the proof for Lemma D.2. That means, as long as we can prove the correctness of Equation 13 in this circumstance, this theorem can be proved.

In the case of an unlimited number of censor instances, the number of survival instances at the previous time must be large or equal to the at-risk instances at the next time, which means:

$$
\frac{n_{k-1}-d_{k-1}}{n_k}\ge 1,\quad\forall t_k\,,
\tag{16}
$$

$$
\Longrightarrow\qquad\prod_{k:\,t_1\le t_k<t_j}\frac{n_{k-1}-d_{k-1}}{n_k}\ge\prod_{k:\,t_1\le t_k<t_j}\frac{n_{k-1}-1-d_{k-1}}{n_k-1}\ge 1\,.
\tag{17}
$$

Therefore, using the above equations, the first term in Equation 13 can be expressed as:

$$
\begin{aligned}
&\int_0^{t_j}S_{\text{KM}(\mathcal{D})}(\,t\,)\,dt \\
&\quad=\int_0^{t_1}\frac{n_0-d_0}{n_0}\,dt+\int_{t_1}^{t_2}\frac{n_0-d_0}{n_0}\frac{n_1-d_1}{n_1}\,dt+\cdots+\int_{t_{j-1}}^{t_j}\frac{n_0-d_0}{n_0}\cdots\frac{n_{j-2}-d_{j-2}}{n_{j-2}}\frac{n_{j-1}-d_{j-1}}{n_{j-1}}\,dt \\
&\quad=\int_0^{t_1}\frac{n_0-d_0}{n_0}\,dt+\int_{t_1}^{t_2}\frac{n_0-d_0}{n_1}\frac{n_1-d_1}{n_0}\,dt+\cdots+\int_{t_{j-1}}^{t_j}\frac{n_0-d_0}{n_1}\cdots\frac{n_{j-2}-d_{j-2}}{n_{j-1}}\frac{n_{j-1}-d_{j-1}}{n_0}\,dt \\
&\quad\ge\int_0^{t_1}\frac{n_0-d_0}{n_0}\,dt+\int_{t_1}^{t_2}\frac{n_1-d_1}{n_0}\,dt+\cdots+\int_{t_{j-1}}^{t_j}\frac{n_{j-1}-d_{j-1}}{n_0}\,dt\,.
\end{aligned}
\tag{18}
$$

The second equality is simply by changing the position of denominators. The inequality holds because of Equation 16. Similarly, the second term in Equation 13 can be expressed as:

$$
\begin{aligned}
&(N-1)\int_0^{t_j}\left(S_{\text{KM}(\mathcal{D})}(\,t\,)-S_{\text{KM}(\mathcal{D}^{-i})}(\,t\,)\right)\,dt \\
&=(N-1)\left(\int_0^{t_1}(\frac{n_0-d_0}{n_0}-\frac{n_0-1-d_0}{n_0-1})\,dt+\int_{t_1}^{t_2}(\frac{n_0-d_0}{n_1}\frac{n_1-d_1}{n_0}-\frac{n_0-1-d_0}{n_1-1}\frac{n_1-1-d_1}{n_0-1})\,dt\right. \\
&\quad\left.+\cdots+\int_{t_{j-1}}^{t_j}(\frac{n_0-d_0}{n_1}\cdots\frac{n_{j-2}-d_{j-2}}{n_{j-1}}\frac{n_{j-1}-d_{j-1}}{n_0}-\frac{n_0-1-d_0}{n_1-1}\cdots\frac{n_{j-2}-1-d_{j-2}}{n_{j-1}-1}\frac{n_{j-1}-1-d_{j-1}}{n_0-1})\,dt\right) \\
&\geq(N-1)\left(\int_0^{t_1}\frac{d_0}{n_0(n_0-1)}\,dt+\int_{t_1}^{t_2}\frac{n_0-n_1+d_1}{n_0(n_0-1)}\,dt+\cdots+\int_{t_{j-1}}^{t_j}\frac{n_0-n_{j-1}+d_{j-1}}{n_0(n_0-1)}\,dt\right) \\
&\geq\int_0^{t_1}(1-\frac{n_0-d_0}{n_0})\,dt+\int_{t_1}^{t_2}(1-\frac{n_1-d_1}{n_0})\,dt+\cdots+\int_{t_{j-1}}^{t_j}(1-\frac{n_{j-1}-d_{j-1}}{n_0})\,dt\,.
\end{aligned}
\tag{19}
$$

The first equality is also by changing the position of denominators. The first inequality holds because of Equation 17. The second inequality follows the rules of Equation 15, as well as $n_1 \leq N$ for any time for the dataset with unlimited censored instances. As Lemma D.2, we can now combine the Equations 18 and 19:

$$
\int_0^{t_j} S_{\text{KM}(\mathcal{D})}(\,t\,)\,dt+(N-1)\int_0^{t_j}\left(S_{\text{KM}(\mathcal{D})}(\,t\,)-S_{\text{KM}(\mathcal{D}^{-i})}(\,t\,)\right)\,dt\geq\int_0^{t_j}1\,dt=t_j\;\geq\;c_i\,,
$$

then we complete the proof. □

D.3. Susceptible to Dataset Size

We observe that the pseudo-observation value for a censored subject in the dataset is not "invariant" under duplication of subjects in the dataset, unlike other MAE-based metrics such as MAE-margin or MAE-IPCW-T. For example, if we have a dataset $\mathcal{D}$ with 100 subjects and one censored subject, i, we can calculate the pseudo-observation value for i using two Kaplan-Meier (KM) estimations. However, if we duplicate every subject in $\mathcal{D}$ and create a new dataset $\mathcal{D}'$ with 200 subjects and two censored subjects (including i), the pseudo-observation values for i in $\mathcal{D}$ will not be the same as the PO for that one i in $\mathcal{D}$, despite the KM curves being identical in both datasets ($S_{\text{KM}(\mathcal{D})}(\,t\,)=S_{\text{KM}(\mathcal{D}')}(\,t\,)$ for all t). We called this property "susceptible to dataset size".

However, we view this property neither as an advantage nor a limitation. In one way, it is arguable that duplicating the subjects violates the i.i.d. assumption of the data, which makes two different datasets and therefore the surrogate event times for the same subjects should be different. In another way, if the KM curves represent the true survival distribution of the datasets, then the same subjects in the same survival distribution should have the same surrogate times, no matter the dataset size.

D.4. Relationship between Pseudo-observation and Other MAE Metrics

In this section, we investigate the properties of pseudo-observations by comparing them with other MAE metrics.

Theorem D.4. *For a survival dataset with only one censored instance $\{\boldsymbol{x}_m, t_m = c_m, \delta_m = 0\}$ (also $\delta_i = 1$ unless $i = m$), the pseudo-observation value and the margin "best-guess" value (both calculated using KM estimators) are the same, $e_{\textit{pseudo-obs}}(m) = e_{\textit{margin}}(m)$. By their definition, it equals to:*

$$
N\times\mathbb{E}_t[S_{\textit{KM}(\mathcal{D})}(\,t\,)]-(N-1)\times\mathbb{E}_t[S_{\textit{KM}(\mathcal{D}^{-m})}(\,t\,)]=c_m+\frac{\int_{c_m}^{\infty}S_{\textit{KM}(\mathcal{D})}(\,t\,)dt}{S_{\textit{KM}(\mathcal{D})}(\,c_m\,)}\,.
$$

Proof. Recall the proof in Lemma D.2, we can have:

$$
\begin{aligned}
e_{\text{pseudo-obs}}(m) &= \int_0^{t_j} S_{\text{KM}(\mathcal{D})}(\,t\,)\,dt + (n-1)\int_0^{t_j}\left(S_{\text{KM}(\mathcal{D})}(\,t\,) - S_{\text{KM}(\mathcal{D}^{-m})}(\,t\,)\right)\,dt \\
&\quad + \int_{t_j}^{\infty} S_{\text{KM}(\mathcal{D})}(\,t\,)\,dt + (n-1)\int_{t_j}^{\infty}\left(S_{\text{KM}(\mathcal{D})}(\,t\,) - S_{\text{KM}(\mathcal{D}^{-m})}(\,t\,)\right)\,dt \\
&= t_j + \int_{t_j}^{\infty} S_{\text{KM}(\mathcal{D})}(\,t\,)\,dt + (n-1)\int_{t_j}^{\infty}\left(S_{\text{KM}(\mathcal{D})}(\,t\,) - S_{\text{KM}(\mathcal{D}^{-m})}(\,t\,)\right)\,dt\,,
\end{aligned}
$$

in the circumstance of only one censored instance. The sum of the first two terms is equal to t_j because of Equation 14 and 15. Then, we can replace the unbiased and biased KM estimation by Equation 11 and 12 (also use the fact that $n_{k-1} - d_{k-1} = n_k$ for all t_k except $k = j$).

$$
\begin{aligned}
e_{\text{pseudo-obs}}(m) &= t_j + \int_{t_j}^{\infty} \frac{n_{j-1} - d_{j-1}}{n_0} \prod_{k:\, t_j \le t_k < t} \frac{n_k - d_k}{n_k}\,dt \\
&\quad + (N-1)\int_{t_j}^{\infty} \left(\frac{n_{j-1} - d_{j-1}}{n_0} - \frac{n_{j-1} - 1 - d_{j-1}}{n_0 - 1}\right) \prod_{k:\, t_j \le t_k < t} \frac{n_k - d_k}{n_k}\,dt \\
&= t_j + \int_{t_j}^{\infty} \left(\frac{n_{j-1} - d_{j-1}}{n_0} + \frac{n_0 - n_{j-1} + d_{j-1}}{n_0}\right) \prod_{k:\, t_j \le t_k < t} \frac{n_k - d_k}{n_k}\,dt \\
&= t_j + \int_{t_j}^{\infty} \prod_{k:\, t_j \le t_k < t} \frac{n_k - d_k}{n_k}\,dt\,.
\end{aligned}
$$

The second equality is from factorization with $N = n_0$. Similarly, we can get a derivation for the margin best-guess time:

$$
\begin{aligned}
e_{\text{margin}}(m) &= c_m + \frac{\int_{c_m}^{\infty} S_{\text{KM}(\mathcal{D})}(\,t\,)\,dt}{S_{\text{KM}(\mathcal{D})}(\,c_m\,)} \\
&= c_m + \frac{\int_{c_m}^{t_j} S_{\text{KM}(\mathcal{D})}(\,c_m\,)\,dt + \int_{t_j}^{\infty} S_{\text{KM}(\mathcal{D})}(\,t\,)\,dt}{S_{\text{KM}(\mathcal{D})}(\,c_m\,)} \\
&= c_m + t_j - c_m + \frac{\int_{t_j}^{\infty} S_{\text{KM}(\mathcal{D})}(\,c_m\,) \prod_{k:\, t_j \le t_k < t} \frac{n_k - d_k}{n_k}\,dt}{S_{\text{KM}(\mathcal{D})}(\,c_m\,)} \\
&= t_j + \int_{t_j}^{\infty} \prod_{k:\, t_j \le t_k < t} \frac{n_k - d_k}{n_k}\,dt\,.
\end{aligned}
$$

Here we complete the proof by showing that the derivation of pseudo-observations and margin best-guess values is the same. Please note that this derivation for the margin best-guess time is not limited to this special dataset with only one censored instance. □

Lemma D.5. *For a survival dataset with arbitrary numbers of censored and event instances, the pseudo-observation value for any censored instance is always higher or equal to the margin best-guess value. Formally:*

$$
N \times \mathbb{E}_t[S_{KM(\mathcal{D})}(\,t\,)] - (N-1) \times \mathbb{E}_t[S_{KM(\mathcal{D}^{-m})}(\,t\,)] \ge c_m + \frac{\int_{c_m}^{\infty} S_{KM(\mathcal{D})}(\,t\,)\,dt}{S_{KM(\mathcal{D})}(\,c_m\,)} = t_j + \int_{t_j}^{\infty} \prod_{k:\, t_j \le t_k < t} \frac{n_k - d_k}{n_k}\,dt\,.
$$

Proof. In the case of arbitrary numbers of censored and event instances, the pseudo-observation value has the following relationship with the KM estimators:

$$
e_{\text{pseudo-obs}}(m) \ge t_j + \int_{t_j}^{\infty} S_{\text{KM}(\mathcal{D})}(\,t\,)\,dt + (N-1)\int_{t_j}^{\infty}\left(S_{\text{KM}(\mathcal{D})}(\,t\,) - S_{\text{KM}(\mathcal{D}^{-m})}(\,t\,)\right)\,dt\,,
$$

The inequality is due to the derivation in Equation 18 and 19. Following the idea in Theorem D.3, it is trivial to prove that the sum of the second and third terms in the above equation is greater than or equal to $\int_{t_j}^{\infty} \prod_{k:\, t_j \le t_k < t} \frac{n_k - d_k}{n_k}\,dt$. Then we complete the proof. □

E. Experimental Details

E.1. Datasets and Preprocessing

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

E.1.1. GBM

GBM 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 dataset. The data from TCGA can be found on `http://firebrowse.org/` or by the instruction in Haider et al. (2020). There are three features (radiation therapy, Karnofsky performance score, and ethnicity) containing the missing values. We will use their median value to fill in the missing values.

E.1.2. SUPPORT

The Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT) dataset (Knaus et al., 1995) comprises 8873 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 median value imputation.

E.1.3. METABRIC

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) (Curtis et al., 2012) contains survival information for breast cancer patients. The feature sets contain genetic and protein expression features. The dataset can be downloaded from (`https://github.com/havakv/pycox`), and it does not have any missing values.

E.1.4. MIMIC-IV

The Medical Information Mart for Intensive Care (MIMIC)-IV (Johnson et al., 2022) dataset is an update to MIMIC-III, which provides critical care data from patients admitted to hospital and intensive care units (ICU). We create two datasets using the MIMIC-IV database.

MIMIC-IV (all-cause mortality) contains patients that are alive at least 24 hours after being admitted to ICU. Their date of death is derived from hospital records or state records, which means, the cause of mortality is not limited to the reason for ICU admission. We follow the instruction from (Han et al., 2022) to make the dataset. However, there are two differences between our dataset and theirs. The first is that their paper used MIMIC-IV v1.4 while we used the latest version, MIMIC-IV v2.0. The second is if a patient got admitted to ICU multiple times, we will only include the last admission while they will consider each visit as a separate data. The SQL code and python code that prepossesses the data from MIMIC-IV database is available in our GitHub repository.

MIMIC-IV (hospital cause mortality) contains patients that are alive at least 24 hours after admitting to the hospital. Their date of death is only derived from hospital records, so the direct cause of death will be the same as the cause of hospital admission. Because time-series lab features are only available for patients admitted to ICU, we can only use the demographic and clinical features to describe each patient. The SQL code and python code that prepossesses the data from MIMIC-IV database is available in our GitHub repository.

E.2. Model Implementation Details and Hyperparameter Choices

In this section, we will describe the implementation of the models utilized in the performance comparison. Table 4 also provides a summary of the model comparison. Because the purpose of this work is to compare evaluation metrics, extensive hyperparameter searches for each model are unnecessary. Rather, we would like that the models display distinguishable performance. Please do not view these results as a definitive evaluation of the robust performance of survival models.

Linear regression (LR) is a regressor model. We implemented an LR model using an Adam optimizer. LR is not a survival prediction model as it cannot handle censored subjects nor generate ISD and risk predictions. Instead, we only used the

Table 4. Comparison between the baseline time-to-event models.

	Survival Curves	Individual Prediction	Time-Prediction	Continuous
LR	✗	✓	✓	N/A
KM	✓	✗	✗	✗
CoxPH	✓†	✓	✗	✓‡
AFT	✓	✓	✓	✓
RSF	✓	✓	✗	✗
GBCM	✓	✓	✗	✗
MTLR	✓	✓	✗	✗
DeepHit	✓	✓	✗	✗
SCA	✓	✓	✗	✓
S-MDN	✓	✓	✓	✓

† Naive CoxPH only predicts risk scores, whereas its Breslow extension allows the model to generate survival curves.
‡ Although CoxPH model assumes the risk is a time-invariant score in continuous time, the baseline hazard function is estimated through the discrete-time Breslow estimator.

uncensored subjects in the training set to train the model. The model generated the estimated event time for the full test set, and we can perform the evaluation on the full test set. The model is implemented using `scikit-learn` packages.

Kaplan Meier (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. We use the median survival time of the training set's population-level survival distribution as the predicted time for all the testing subjects and perform the evaluation. The model is implemented using `lifelines` packages.

CoxPH (Cox, 1972) with Breslow estimator (Breslow, 1975) is a semi-parametric model. It consists of a population-level baseline hazard function (non-parametric) and a partial hazard function (parametric). In the model implementation, the population-level baseline hazard function is estimated using Breslow method (Breslow, 1975). And the partial hazard function is estimated by a linear function. The model is optimized using partial likelihood loss (Cox, 1975) and Adam optimizer. The model is implemented in the code base attached. The early stop technique is applied to the model via validating on a separate validation set

AFT (Stute, 1993) with Weibull distribution is a parametric model with two estimated coefficients (a scale parameter and a shape parameter). We add a small l2 penalty to the loss for the model optimization. The method is implemented using `lifelines` packages.

GBM-C (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. The method is implemented using `scikit-survival` packages.

RSF (Ishwaran et al., 2008) is also an ensemble estimator that fits a number of survival trees on bootstrapping datasets. We use 50 trees with 3 minimal samples per leaf to fit the model. The method is implemented using `scikit-survival` packages.

MTLR (Yu et al., 2011) is a discrete model that directly models the survival distribution for each individual. 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). The early stop technique is applied to the model via validating on a separate validation set The model is implemented in the code base attached.

DeepHit (Lee et al., 2018) is also a discrete model. It models the probability density function of the event for each individual (and PDF can be used to calculate the survival distribution accordingly). The number of discrete times is determined by the square root of number of uncensored patients, just like MTLR. However, the time interval is uniformly split from time zero to the last observed time, as in the original paper (Lee et al., 2018). Early stopping is also performed during the optimization. The model is implemented using `pycox` packages.

SCA (Chapfuwa et al., 2020) models the covariates into a mixture-of-distributions latent space. And each component in the latent space will be used to stochastically predict/sample the survival distribution. We will use a three-hidden-layer structure

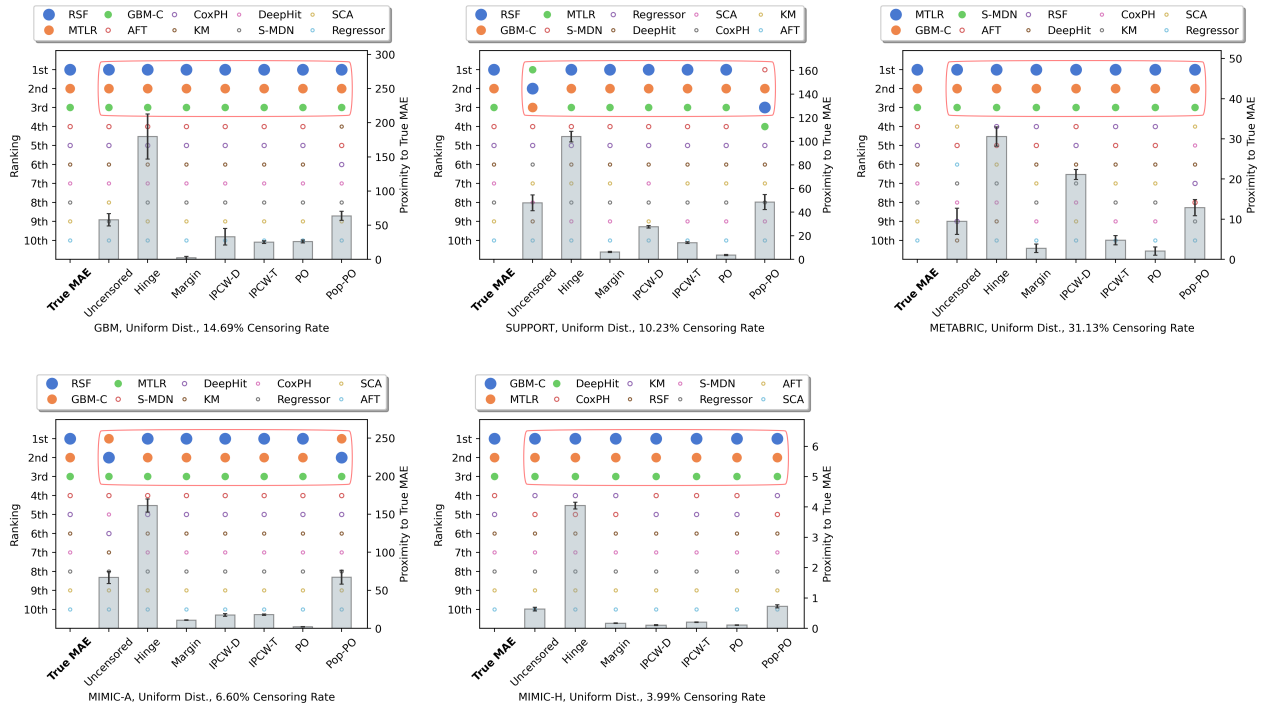

Figure 8. Evaluation metrics comparison on uniform censoring in terms of ranking accuracy (left axis) and proximity to true MAE (right axis).

with dimensions of [50, 50, 50]. We set the number of components to 25, kept the probability for weights equal to 0.8, and set the sample size to 200. Early stopping is also performed with at least 10000 epochs for guaranteed improvement. The model is implemented using the code base in `https://github.com/paidamoyo/survival_cluster_analysis`.

S-MDN (Han et al., 2022) uses Mixture Density Networks to model the survival distributions. The model architecture in the experiment has one hidden layer with a size of 15. The number of components is set to 15, and use residual as the initial type. The model is optimized via RMSprop optimizer with early stopping. The model is implemented using the code base in `https://github.com/XintianHan/Survival-MDN`.

For further details, we refer to the code base attached.

F. Complete Results

F.1. Uniform Censorship

The five subplots in Figure 8 demonstrate the metrics performance for uniform censoring distributions. The last column in each subplot shows the ablation study of MAE-population-pseudo-observation (MAE-Pop-PO), in addition to the true MAE and six MAE-inspired metrics presented in Section 3. All the MAE-inspired metrics discussed in Section 3 can accurately identify the top-three models in all five datasets. While MAE-margin is the closest one to the true MAE score in the GBM dataset, MAE-PO has the smallest difference in SUPPORT, METABRIC, MIMIC-A, and MIMIC-H in terms of true MAE proximity. The results of MAE-Pop-PO show it neither has advantages in ranking the models (incorrectly ranking the top-three models for SUPPORT) nor can approximate the true MAE value (second largest in 4 subplots). We can conclude that for the uniform distribution, MAE-PO is the best here with MAE-margin as the second best one.

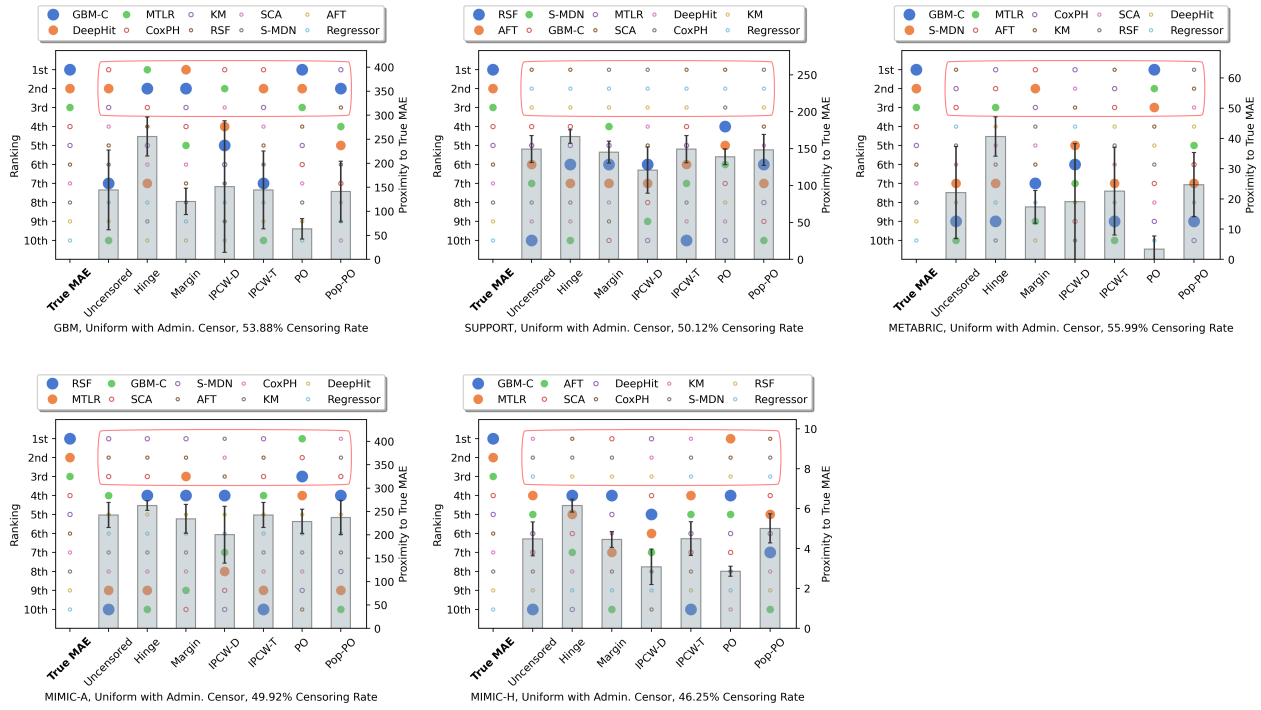

Figure 9. Evaluation metrics comparison on uniform with administrative censoring in terms of ranking accuracy (left axis) and proximity to true MAE (right axis).

F.2. Uniform with Administrative Censorship

The five subplots in Figure 9 demonstrate the metrics performance for uniform censoring distributions with administrative censoring. Due to the large percentage of administrative censoring, all semi-synthetic datasets have very large percentage censoring rates. The MAE-PO is again the best here, in both ranking performance (as it is the only one that correctly identifies the top-three models in GBM and METABRIC, the only one that identifies two of the top-three models for MIMIC-A, and the only one that identifies one of the top-three models for MIMIC-H) and closeness to the true MAE (significantly better in GBM, METABRIC and MIMIC-H with p-value < 0.05, and one of the best in SUPPORT and MIMIC-A).

MAE-margin is the runner-up as it can identify parts of the best-performing models and has the second closest difference to true MAE. Between IPCW-D and IPCW-T, the performance does not have a significant difference in both ranking and proximity to true MAE. However, IPCW-D is associated with quite large error bars, which may be because the accuracy of later uncensored subjects will dominate the score (as we discussed in Section 3.4).

Note that the GBM-C model, while doing very well (best in GBM and METABRIC) for the true MAE, does not achieve a good ranking for uncensored subjects (7th in GBM, 9th in METABRIC, and 10th in MIMIC-H for MAE-uncensored score). Most other MAE variants also consider GBM-C as an "inefficient" model, while only the pseudo-observation can disclose GBM-C's true performance (1st in GBM and METABRIC, and 4th in MIMIC-H).

As to the results of MAE-Pop-PO, it has relatively large errors to the true MAE, and it shows no benefit for identifying the top models.

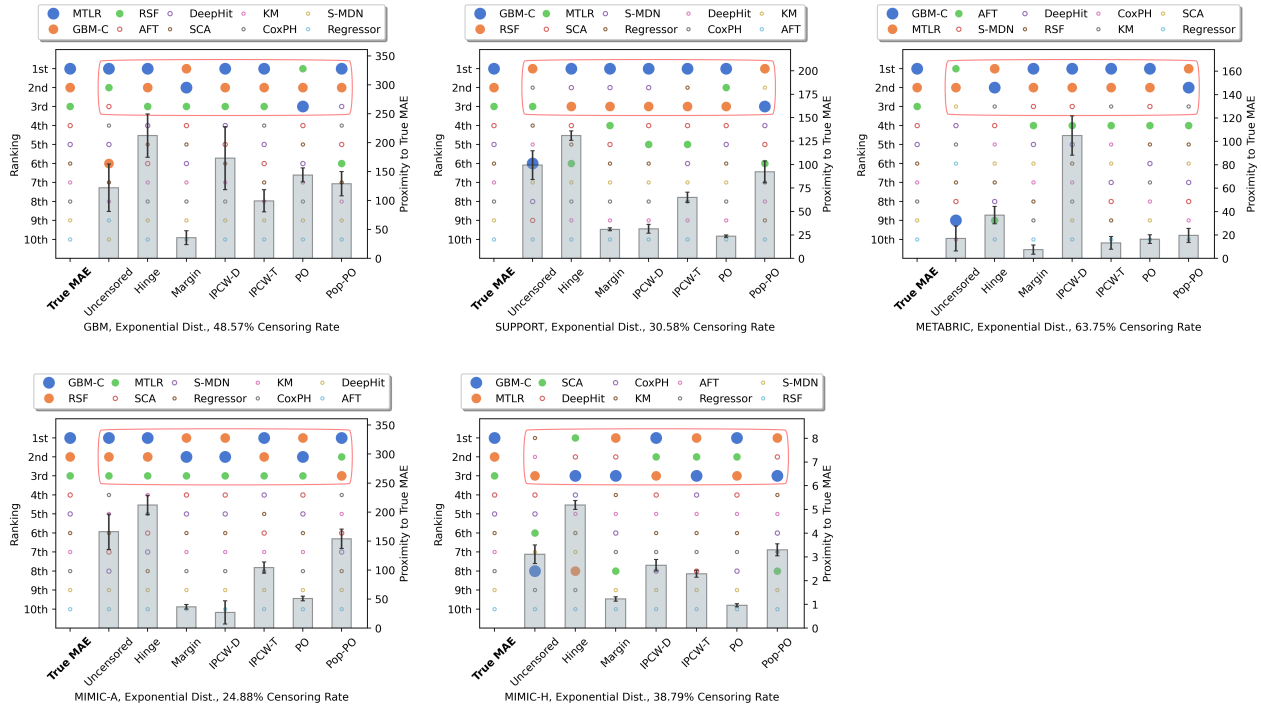

Figure 10. Evaluation metrics comparison on exponential censoring in terms of ranking accuracy (left axis) and proximity to true MAE (right axis).

F.3. Exponential Censorship

The five subplots in Figure 10 demonstrate the metrics performance for exponential censoring distributions. MAE-margin and MAE-PO show comparable performance in these five semi-synthetic datasets. MAE-margin correctly identifies the top models in GBM and MIMIC-A datasets and has a significantly smaller difference to true MAE compared to MAE-PO (p-value < 0.05) on GBM, METABRIC, and MIMIC-A datasets. MAE-PO, on the other side, excels in identifying all the top-three models in all five semi-synthetic datasets, and has a significantly smaller difference to true MAE (p-value < 0.05) on SUPPORT and MIMIC-H datasets.

We notice that MAE-IPCW-D is always associated with a larger standard deviation when it comes to the proximity to true MAE, this is again due to the reason we discussed in Section 3.4. MAE-Pop-PO does not provide any advantages when ranking models, nor can it approximate the true MAE value.

F.4. Feature-Independent Original Censorship

The five subplots in Figure 11 demonstrate the metrics performance for feature-independent original censoring distribution. Among all the evaluation metrics, margin and pseudo-observation perform equally the best for identifying the top three performing models (correctly identifying the best models for GBM, SUPPORT, METABRIC, and MIMIC-A, while recognizing two of the top-three model for MIMIC-H). MAE-PO has a slight advantage in the proximity to true MAE. Its value is closer to the true MAE on GBM and significantly closer on METABRIC and MIMIC-A datasets. In addition, we also observe that IPCW-D is always associated with a large variance (reason explained in Section 3.4). As to the results of MAE-Pop-PO, it again shows no promising results compared to MAE-margin or MAE-PO.

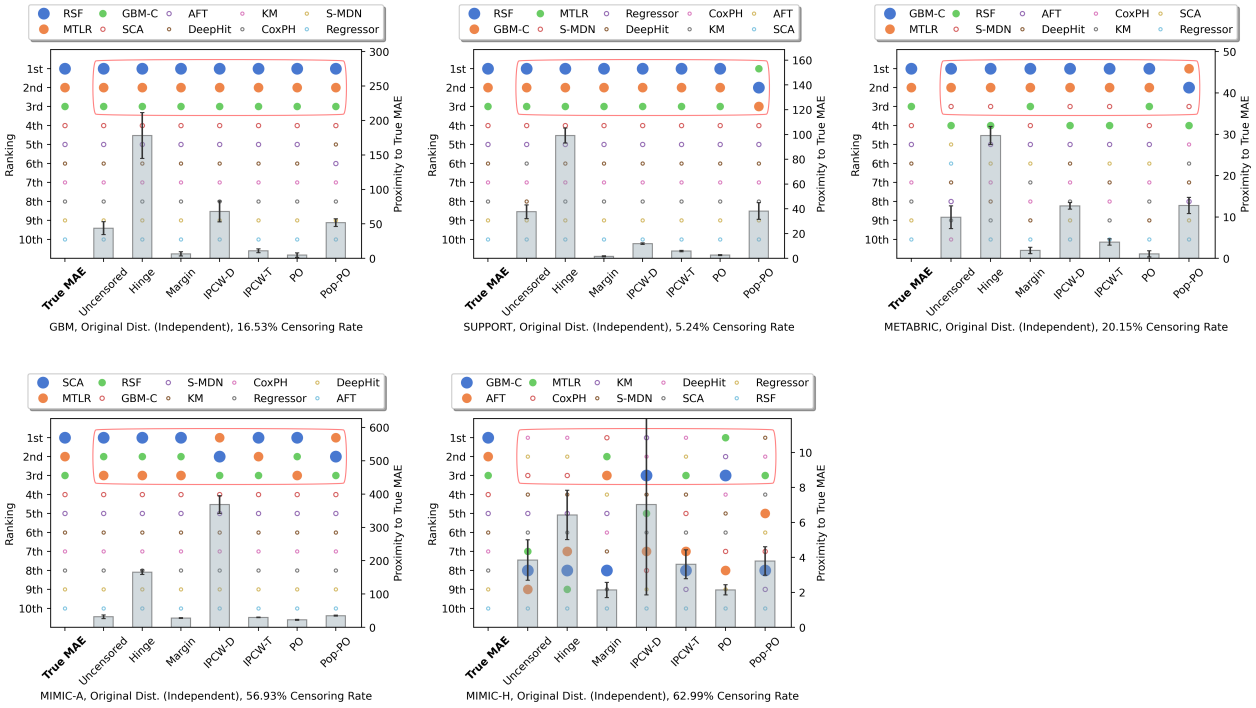

Figure 11. Evaluation metrics comparison on feature-independent original censoring in terms of ranking accuracy (left axis) and proximity to true MAE (right axis).

F.5. Feature-Dependent Original Censorship

The five subplots in Figure 12 demonstrate the metrics performance for feature-independent original censoring distribution. MAE-uncensored performs the best on GBM datasets, which may be due to the low synthetic censoring rate of this dataset (8.37% censoring rate), meaning the whole dataset could be approximately represented by the uncensored population. Among the MAE metrics that can handle the censored subjects, MAE-margin, IPCW-T, and MAE-PO perform equally well on GBM and MIMIC-A datasets. For the METABRIC dataset, pseudo-observation is the best metric as it has the significantly lowest error to true MAE among all the metrics that can identify the top-three performing models. Pseudo-observation is also the optimal metric for the MIMIC-H dataset, as it is the only metric that can identify the top three models (GBM-C, MTLR, and SCA).

F.6. External GBM Dataset Censorship

The four subplots in Figure 13 demonstrate the metrics performance for external dataset censor distribution using the GBM dataset. We only have four subplots because the GBM dataset with external dataset censoring will just be the same as feature-independent original censorship. MAE-PO is the optimal metric in all four semi-synthetic datasets, as it correctly identifies the top-three models in all cases, and has the significantly lowest error to true MAE among all other MAE-inspired metrics.

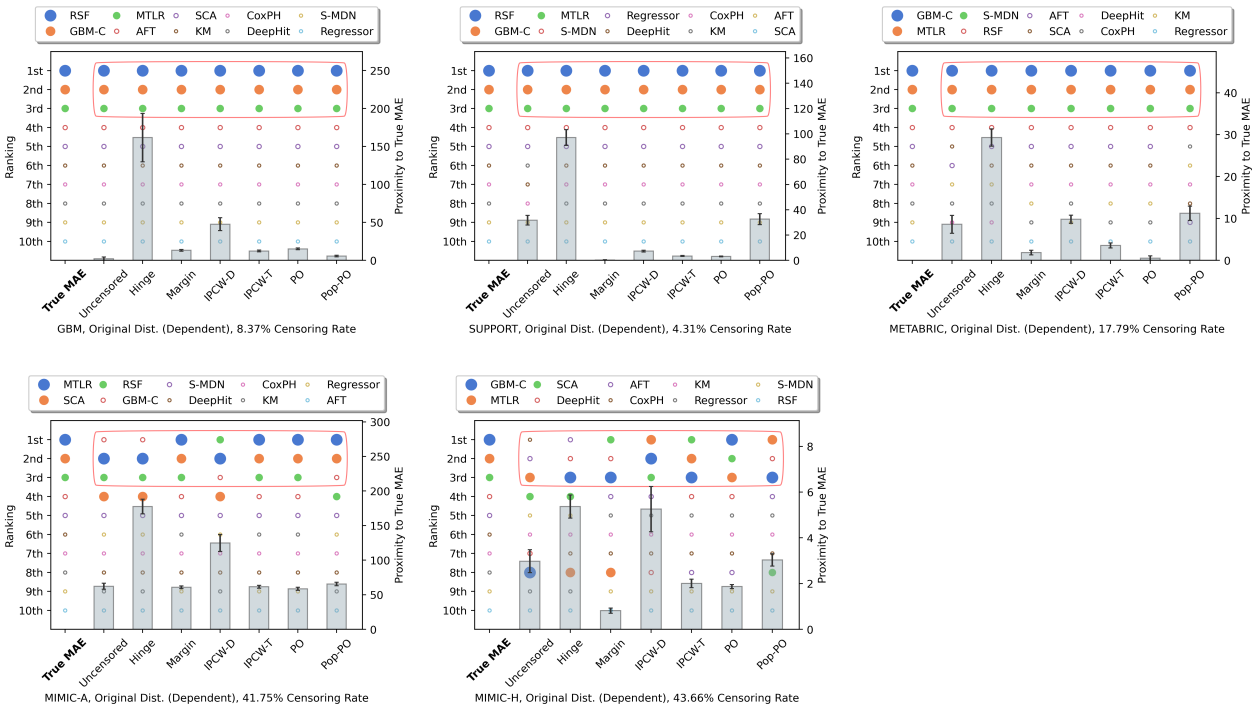

Figure 12. Evaluation metrics comparison on feature-dependent original censoring in terms of ranking accuracy (left axis) and proximity to true MAE (right axis).

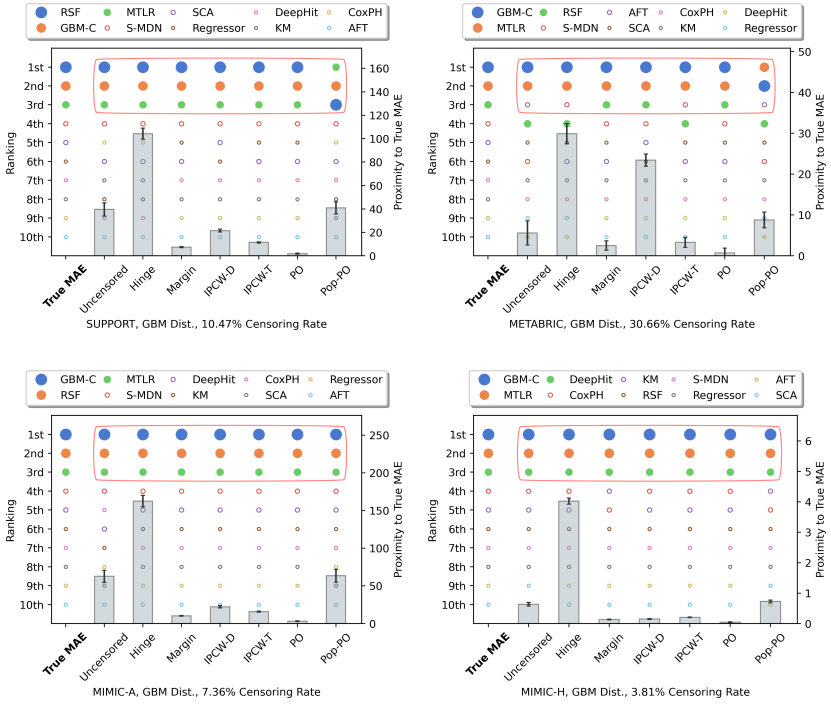

Figure 13. Evaluation metrics comparison on censoring distribution from GBM dataset in terms of ranking accuracy (left axis) and proximity to true MAE (right axis).