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]

[1]University of Alberta, Canada [2]Alberta Machine Intelligence Institute, Canada [3]New York University, USA [4]Duke University, USA.

"All models are wrong, but some are useful." — Dr. George Box

Objective

What is an appropriate metric for evaluating survival models?

Survival Analysis Background

Survival dataset $\mathcal{D} = \{(\boldsymbol{x}_i, t_i, \delta_i)\}_{i=1}^N$

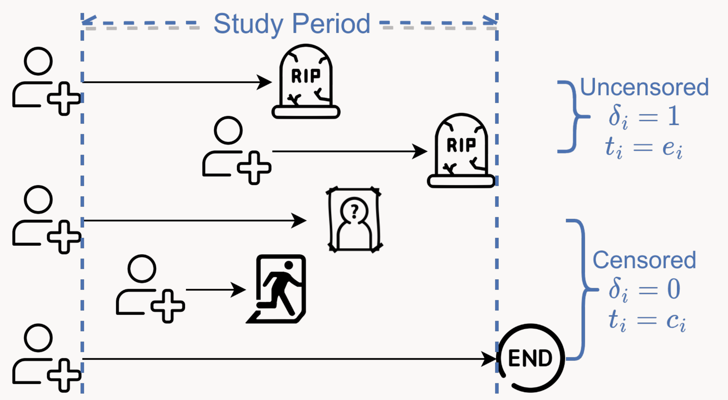
Features $\boldsymbol{x}_i$, observed time t_i, event indicator δ_i.

Each patient i has an event time e_i and a censoring time c_i.

$$t_i \triangleq \min\{e_i, c_i\} \quad \text{and} \quad \delta_i \triangleq \mathbb{1}[e_i \le c_i]$$

A subject is **right-censored** iff s/he has not experienced an event at the observed time.

Assumption: **Independent censoring**, $e_i \perp c_i \mid \boldsymbol{x}_i$



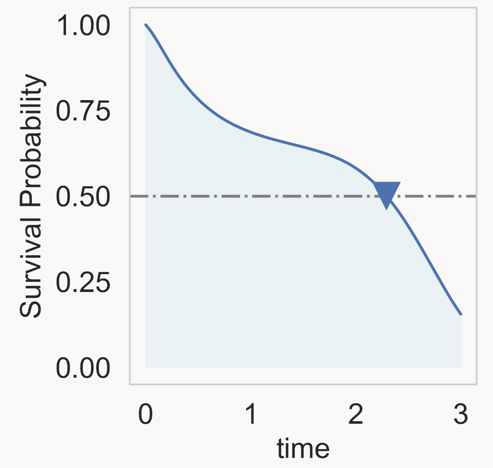
Individual Survival Distribution (ISD) is a probability curve for all future time points for a patient:

$$S(t \mid \boldsymbol{x}_i) = P(T > t \mid \mathbf{X} = \boldsymbol{x}_i)$$

A predicted event time $\hat{t}_i$ can then be represented by either mean survival time (blue area) or median survival time (triangle):

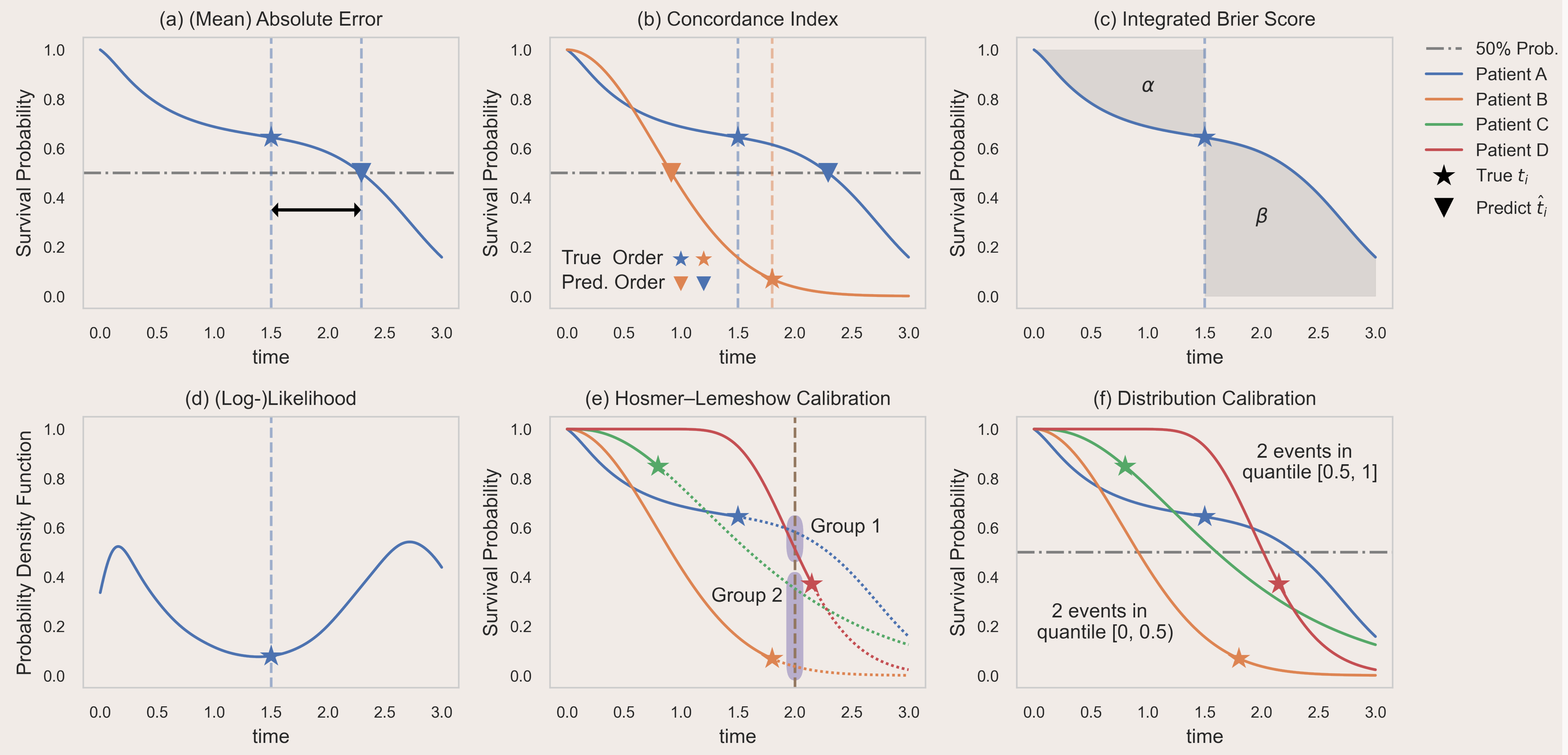
$$\hat{t}_{i,\text{mean}} = \mathbb{E}_t[S(t \mid \boldsymbol{x}_i)] = \int_0^\infty S(t \mid \boldsymbol{x}_i)\, dt$$

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



Evaluation Metrics for Survival Analysis (for uncensored subjects)

a. **Mean Absolute Error (MAE)**: the error between true times and predicted times, $|t_i - \hat{t}_i|$.
b. **Concordance Index** [1]: the ranking accuracy of all the order pairs.
c. **Integrated Brier Score** [2]: the probability accuracy over all time points (shaded areas).
d. **Log-Likelihood**: the magnitude of the predicted probability at event times.
e. **Hosmer-Lemeshow Calibration** [3]: if expected and observed event rates are similar over groups.
f. **Distribution Calibration** [4]: if subjects in each probability quantile bin are uniformly distributed.

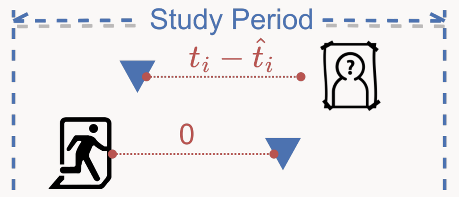


Handling Right-Censoring in MAE

How to apply MAE on censored subjects?

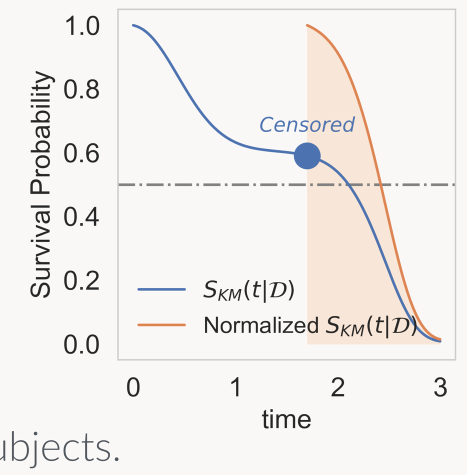
1. **Uncensored** simply excludes all censored subjects.
2. **Hinge** considers only the early prediction error.

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



3. **Margin** [4] assigns a surrogate value to each censored subject using the Kaplan-Meier estimator, $S_{\text{KM}(\mathcal{D})}(t)$.

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

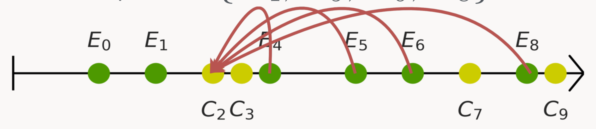


4. **IPCW-D** (proposed) uses inverse probability censoring weight $G(t_i)$, to uniformly transfer a censored subject's weights to relevant uncensored subjects.

$$\mathcal{R}_{\text{MAE-IPCW-D}}(\hat{t}_i, t_i, \delta_i) = \frac{|t_i - \hat{t}_i| \cdot \mathbb{1}_{\delta_i = 1}}{G(t_i)}$$

5. **IPCW-T** (proposed) uses the average over the times of all subsequent uncensored subjects as the surrogate time for the censored subject. (C_2 is distributed over the subsequent $\{E_4, E_5, E_6, E_8\}$)

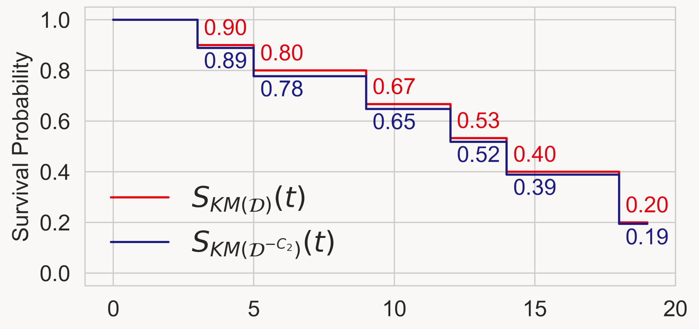
$$e_{\text{IPCW}}(t_i, \mathcal{D}) = \frac{\sum_{j \in \mathcal{D}} \mathbb{1}_{t_i < t_j} \cdot \mathbb{1}_{\delta_j = 1} \cdot t_j}{\sum_{j \in \mathcal{D}} \mathbb{1}_{t_i < t_j} \cdot \mathbb{1}_{\delta_j = 1}}$$



6. **PO** (proposed) uses pseudo-observations to estimate the surrogate event values.

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

Intuition: How much a censored subject counts towards the KM?



We apply a **weighting scheme**, for Margin, IPCW-T, and PO, to measure the trustworthiness of the surrogate values.

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

$\omega_i = 1 - S_{\text{KM}(\mathcal{D})}(t_i)$ for censored subjects, and $\omega_i = 1$ for uncensored subjects.

Theoretical Analysis

Why we prefer MAE?

- MAE is the most appropriate metric for quantifying the **time-to-event accuracy**.
- Time-to-event precision **cannot be covered** by other metrics.
- The model preference between MAE and other metrics might be **distinct**.

Is MAE proper?

$\Rightarrow$it is **a proper scoring rule** if we use **median survival time** of ISD as the predicted time.

(Definition) Proper scoring rule if $\mathbb{E}_{i \sim \mathcal{D}}\, \mathcal{R}(S_{\text{true}}(t \mid \boldsymbol{x}_i)), t_i, \delta_i) \le \mathbb{E}_{i \sim \mathcal{D}}\, \mathcal{R}(S_m(t \mid \boldsymbol{x}_i)), t_i, \delta_i)$

Authenticity of Pseudo-observation

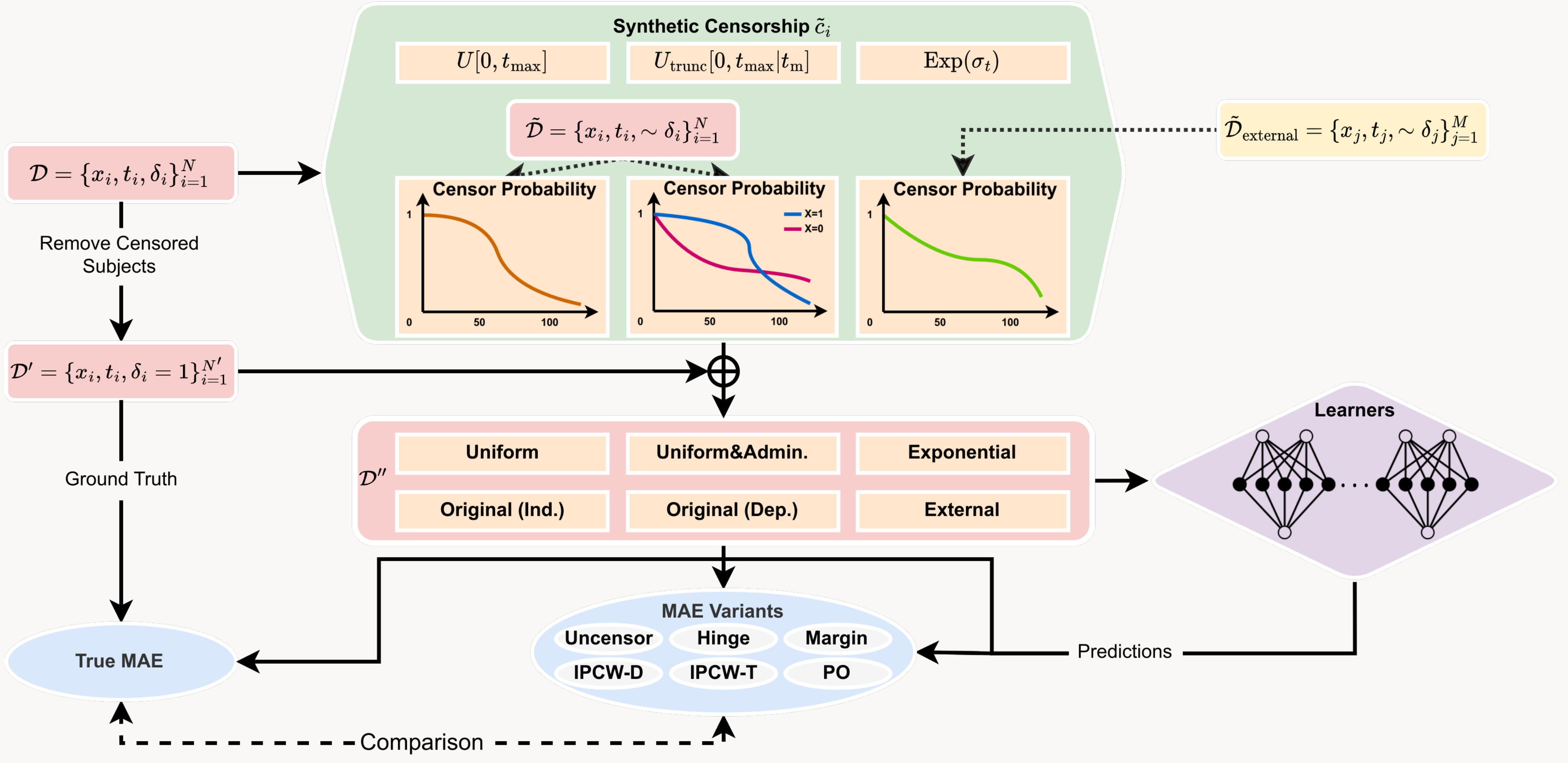
The pseudo-observation value for any censored instance is **lower bound** by its censoring time:

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

Evaluating the Evaluation Metrics

To evaluate the MAE-inspired evaluation metrics, we need to know the **true MAE**.

Not available in a real-world survival dataset? $\Rightarrow$ Produce a synthetic one.

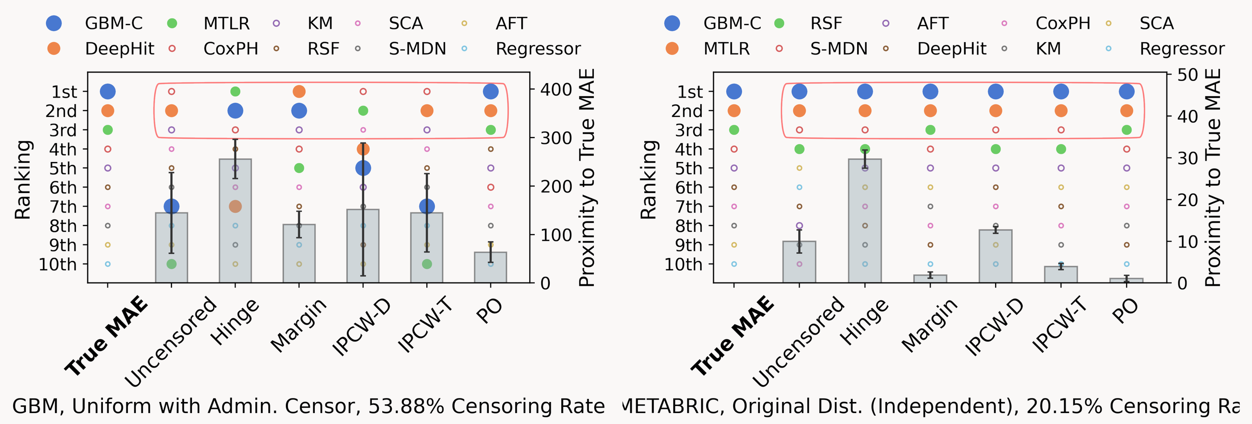


Property: **real-world covariates**, **close-to-reality event distribution**, **close-to-reality censor distribution**.

Empirical Performance

Desired MAE-variant should

- **accurately rank** the performance of models;
- generate performance score **closely approximate** the true MAE.



Summary of metric performance by counting the number of times each metric is best. *Includes ties.

	Uniform	Uniform&Admin.	Exponential	Original(Ind.)	Original(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*

References

[1] Harrell et al. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med

[2] Graf et al. Assessment and comparison of prognostic classification schemes for survival data. Stat Med

[3] Hosmer et al. Goodness of fit tests for the multiple logistic regression model. Commun. Stat. Theory Methods

[4] Haider et al. Effective ways to build and evaluate individual survival distributions. JMLR

Paper

Code

shiang@ualberta.ca

Proceedings of ICML-23