When Can We Trust Survival Model Evaluation?

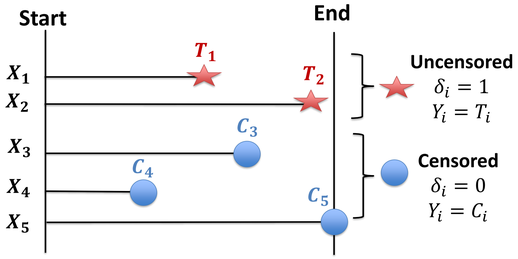
Ghanem Bahrini[1,2] , Sebastien Razakarivony[1], Jean-François Dupuy[2] ,Valerie Gares[3], Morgane Barbet-Massin[1]

[1]Safran Tech, Safran Group, France, [2] INSA Rennes, France, [3] INRIA Rennes, France

"When censoring hides the outcome, evaluation can hide the truth."

1. Survival Analysis

In survival data, some event times are observed, while others are hidden by right-censoring.



When event times are censored, model evaluation itself becomes uncertain.

In practice, survival models are evaluated using censored outcomes:

$$Y_i = \min(T_i, C_i), \quad \delta_i = 1(T_i \leq C_i)$$

2. Core Question

Can standard survival metrics still be trusted when censoring conditions vary?

We study whether censoring distorts:

Metric values & Model rankings

C-index, **IBS**, and **calibration** scores handle censored data, but their reliability still depends on : **Censoring rate & censoring mechanism**

Study Design Overview

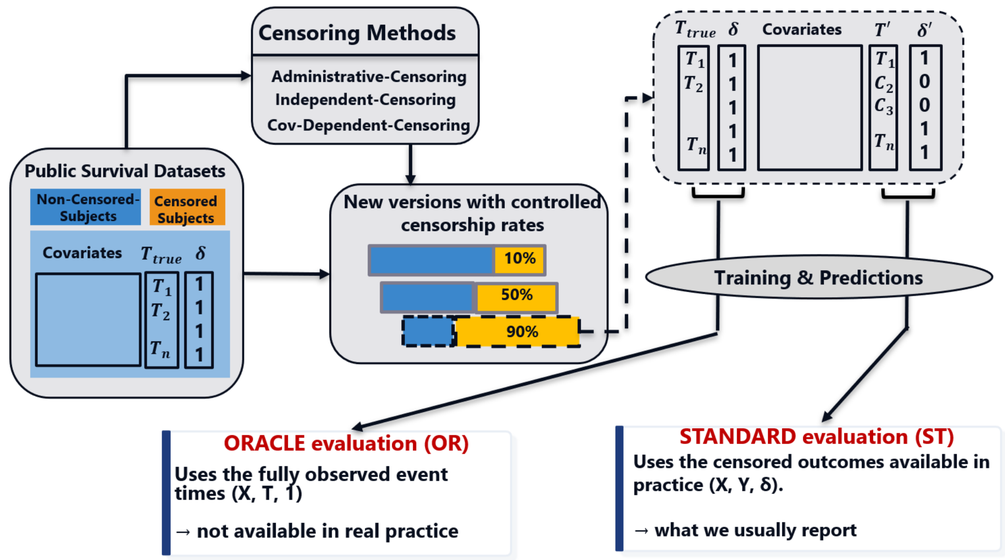
- **5 public datasets :** NACD · GBMLGG · METABRIC · PBC · FLCHAIN
- **3 censoring mechanisms :** AC / IC / CDC
- **5 censoring rates :** 10% · 30% · 50% · 70% · 90%
- **7 survival models :** CoxPH · Weibull AFT · RSF · DeepSurv · DeepHit · NMTLR · DCM

$d : dataset \mid c : censoring\ mecanism \mid \rho : censoring\ rate$

Code & Paper

ghanem.bahrini@safrangroup.com
ghanem.bahrini@insa-rennes.fr

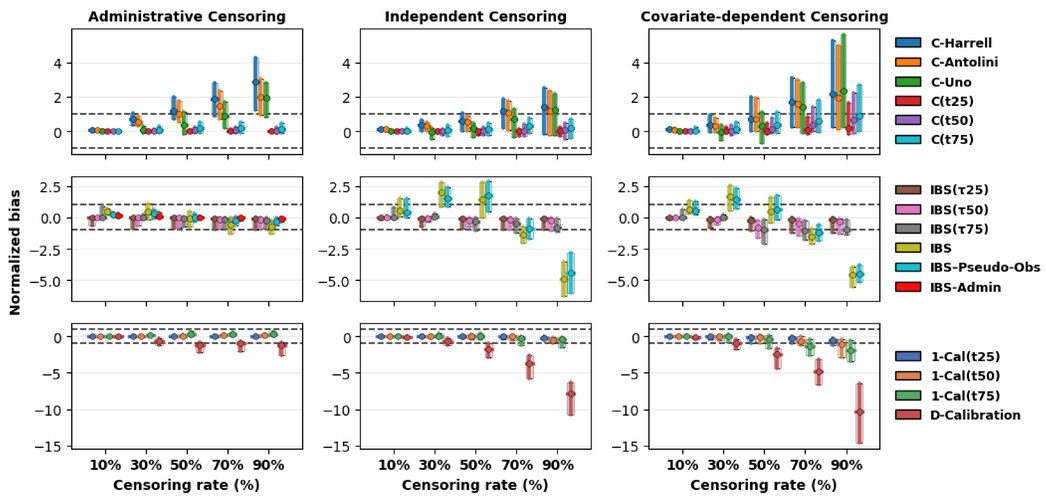
3. Key Idea



4. Numerical Metric Bias

$$B_r(d, c, \rho) = \frac{m_r^{ST}(d, c, \rho) - m_r^{OR}(d, c, \rho)}{IQR\left(m^{OR}(d, c, \rho)\right) + \epsilon}$$

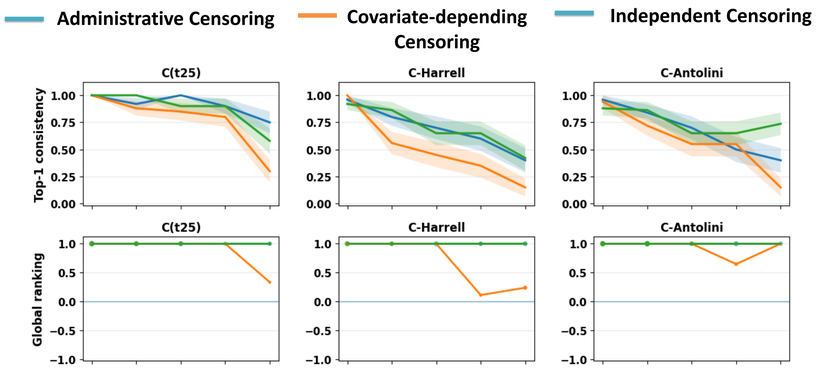
$|B_r| > 1 \Rightarrow ST - OR$ gap exceeds natural oracle variability



1. **Censoring rate** is the main driver of metric **distortion**.
2. **The censoring mechanism** shapes how metrics fail.
3. **Global C-index metrics** tend to **overestimate** discrimination.
4. **Fixed-horizon** concordance is more robust.
5. IBS benefits from **horizon truncation** under **independent censoring.**
6. Calibration can look **artificially good** under censoring.

5. Ranking Preservation

- **Top-1 consistency :** does ST select the same best model as OR?
- **Global agreement :** do statistically significant pairwise dominances agree ?



1. **Top-1 model selection becomes unstable** as censoring increases.
2. **A different "best model" does not always mean a reliable ranking change**: leading models are often not statistically separable.
3. **Global agreement is more robust** because it compares only statistically significant model pairs.
4. **C-index metrics often preserve global ranking under AC/IC**, despite Top-1 volatility.
5. **IBS and D-Calibration show stronger ranking failures**, especially under IC and CDC.

6. Recommendations

- Always report censoring rate and mechanism.
- Avoid Top-1-only model selection.
- Use statistical tests for model comparisons.
- Prefer early-horizon / truncated metrics when late-time evaluation is unreliable.
- Be especially cautious under covariate-dependent censoring.

7. Perspectives

- Making calibration metrics more robust to censoring
  Goal : reduce artificial calibration effects under heavy and covariate-dependent censoring
- Extend the framework to **competing risks** settings.