XGBoost-Based vs. AFT Model Imputation: Addressing Interval Censoring in Time-to-Event Data

Gustavo Soutinho and Luís Meira-Machado
Portucalense University

23rd International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2025)
September 20, 2025

What is Survival Analysis

- A branch of statistics that studies the time until an event happens.
- The event may be failure, death, relapse, etc.
- Widely applied in medicine, engineering, economics, and social sciences to analyze time-to-event data.

Mortality model

Mortality model for survival analysis.

Let T denote the survival times and C a univariate right-censoring which we assume to be independent of T.

Because of censoring we only observe $(\widetilde{T}, \Delta)$ where $\widetilde{T} = \min(T, C)$, $\Delta = I(T \leq C)$.

Survival Function $S(t)$

The survival function, $S(t)$, is defined as the probability that an individual or object survives beyond a given time t:

$$S(t) = P(T > t)$$

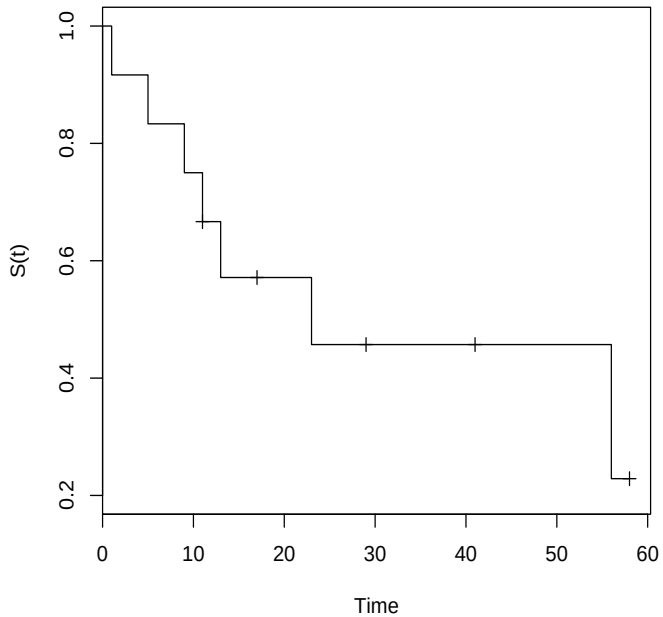
where T represents the time to the event of interest.

- This function can be empirically estimated using Kaplan-Meier estimator for right-censored data.
- This occurs when we don't know the exact time of an event, only that it happened after a certain observed time.

Kaplan-Meier estimator

$S(T > y)$ may be consistently estimated by the Kaplan-Meier estimator (Kaplan and Meier, 1958):

$$\widehat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \equiv \prod_{i=1}^{n} \left(1 - \frac{\Delta_{[i]}}{n - i + 1}\right)^{I(\widetilde{T}_{(i)} \leq t)}$$



time: 1, 5, 9, 11, 11, 13, 17, 23, 29, 41, 56, 58
event:
1, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0

Kaplan-Meier survival curve

Cox proportional hazard model

The Cox proportional hazard model (Cox, 1972) can be defined as follows

$$\lambda(t, x_i) = \lambda_0(t) \exp\Big(\sum_{j=1}^{q} \beta_j x_{ij}\Big) \tag{1}$$

where λ_0 are the baseline hazard functions for each transition, $\beta = (\beta_1, \cdots, \beta_q)'$ is the vector of unknown coefficients and $X = (x_1, \cdots, x_q)'$ is the vector of the covariates.

- Traditionally, $\lambda_0(t)$ remains unspecified and the coefficients β are obtained through partial likelihood (Cox (1975)) without regarding to the baseline hazard function.
- It is also possible to consider parametric models of the baseline hazard from standard survival distributions such as exponential, weibull or gamma.

Interval-Censored Data

The exact event time is unknown:

- Only known to happen between two times: (L_i, R_i)
- Kaplan-Meier estimator and traditional regression models are not suitable.
- They can lead to biased survival estimates and inaccurate hazard ratios.
- The Turnbull estimator offers a nonparametric approach especially well suited for interval-censored data, including scenarios that also involve right-censoring.
- Survival estimates are obtained from an iterative algorithm.

Imputation methods for survival estimation

Common methods are:

- left imputation: setting the event time to the start of the interval (L_i)
- right imputation: setting the event time to the end of the interval (R_i)
- Midpoint imputation: setting the event time to the middle of the interval ($(L_i + R_i)/2$)
- Making use of auxiliary information from covariates through Cox or Accelerated Failure Times (AFT) models.

XGBoost-based imputation method

- XGBoost-based imputation leverages the XGBoost algorithm, a powerful and efficient gradient boosting method.
- This builds an ensemble of decision trees to model the relationships between variables, allowing it to capture complex patterns and dependencies in the data.
- It provides accurate and robust imputations, even in the presence of missing values, by exploiting both linear and non-linear relationships.

The Scaled Linear Redistribution Method

- One of the limitations of these methods is that the predicted times may fall outside the range of the observed values.
- To address this, we propose a new imputation method-the Scaled Linear Redistribution Method-designed to ensure that imputed values for interval-censored data remain within their respective bounds while preserving the natural variability in the data.
- This composite strategy combines stochastic imputation with predictive modeling, while introducing a novel rescaling step that enforces consistency with censoring intervals and yields statistically coherent imputations.

Step-by-Step Procedure

1. **Initial Imputation:**
   For each interval-censored case, draw $T_i^{(0)} \sim U[L_i, R_i]$.
2. **Model Prediction:**
   Fit an AFT model or Machine Learning model (e.g., XGBoost).
   Predict survival times (may fall outside $[L_i, R_i]$).
3. **Repeat M Times:**
   Generate M sets of predicted times via simulation and re-fitting.
   Builds a distribution of predictions for each individual.
4. **Rescale Predicted Times:**
   Transform predictions to fit inside $[L_i, R_i]$ while preserving shape:

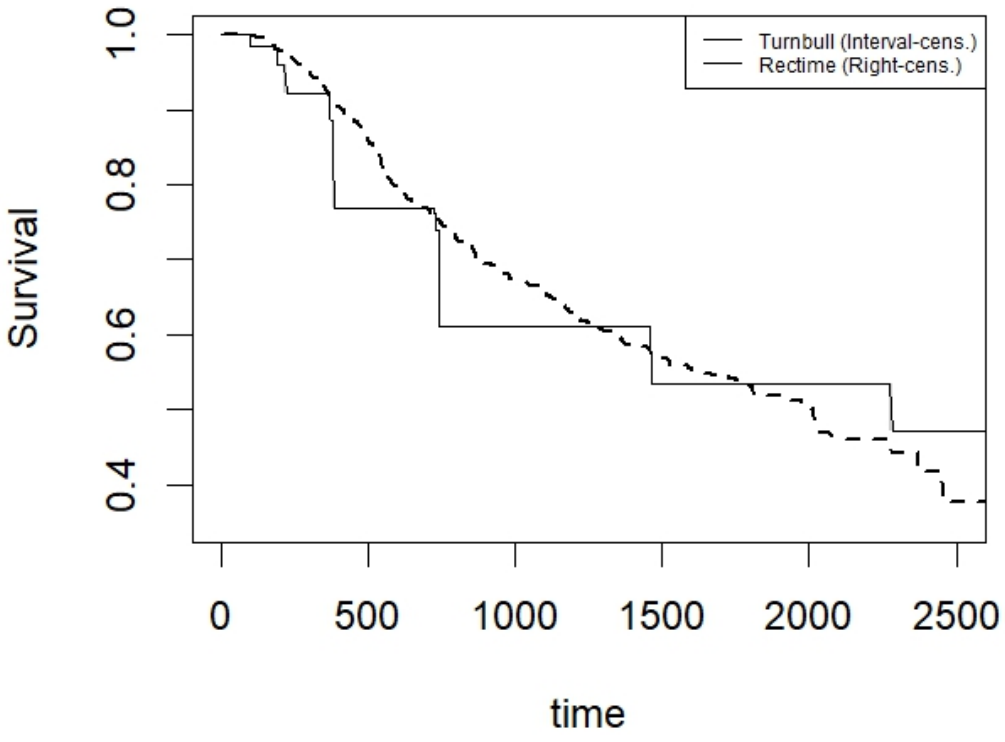
$$T_i^{\text{scaled}} = L_i + \frac{\hat{T}_i - \min(\hat{T}_i)}{\max(\hat{T}_i) - \min(\hat{T}_i)} (R_i - L_i)$$

5. **Final Imputation & Uncertainty:**
   Final imputed time: median of scaled predictions.
   Estimate uncertainty via standard deviation (SD) and standard error (SE) of M predictions.

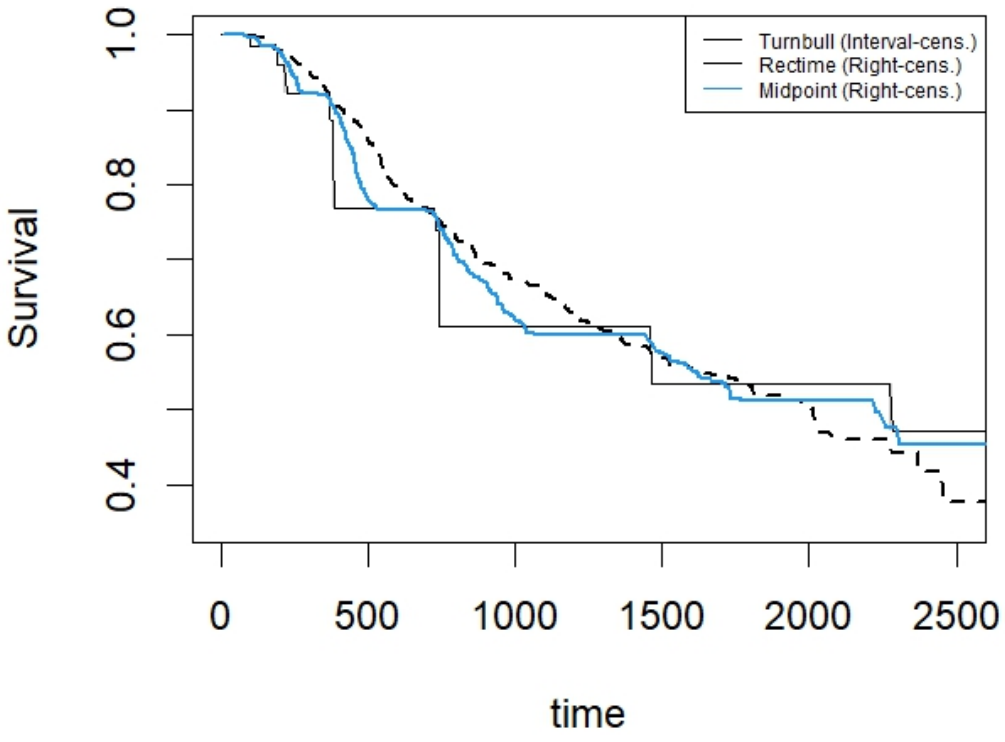
Case study: German Breast Cancer Study Group

- Study Overview: Prospective study (1984-1989) of 686 women with operable primary breast cancer, analyzing recurrence and survival.
- Event of Interest: Time to recurrence (299 experienced recurrence).
- Covariates: Age, menopausal status, tumor size, grade, lymph nodes, hormone therapy, ER/PR levels.
- Censoring Strategy: Synthetic follow-up times created to simulate interval censoring.
- Purpose: Estimate time to recurrence using various imputation methods.

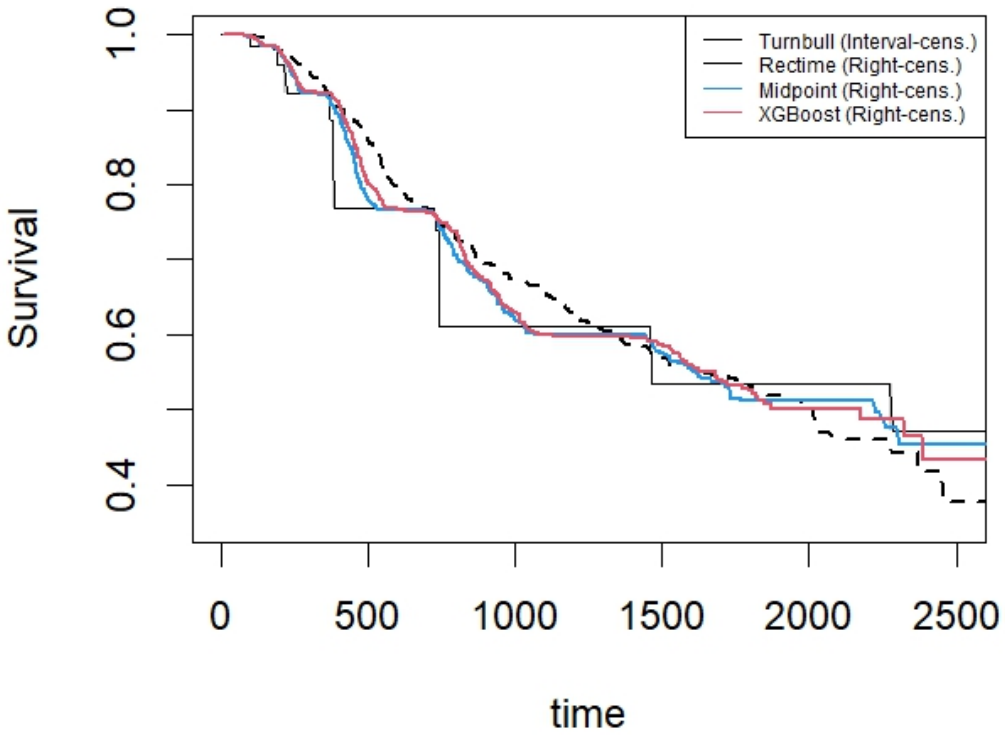
Survival curves - Comparison of results



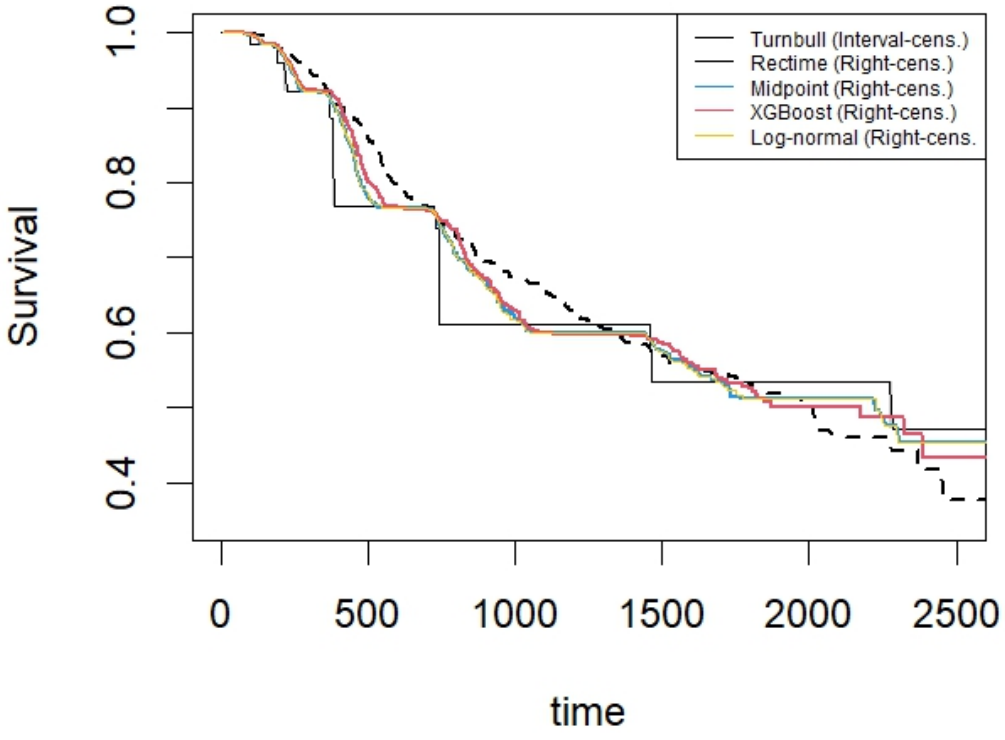
Survival curves - Comparison of results



Survival curves - Comparison of results



Survival curves - Comparison of results



Conclusions and Future Research

- Interval-censored data are challenging in survival analysis because event times are only known to lie within intervals.
- This study introduces a new approach to imputation of times to events in interval censoring data.
- Using synthetic data with known event times, we compare the new approach with classical methods such as midpoint and machine learning.
- Turnbull is conservative, but midpoint, XGBoost, and log-normal give similar survival curves, suggesting reasonable right-censoring approximations.
- As a future work we intend to assess the performance of the imputation methods using simulation studies.

- B. W. Turnbull, J. Roy. Stat. Soc Series B, **38**, 290–295 (1976).
- D. R. Cox, Journal of the Royal Statistical Society Series B (Methodological) **34**, 187–202 (1972)
- J. W. Bartlett, R. Keogh, E. F. Bonneville, and C. T. Ekstrom, *smcfcs: Substantive Model Compatible Fully Conditional Specification*, R package (2024).
- G. Gomes, S. R. Giolo, and E. A. Colosimo, Statistical Modelling **9**, 269–287 (2009).
- P. Wang, Y. Li, and C. K. Reddy, ACM Computing Surveys **51**, 1–36 (2019).
- H. Kvamme, and O. Borgan, Lifetime Data Anal. **27**, 710–736 (2021).
- A. Barnwal, H. Cho, and T. Hocking, Journal of Computational and Graphical Statistics **31**, 1292–1302 (2022)
- Y. Deng, and T. Lumley, Journal of Computational and Graphical Statistics **33**, 352–363 (2023).
- Z. Jinbo, L. Yufu, and M. Haitao, Front Artif Intell **8** (2025)

Thank you for your attention!