AIMS Mathematics, 11(3): 5719–5737.
DOI: 10.3934/math.2026235
Received: 05 November 2025
Revised: 31 January 2026
Accepted: 11 February 2026
Published: 06 March 2026

https://www.aimspress.com/journal/Math

Research article

Imputation strategies for interval-censored data: from AFT models to machine learning and scaled redistribution

Gustavo Soutinho[1,*] **and Luís Meira-Machado**[2]

[1] Department of Science and Technology, Portucalense University, R. Dr. António Bernardino de Almeida 541, 4200-072 Porto, Portugal

[2] Centre of Mathematics, University of Minho, Campus de Azurém, Edifício 12, 4800-058 Guimarães, Portugal

* **Correspondence:** Email: gustavo.soutinho@upt.pt.

Abstract: Interval-censored data pose challenges in survival analysis because event times are only known to occur within observation intervals. Traditional strategies, such as midpoint imputation, often fail to capture the uncertainty inherent to this censoring. This study compares classical, model-based, and machine learning approaches for imputing interval-censored event times. Specifically, we evaluate (i) standard midpoint imputation, (ii) accelerated failure time (AFT) model–based imputation, (iii) a machine learning method using XGBoost, and (iv) a new scaled linear redistribution method that constrains model-based imputations within censoring bounds while preserving their relative variability. A comprehensive simulation study under varying levels of right censoring was carried out to assess bias, accuracy, and concordance. Three real datasets were then analyzed to illustrate the practical behavior of the imputation methods. Results show that the XGBoost-based imputation shows stable performance across the different censoring scenarios considered, yielding survival estimates close to those of the nonparametric Turnbull estimator. The midpoint method performs adequately when intervals are short or censoring is mild, whereas parametric models are more sensitive to distributional assumptions and may yield biased estimates under heavy censoring. Analyses of real data further revealed greater variability among parametric models under high right censoring and a flattening of survival curves when censoring occurs, mainly at long event times. The proposed scaled linear redistribution method provides a way to map model-based predictions back to their observed censoring intervals while retaining their relative dispersion. The methods considered display complementary strengths across censoring regimes, with no single approach uniformly dominating.

Keywords: interval-censored data; machine learning; XGBoost; imputation methods
Mathematics Subject Classification: 62N02, 68T05, 68T09, 62J99, 62R07

1. Introduction

Survival analysis provides a statistical framework for studying the time until an event of interest occurs, such as death, equipment failure, or disease onset. It is widely applied in medicine, engineering, and the social sciences, particularly in situations where event times are only partially observed, resulting in censored data [1–3].

Interval censoring arises when the exact event time is unknown but known to lie between two observation times. This form of censoring commonly appears in longitudinal studies, leading to incomplete observations that pose analytical challenges. Addressing interval censoring appropriately is essential to reduce bias and obtain reliable estimates of the survival function. Since failure times are not directly observed, classical estimators such as Kaplan–Meier require adaptations that may introduce bias [4, 5].

Several approaches have been proposed for estimating survival functions under interval censoring. A widely used nonparametric option is the Turnbull estimator [6], which extends the Kaplan–Meier method to accommodate various censoring patterns without strong assumptions. However, this estimator cannot incorporate covariates or estimate hazard ratios, limiting its use in regression contexts. Parametric and semiparametric models, such as adaptations of the Weibull and Cox models [7–9], provide alternative frameworks but may produce biased results when model assumptions are violated or the data structure is complex.

Imputation provides a convenient strategy for handling censored observations by replacing incomplete records with plausible event-time values. Common strategies include midpoint, lower- or upper-bound, and model-based imputations that use fitted distributions to predict event times. Multiple imputation, in particular, allows the use of covariate information to generate plausible event times. The `smcfcs` package in R [10] implements this approach, producing imputations compatible with the assumed survival model and helping maintain model validity in complex censoring settings [5].

Recent developments in machine learning have expanded the range of tools available for survival analysis, particularly for problems involving censored or partially observed outcomes [11, 12]. Among these methods, XGBoost has received considerable attention because of its flexibility and its strong empirical predictive performance in high-dimensional settings [13]. Its boosting structure allows nonlinear effects and complex interactions between covariates to be represented without imposing a specific parametric form [14]. In addition, the ability to handle heterogeneous data, incorporate regularization to control overfitting, and exploit parallel computation makes XGBoost a particularly appealing, data-driven alternative for imputing incomplete observations in large-scale and longitudinal survival studies. An efficient implementation is available through the `xgboost` package in R.

More broadly, a number of contributions in the machine learning literature have proposed ways of adapting standard supervised learning algorithms to censored survival outcomes through so-called "uncensoring" procedures. For example, [15] proposed a likelihood-based preprocessing approach that replaces right-censored observations by pseudo–event times, allowing the subsequent application of conventional learning algorithms. However, most existing uncensoring techniques are tailored to right-censored data and rely on estimated survival distributions, and they are not specifically designed for interval-censored settings in which the event time is only known to lie within inspection intervals. Moreover, systematic comparative evaluations of different uncensoring or imputation strategies under interval censoring are still relatively scarce. The present work contributes to this line of research by

presenting a unified comparison of classical, model-based, and machine-learning–driven imputation procedures for interval-censored data, and by introducing a redistribution mechanism that enforces coherence with the observed censoring bounds.

Related work has also investigated the integration of boosting ideas within classical survival models. In particular, [16, 17] proposed boosting-based accelerated failure time (AFT) models that increase flexibility, allow automatic variable selection, and relax strict parametric assumptions while retaining interpretability. Although these approaches provide a useful bridge between traditional AFT modeling and modern machine learning, their primary focus is on direct survival modeling rather than on the imputation of latent event times under interval censoring, which is the central concern of the present study.

Despite their usefulness, AFT-based imputations share an important limitation: The imputed event times may fall outside the observed censoring intervals, thereby violating the logical constraints of the data. Such inconsistencies can distort survival estimates and reduce interpretability. To address this, we propose the *scaled linear redistribution method*, a new approach that ensures imputed values remain within their censoring intervals while preserving their relative variability. By combining predictive modeling with a rescaling step, this method maintains coherence with censoring bounds and provides a statistically consistent alternative to existing imputation techniques.

The main goal of this paper is to compare several imputation-based strategies for regression modeling under interval censoring. We contrast traditional midpoint imputation with model-based approaches using AFT models and a machine learning method based on XGBoost. To maintain consistency with censoring intervals, imputations are further adjusted using the proposed scaled linear redistribution method. This unified adjustment allows for a fair comparison of the methods and offers practical insight into their relative performance in survival analysis. To this end, we conduct a simulation study under different censoring levels and complement it with applications to three real datasets, enabling a comprehensive evaluation of the proposed strategies.

The remainder of the paper is organized as follows. Sections 2 and 3 present the theoretical background for survival curve estimation and the imputation methods. Sections 4 and 5 report the simulation results and real-data applications. Finally, Section 6 summarizes the main conclusions and outlines directions for future research.

2. Estimation of survival

2.1. Notation

Let T denote the true but unobserved event time, known only to lie within an interval $(L, R]$, where L and R are the lower and upper bounds, respectively. If the event is known to occur after time L, the upper bound R is set to ∞, indicating right censoring. When the event time is observed exactly, the interval collapses to a single point, i.e., $L = R$.

The censoring indicator δ specifies the observation type: $\delta = 1$ for exact event times ($L = R$), $\delta = 2$ for interval censoring ($L < R$), and $\delta = 0$ for right censoring ($R = \infty$).

The estimated survival function, $\hat{S}(t)$, represents the probability that the event occurs after time t and serves as the main quantity for evaluating the performance of the imputation methods.

2.2. Kaplan-Meier estimator

The Kaplan–Meier estimator [18], also known as the product-limit estimator, is a well-known nonparametric method for estimating the survival function $S(t)$. It is particularly suited to right-censored data. Assume, therefore, that each event time is either exactly observed ($L = R$) or right-censored ($R = \infty$, $\delta = 0$).

Let $L_{(1)} < L_{(2)} < \cdots < L_{(k)}$ denote the ordered, distinct event times from a sample of n individuals. The Kaplan–Meier product-limit estimator of the survival function can be expressed using the Kaplan–Meier weights [19] as

$$\hat{S}(t) = 1 - \sum_{i=1}^{n} W_i \, I(L_{(i)} \le t),$$

where W_i is the weight associated with the ith ordered event time, defined as

$$W_i = \frac{\delta_{[i]}}{n - i + 1} \prod_{j=1}^{i-1} \left(1 - \frac{\delta_{[j]}}{n - j + 1}\right).$$

Although the Kaplan–Meier estimator is widely used, it cannot directly accommodate interval-censored data. In such cases, the Turnbull estimator provides a more general nonparametric alternative, as discussed in the next section.

2.3. Turnbull estimator

The Turnbull estimator [6] extends the Kaplan–Meier method to interval-censored data, where the event time T is known only to lie within an interval $(L_i, R_i]$. It is based on constructing disjoint *Turnbull intervals* $\{I_j\}_{j=1}^{m}$, obtained from the ordered set of unique censoring endpoints $\{\tau_0 < \tau_1 < \cdots < \tau_m\}$. Each interval $I_j = (\tau_{j-1}, \tau_j]$ represents a possible location of the event.

The probability mass assigned to interval I_j is

$$p_j = P(\tau_{j-1} < T \le \tau_j),$$

and the nonparametric likelihood for a sample of n subjects is

$$L(p_1, \ldots, p_m) = \prod_{i=1}^{n} \left(\sum_{j=1}^{m} \delta_{ij} \, p_j \right),$$

where $\delta_{ij} = 1$ if $I_j \subset (L_i, R_i]$, and $\delta_{ij} = 0$ otherwise.

Maximization is carried out using Turnbull's self-consistency algorithm, an iterative procedure that alternates between redistributing fractional counts across compatible intervals and updating the probabilities until convergence.

The resulting estimate of the survival function is

$$\hat{S}(t) = 1 - \sum_{j:\, \tau_j \le t} \hat{p}_j,$$

yielding a nonparametric maximum likelihood estimator that generalizes the Kaplan–Meier approach to interval-censored data.

Although the algorithm can be computationally demanding, particularly for large or complex datasets, its capacity to accommodate different types of censoring makes it a valuable complement to the Kaplan–Meier estimator, which is mainly designed for right-censored data. In this study, both approaches, the Kaplan–Meier estimator combined with imputation strategies and the Turnbull estimator, serve as baseline methods for comparing survival estimation techniques.

3. Imputation methods

3.1. Classical methods

Imputing interval-censored observations $(L_i, R_i]$, for $i = 1, \ldots, n$, where $L_i < R_i$ and R_i is finite, presents a particular challenge. Conventional approaches typically replace each censored interval with a single representative value. The most common options are: *left imputation*, which sets the event time to the start of the interval (L_i) and assumes that the event occurred as early as possible; *right imputation*, which sets the event time to the end of the interval (R_i) and assumes that it occurred as late as possible; and *midpoint imputation*, which uses the midpoint of the interval ($(L_i + R_i)/2$), assuming that the event time is uniformly distributed within the interval.

Although straightforward, these methods ignore the variability and uncertainty inherent in interval-censored data. They can also introduce systematic bias, particularly when event times deviate from a uniform distribution or are strongly influenced by covariates.

3.2. Accelerated failure time (AFT) model

Alternatively, the survival outcome can be treated as missing and predicted using the relationships between the covariates and the observed survival data. The interval-censored response is then handled as a missing value to be filled in by imputation, with the key requirement that the imputed time remains consistent with the observed interval $(L_i, R_i]$.

Standard survival models, such as AFT models, are widely used in survival analysis due to their flexibility and ease of interpretation. In this study, this type of model is also used to obtain imputed values based on auxiliary information from covariates.

The AFT model provides an alternative to proportional hazards models, such as the Cox model, by directly describing the effect of covariates on survival time [20]. Specifically, it models how predictors accelerate or decelerate the time until an event occurs, rather than modeling the hazard function. The model can be expressed as

$$\log(T) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon,$$

where T denotes the event time, $X_1, X_2, \ldots, X_p$ are covariates, β_i are coefficients representing the effect of the ith covariate on log-survival time, and ε is an error term following a specified distribution. In particular, a normal distribution for ε corresponds to a log-normal AFT model, whereas an extreme-value distribution gives rise to the Weibull AFT model, with the exponential model as a limiting case. These alternative specifications are considered here to evaluate the sensitivity of the imputation results to the assumed error distribution.

The interpretation of the coefficients is straightforward: positive values of β_i indicate that the covariate prolongs survival time, while negative values suggest an acceleration toward the event. In

practice, the AFT model offers a direct and interpretable framework for assessing how covariates influence event timing, making it a valuable tool for generating model-based imputations.

We adopt the AFT formulation instead of the Cox proportional hazards model because the main objective of this study is the imputation of latent event times, rather than inference on hazard ratios. The AFT model directly relates survival time to covariates, which is convenient for generating stochastic predictions within censoring intervals. In contrast, the Cox model is specified in terms of the hazard function and does not provide event-time predictions without further assumptions on the baseline hazard, making it less natural for the present imputation framework.

3.3. XGBoost-based imputation method

XGBoost is an efficient implementation of gradient tree boosting that has been widely used in regression and prediction problems involving complex and high-dimensional covariate spaces [21]. In the present context, it is employed as a flexible non-linear regression tool to model the relationship between covariates and survival times, without imposing proportional hazards assumptions or a specific parametric form.

The method builds an ensemble of regression trees in an additive manner. At boosting iteration t, a new tree is fitted to the current residuals by minimizing a regularized objective function of the form

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{K} \Omega(f_k),$$

where $l(\cdot)$ denotes a loss function, y_i is the target value, and $\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i)$ is the prediction for observation i at iteration t. Each f_k is a regression tree belonging to the function space $\mathcal{F}$, and the regularization term $\Omega(f_k)$ penalizes model complexity through the number of terminal nodes and the L_2 norm of the leaf weights.

Here, XGBoost is not used through its built-in mechanism for handling missing covariates via sparsity-aware split finding. Instead, the fitted ensemble is used as a global prediction model for the latent event times T_i corresponding to interval-censored observations. At each iteration of the imputation procedure, the model is refitted to the current set of pseudo–event times, yielding updated predictions that reflect the dependence structure between covariates and survival outcomes.

Although computationally more demanding than simple deterministic imputations, this approach can accommodate nonlinear effects and high-order interactions among covariates. Its performance depends on appropriate tuning of hyperparameters, such as the learning rate and tree depth, which is carried out in the simulation and data analyses.

Deep neural networks could in principle also be considered for this task. However, they typically require larger sample sizes, more complex model specification, and substantially higher computational cost. Given the tabular structure of the data and the sample sizes considered in this study, XGBoost provides a suitable compromise between flexibility, stability, and interpretability. Extensions to deep learning–based imputation strategies are left for future work.

3.4. The scaled linear redistribution method

A key difficulty in the imputation of interval-censored data is that model-based predictions of event times may fall outside the observed censoring bounds. To overcome this issue, we introduce the *scaled*

linear redistribution method, an imputation procedure that enforces compatibility between predicted event times and the observed censoring intervals, while preserving the relative dispersion induced by the predictive model.

The proposed method proceeds as follows:

(1) For each individual with interval-censored data, an initial survival time $T_i^{(0)}$ is generated by sampling from a uniform distribution over the observed interval:

$$T_i^{(0)} \sim U(L_i, R_i).$$

(2) A predictive model, either a parametric AFT model or a machine learning model such as XGBoost, is then fitted using these initially imputed times. The fitted model is used to obtain updated predictions $\hat{T}_i$ for all interval-censored observations.

(3) Steps 1 and 2 are iterated M times (with $M = 1000$ in this study). At iteration m, the model is refitted using the imputed values from the previous step, generating a new set of predictions. This procedure yields, for each individual, a collection of predicted event times $\{\hat{T}_i^{(1)}, \ldots, \hat{T}_i^{(M)}\}$ that reflects the stochastic variability induced by both the censoring mechanism and the predictive model.

(4) Since the predicted values $\hat{T}_i^{(m)}$ are not guaranteed to lie within the corresponding censoring interval $(L_i, R_i]$, a linear rescaling is applied. For each subject i, the rescaled values are defined by

$$T_{i,m}^{\text{scaled}} = L_i + \frac{\hat{T}_i^{(m)} - \min_{1\le m\le M} \hat{T}_i^{(m)}}{\max_{1\le m\le M} \hat{T}_i^{(m)} - \min_{1\le m\le M} \hat{T}_i^{(m)}}(R_i - L_i),$$

which maps the empirical range of the predicted values onto the interval $(L_i, R_i]$ while preserving their relative ordering and dispersion.

(5) The final imputed event time for subject i is taken as a robust summary of the rescaled distribution:

$$T_i^{\text{final}} = \text{median}\left(T_{i,1}^{\text{scaled}}, \ldots, T_{i,M}^{\text{scaled}}\right).$$

The median is preferred to the mean in order to reduce the influence of extreme values, which may arise from the iterative stochastic imputation process.

To quantify the uncertainty associated with the imputation, the variability of the rescaled samples is further summarized through their standard deviation (SD) and standard error (SE).

The scaled linear redistribution method preserves the variability of the imputed distribution while constraining values to lie strictly within their censoring intervals $(L, R]$. By integrating covariate information and survival modeling into the imputation step, the method produces coherent and data-driven imputations. Unlike midpoint or boundary-based strategies, which may distort survival estimates, this approach maintains stochasticity, respects censoring information, and yields more realistic and robust imputations.

4. Simulation studies

This section describes the simulation framework used to evaluate the performance of different imputation methods under interval censoring. The procedure follows a sequential design that begins

with the generation of realistic covariates, proceeds with the simulation of true survival times under a log-normal AFT model, and finally introduces censoring mechanisms through predefined inspection windows. This setup provides a controlled environment to assess and compare the accuracy of the proposed imputation strategies. An important feature of the simulation design is the inclusion of a multivariate covariate structure, which allows us to examine how the different imputation methods use auxiliary information when recovering latent event times under interval censoring.

Step 1: Simulating covariates. To reproduce the characteristics of a realistic clinical dataset, we generated covariates resembling those from the GBCSG trial. These included: age (normally distributed, mean 55 and standard deviation 10), menopausal status (Bernoulli with probability 0.6), tumor size in millimeters (normally distributed, mean 30 and standard deviation 15), histologic grade (categorical with three levels), number of positive lymph nodes (Poisson with mean 3), and hormone therapy status (Bernoulli with probability 0.5). For histologic grade, dummy variables were created to represent grades 2 and 3.

Step 2: Generating true survival times. True event times were simulated under an AFT model with a log-normal baseline. A linear predictor was first defined as a weighted combination of the covariates, with effect sizes chosen to reflect plausible clinical relationships. The log-survival times were then generated as

$$\log T = \eta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2),$$

where the linear predictor η is defined through a standard regression structure:

$$\eta = \beta_0 + \sum_j \beta_j X_{ij},$$

with X_{ij} denoting the value of the jth covariate for individual i (such as age, menopausal status, and other clinical variables), and β_j the corresponding regression coefficients. This formulation ensures that the simulated survival times depend directly on covariate profiles, thereby reflecting realistic regression-based survival processes commonly encountered in applied settings.

Exponentiating $\log T$ produced the true survival times T_{true}. This framework enables covariates to influence survival outcomes in a regression-consistent manner while maintaining stochastic variability across individuals.

Step 3: Defining follow-up windows. In practice, event times are rarely observed exactly but are known to occur between consecutive follow-up visits. To mimic this, we generated individual-specific visit schedules rather than adopting a fixed design. The number of follow-up visits was determined adaptively based on each subject's risk profile, obtained from the AFT model's linear predictor. Higher-risk individuals were assigned more frequent visits. In this setup, the number of visits ranged from two to three per subject.

Step 4: Introducing censoring mechanisms. Follow-up visit times were sampled from a uniform distribution on $[0, c \cdot T_i]$, where T_i represents the true event time and c is a scaling factor controlling the level of right censoring. Based on these visits, the observed interval $(L_i, R_i]$ was defined for each subject. Two censoring mechanisms were then identified:

- Interval censoring: If T_i occurred between two visits, the event was recorded as $(L_i, R_i]$;
- Right censoring: If T_i exceeded the last visit time, the subject was considered right-censored with $R_i = \infty$ and L_i equal to the last visit.

4.1. Simulation scenarios

To examine the impact of different censoring intensities, we varied the scaling factor c, which directly affects the probability of right censoring. Four scenarios were defined:

- Scenario I: $c = 30$, resulting in virtually no right censoring.
- Scenario II: $c = 3.5$, producing approximately 8% right censoring.
- Scenario III: $c = 2.6$, producing approximately 15% right censoring.
- Scenario IV: $c = 1.7$, producing approximately 32% right censoring.

These scenarios were constructed to represent increasing levels of interval and right censoring under irregular follow-up schemes. The scaling factor c was chosen so as to generate a gradual increase in the proportion of right-censored observations, from virtually no censoring to approximately 32%. This range spans settings with nearly complete observation to situations with substantial information loss, allowing an evaluation of the sensitivity of the imputation methods to different censoring intensities.

The scaling factor c was calibrated to generate right-censoring levels ranging from 0% to 32%. This range was selected to represent a spectrum from ideal observation conditions to realistic clinical settings with significant information loss, allowing for a comprehensive evaluation of each method's robustness under varying degrees of data sparsity.

4.2. Evaluation metrics for imputation times

To evaluate the performance of the imputation methods, we considered several complementary metrics that capture different dimensions of accuracy, variability, and discriminative ability. Let T_i denote the true event time for individual i and $\hat{T}_i$ its corresponding imputed value. For each simulated dataset, the evaluation was performed across n individuals and repeated over R replications.

The *bias* quantifies the systematic deviation of the imputed times from the true event times, whereas the *root mean squared error* (RMSE) combines both bias and variance into a single measure of accuracy:

$$\widehat{\text{Bias}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{T}_i - T_i), \qquad \widehat{\text{RMSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{T}_i - T_i)^2}.$$

As a robust alternative less sensitive to outliers, we also computed the *median absolute deviation* (MAD) between true and imputed times:

$$\text{MAD} = \text{median}\left(|\hat{T}_i - T_i|\right).$$

Finally, we assessed the discriminative power of the imputations using the concordance index (C-index), which measures the proportion of correctly ordered pairs of individuals based on their imputed and true event times. This metric is widely used in survival analysis to quantify predictive concordance.

Together, these measures provide a comprehensive evaluation of performance, encompassing absolute accuracy (Bias, RMSE), robustness (MAD), and ranking ability (C-index). Results were summarized across 1000 replications to assess the stability and reliability of each imputation method.

4.3. Main results

Table 1 summarizes the results obtained from the simulated datasets. It reports the performance of the different imputation methods – midpoint, XGBoost, log-normal, Weibull, exponential, Gaussian, and the unscaled XGBoost approach (i.e., without the scaled linear redistribution step) – across the four censoring scenarios (I–IV). The first two columns identify the simulation scenario and the evaluation metric used to assess imputation quality: bias, RMSE, MAD, and the C-index. These measures capture complementary aspects of performance, reflecting both the accuracy and the rank consistency of the imputed event times. Together, they enable a comprehensive comparison of the methods under varying degrees of right censoring.

Table 1. Comparison of imputed event times across the midpoint, XGBoost, log-normal, Weibull, exponential, Gaussian, and the unscaled XGBoost method (i.e., without the scaled linear redistribution step).

Scenario	Metric	Midpoint	XGBoost	Log-normal	Weibull	Exponential	Gaussian	XGBoost (unscaled)
I	Bias	2109.171	1344.481	1872.829	1871.917	1846.230	2211.737	1315.367
	RMSE	8244.104	4939.985	7353.937	7639.078	7632.824	8784.523	5113.214
	MAD	372.204	258.595	321.956	317.253	303.306	378.345	280.932
	C-Index	0.861	0.858	0.861	0.860	0.861	0.860	0.803
II	Bias	192.426	160.84	178.921	196.123	182.660	178.213	234.521
	RMSE	1209.229	841.511	1060.554	1123.300	1121.027	973.082	1075.073
	MAD	30.367	28.532	29.713	29.551	30.337	30.703	42.355
	C-Index	0.920	0.920	0.919	0.924	0.924	0.916	0.855
III	Bias	125.876	126.524	128.556	126.372	133.501	124.98	187.159
	RMSE	534.855	378.606	525.033	547.658	549.347	521.781	805.635
	MAD	23.715	25.102	25.077	23.597	25.932	24.121	37.061
	C-Index	0.927	0.927	0.925	0.925	0.924	0.926	0.867
IV	Bias	132.041	154.235	132.615	133.870	140.542	127.149	188.593
	RMSE	516.000	660.621	501.178	574.118	575.710	491.355	936.918
	MAD	21.113	24.709	21.538	22.337	23.569	20.703	27.121
	C-Index	0.928	0.928	0.927	0.928	0.927	0.927	0.887

Scenario I: In the absence of censoring, the imputation methods exhibited generally similar performance, although some differences emerged in bias and overall accuracy. The XGBoost-based approach achieved the smallest bias (1344.5) and lowest overall error (RMSE = 4940), followed by the log-normal model (bias = 1872.8, RMSE $\approx$ 7353.9). The midpoint method showed higher error levels (bias = 2109.2, RMSE $\approx$ 8244.1), while the Gaussian model displayed the poorest accuracy across all metrics (bias = 2211.7, RMSE $\approx$ 8784.5). In this setting without censoring, the standard (unscaled) XGBoost model attains a slightly smaller bias (1315.4) than the scaled version. However, its RMSE and MAD are larger, indicating poorer overall accuracy and illustrating the advantage of enforcing the redistribution step even in the absence of censoring. The MAD confirmed this ranking, with XGBoost producing the most precise imputations (MAD = 258.6) and the Gaussian model exhibiting the greatest dispersion (MAD = 378.3).

Scenario II: With the introduction of low right censoring (approximately 8%), all imputation methods showed a marked reduction in both bias and overall error compared with the fully observed case. This improvement likely results from the censoring constraint, which limits the range of possible event times and reduces extreme deviations between imputed and true values. The XGBoost-based approach again produced the smallest bias (160.8) and lowest overall error (RMSE $\approx$ 841.5),

confirming its numerical accuracy and stability under moderate information loss. The log-normal and Gaussian models performed comparably well (bias $\approx$ 179, RMSE $\approx$ 973–1061), whereas the midpoint and Weibull imputations were slightly less accurate (bias $\approx$ 192–196, RMSE $\approx$ 1121–1209). The MAD remained low for all methods, indicating limited imputation variability at the individual level. The XGBoost-based method again achieved the smallest MAD (28.5), corresponding to the highest precision among the considered imputation approaches.

Scenario III: With a censoring level of approximately 15%, the bias further decreased across all methods, resulting in greater convergence of performance. The reduction in systematic error narrowed the differences among models, with the Gaussian method now emerging as one of the best-performing options. The Weibull and midpoint imputations also improved noticeably, surpassing the XGBoost and log-normal models that had previously shown the lowest errors. Although overall differences were small, the exponential model displayed slightly higher bias than the others. The XGBoost-based approach remained competitive, with a bias of 126.524 and the smallest overall error (RMSE $\approx$ 378.606), closely followed by the Gaussian and log-normal models (bias $\approx$ 125–129, RMSE $\approx$ 520). The MAD stayed low and nearly identical across methods (23–26 units), indicating that individual-level imputation variability remained minimal even under moderate censoring.

Scenario IV: Under the highest level of right censoring (approximately 32%), all imputation methods showed a slight increase in bias. Among the approaches, the Gaussian model achieved the lowest bias and RMSE, confirming its stability under heavy censoring. The midpoint, log-normal, and Weibull imputations performed similarly, with the log-normal model yielding the best RMSE among them, likely due to its parametric form, which better accommodates the restricted range of observed event times. The XGBoost-based method presented the largest RMSE, indicating a modest decline in predictive accuracy as censoring intensified, consistent with the trend observed in Scenario III. Despite this, the MAD remained low and similar across methods (20–24 units), suggesting that individual-level imputation variability was largely unaffected by the higher degree of censoring.

Finally, the concordance index (C-index) values were similar across methods within each scenario. For the approaches incorporating the scaled linear redistribution step, the C-index increased from approximately 0.86 in Scenario I to about 0.93 in Scenario IV. In contrast, the unscaled XGBoost method consistently produced lower concordance values, ranging from roughly 0.80 to 0.89. The increase in the C-index with higher censoring levels reflects a stabilization of the ordering of event times, as heavier censoring reduces the effective variability of the survival distribution. This behaviour should therefore be interpreted mainly as a consequence of reduced dispersion rather than as an intrinsic improvement in predictive discrimination.

Overall comparison across scenarios I–IV: The absolute accuracy measures, namely bias, RMSE and MAD, showed a consistent pattern across the four scenarios. Bias was highest in Scenario I (no right censoring), decreased sharply with the introduction of moderate censoring in Scenario II (approximately 8%), continued to decline at 15% censoring (Scenario III), and increased slightly under the highest censoring level (32%, Scenario IV).

This evolution reflects the combined influence of several mechanisms. As censoring increases, the effective range of observed survival times shortens, which naturally reduces the potential deviation between true and imputed values. The presence of censoring also imposes a truncation on the imputation process, since imputed times are restricted to narrower intervals. This acts as a form of regularization that limits large positive errors. In addition, heavier censoring compresses the

distribution of observed data, leading to smaller average bias simply because the overall scale of survival times decreases. Because bias represents the mean deviation across individuals, a shorter time support will naturally produce smaller absolute values even if the relative estimation accuracy remains unchanged. Therefore, the observed decline in bias with increasing censoring levels should not be seen as a genuine improvement in imputation quality but rather as a by-product of these distributional and mechanical effects inherent to censored data.

Across all scenarios, for the methods incorporating the scaled linear redistribution step, the parametric models based on Gaussian and log-normal assumptions generally achieved the lowest RMSE and bias under moderate to high censoring, demonstrating good robustness and accuracy. In contrast, in the absence of censoring (Scenario I), the Gaussian model showed higher deviations, performing slightly worse than XGBoost and other parametric alternatives. The XGBoost-based approach remained competitive under low to moderate censoring but exhibited some loss in predictive accuracy under heavy censoring. The midpoint method, although simple and consistent, tended to produce larger deviations from the true event times, particularly in settings with little or no censoring, reflecting its limited ability to adapt to varying interval widths.

A comparison with the standard, unscaled XGBoost model shows that unconstrained predictions are associated with larger bias and RMSE, especially as the proportion of right-censored observations increases. Although differences are small in the fully observed setting, they become more pronounced under moderate and high censoring. This confirms the role of the scaled linear redistribution step in enforcing compatibility with the censoring intervals and in improving overall imputation accuracy.

For all methods, the MAD remained low and stable, suggesting limited individual-level imputation variability. Overall, all approaches preserved the relative ordering of event times effectively, as shown by consistently high C-index values, while model-based and machine learning methods provided clear advantages in absolute accuracy and robustness under most censoring conditions.

5. Application to real datasets

To illustrate the practical use of the proposed methodology, we applied it to three real datasets. The aim was to evaluate how different imputation methods influence the estimated survival curves. We began with the original interval-censored data, defined by lower and upper bounds denoted by `left` and `right`, and estimated the nonparametric survival function using the Turnbull estimator, which served as a reference for comparison. We then compared these results with the survival functions obtained from the imputed event times using the nonparametric Kaplan–Meier estimator. For each imputation approach—midpoint, XGBoost-based, log-normal, Weibull, exponential, and Gaussian—we used the corresponding imputed times together with the event indicator `status` to compute Kaplan–Meier survival curves via the `survfit()` function in R. This procedure enabled a direct comparison between the original interval-censored survival distribution, represented by the Turnbull estimator, and the survival curves derived from each imputation method. By examining these curves, we assessed how well each approach reproduced the underlying survival pattern and whether systematic deviations arose from the imputation process.

IR_diabetes dataset. The first dataset, `IR_diabetes`, available in the `icenReg` package, contains interval-censored survival times for 731 patients with type 1 diabetes, measuring the period from disease onset to the development of diabetic nephropathy [22]. The exact onset time is unknown;

instead, patients were examined at clinical visits, and the event was known to have occurred between two consecutive visits. In addition to the interval bounds, the dataset includes gender as a covariate. Importantly, no right censoring is present: All patients experienced the event between follow-up visits, which ranged from 2 to 44 months. This characteristic provides a useful baseline for evaluating how the absence of right censoring affects the survival curves obtained from different imputation methods.

As shown in Figure 1 (top left panel), all survival curves, whether based on midpoint, XGBoost-based or parametric imputations (log-normal, Weibull, exponential or Gaussian), are nearly indistinguishable and closely follow the nonparametric Turnbull estimator. This agreement indicates that, when no right censoring is present, all imputation methods recover event times that align well with the true underlying survival distribution. The Turnbull curve overlaps with the Kaplan–Meier estimates based on the imputed times, confirming that no substantial bias was introduced by the imputation process.

Bcos dataset. The second dataset, `bcos`, involves data from a breast cancer study originally analyzed by [23], which has become a benchmark example in the interval-censoring literature. Available in the `interval` package, it records the time to cosmetic deterioration following breast-conserving therapy. Two treatment groups were compared: 46 patients received radiotherapy alone, and 48 received both radiotherapy and chemotherapy. Deterioration was only detected at scheduled assessments, leading to interval-censored times for the 56 women who experienced the event. The remaining 38 patients did not experience deterioration during follow-up and were therefore right-censored, corresponding to roughly 40% of the sample. This dataset represents a realistic setting for evaluating the performance of imputation-based survival methods in the presence of substantial censoring.

Table 2 reports the main characteristics of the three real-world datasets, namely the sample sizes, the proportion of right-censored observations, and the covariates used in the imputation models.

Table 2. Main characteristics of the three real-world datasets used in the application study.

Dataset	Size	Right censored	% Right censored	Covariates
`IR_diabetes`	731	0	0%	Gender
`Bcos`	94	38	$\approx 40\%$	Treatment group
`MiceData`	144	82	$\approx 57\%$	Environmental condition

As shown in Figure 1 (top right panel), larger deviations appear between the imputed survival curves and the Turnbull estimate. The parametric approaches, particularly those based on exponential and Weibull distributions, tend to produce smoother and slightly more optimistic survival estimates at later times, whereas the Turnbull curve shows a steeper decline. This divergence reflects the increased uncertainty in event-time estimation under heavier censoring and the sensitivity of model-based imputations to distributional assumptions. The XGBoost-based method appears to provide a compromise, capturing the overall shape of the Turnbull curve while avoiding excessive smoothing.

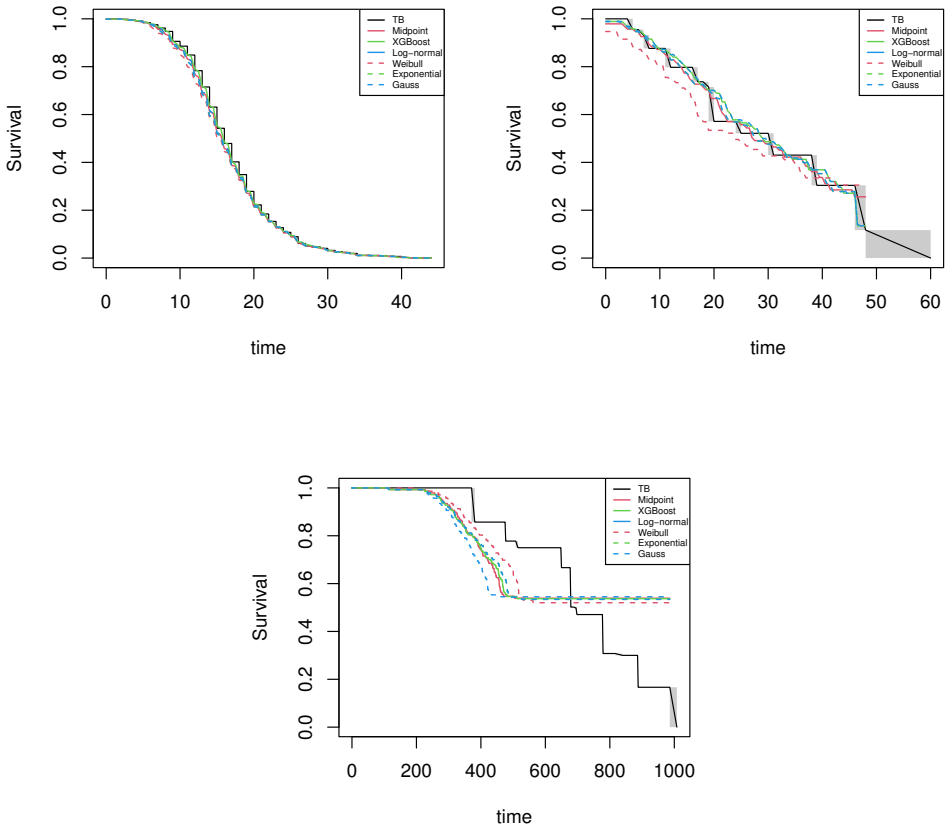

Figure 1. Kaplan–Meier survival curves estimated from imputed event times obtained using midpoint imputation, XGBoost-based imputation and parametric models (log-normal, Weibull, exponential and Gaussian) for three datasets: `IR_diabetes` (top left panel), `bcos` (top right panel) and `miceData` (bottom panel). Turnbull nonparametric estimates are also shown for comparison.

Lung tumor dataset. Finally, the lung tumor dataset from Hoel and Walburg (1972) [24], `miceData`, available in the `icenReg` package, records the development of lung tumors in mice exposed to two environmental conditions: conventional and germ-free environments. Mice were sacrificed at predetermined time points and examined for tumors, resulting in current-status interval-censored data. If a tumor was detected at sacrifice, the event time was considered left-censored, as it must have occurred before observation. Conversely, if no tumor was found, the event time was right-censored, since the tumor had not yet developed. This dataset provides a classic example for assessing the effects of environmental conditions on tumor onset under interval censoring. The dataset includes 144 individuals, of whom 62 experienced the event and 82 were censored, corresponding to 56.9% of right-censored observations.

Figure 1 (bottom panel) shows that, for this dataset, the survival curves derived from imputed event times tend to flatten after approximately 500 time units, while the Turnbull estimator continues to decline. This divergence can be explained by the data structure: among the 144 records, most right-censored observations (those with infinite upper bounds) occur at the longest follow-up times,

typically beyond 400 units. As a result, the imputation methods (midpoint, XGBoost-based and parametric models) treat these cases as late or effectively unobserved events, leading to artificially stable survival probabilities after the last observed failures. In contrast, the Turnbull estimator, which explicitly accounts for interval censoring without imputing missing times, continues to decline gradually, reflecting the remaining uncertainty regarding late events. This pattern highlights how the accumulation of right-censored observations near the end of follow-up can bias imputation-based estimates upward, creating a plateau in the survival function.

Overall, while the imputation methods produce similar survival trends during the early and middle follow-up periods, their limitations in capturing the tail behavior of heavily right-censored data underscore the advantage of interval-censoring methods such as the Turnbull estimator, particularly when censoring is clustered toward the end of observation.

5.1. Overall findings

Overall, greater dispersion among the survival curves is observed as the proportion of right censoring increases, as seen in the `bcos` dataset. This difference is particularly evident in parametric models, such as the Weibull model, which show higher sensitivity to the assumed distributional form and tend to deviate from the Turnbull estimates. In contrast, the XGBoost, exponential, and log-normal models maintain close agreement with the nonparametric estimates, demonstrating robustness to increasing censoring and the ability to recover consistent survival patterns even under substantial uncertainty. These findings partly confirm the results of the simulation study, where the midpoint method provided good approximations when intervals were short or censoring was absent. Parametric models, while appropriate when the underlying distribution is correctly specified, were more sensitive to misspecification and may produce biased estimates, particularly under high levels of censoring.

In extreme cases, such as in the `miceData`, right censoring is concentrated at longer event times, leading to a flattening of the imputed curves after approximately 500 time units. This pattern mirrors the behavior observed in Scenario IV of the simulations, where high right censoring resulted in an underestimation of mortality at longer times when deterministic imputation methods were applied. The Turnbull estimator, by explicitly accounting for interval uncertainty, continues to decline gradually, providing a more realistic representation of long-term survival and underscoring the importance of methods that properly accommodate interval censoring.

In the `IR_diabetes` dataset, the survival curves obtained from different imputation methods are almost identical to one another and to the Turnbull estimate. This apparent agreement does not contradict the simulation finding that bias can be large in the absence of censoring; instead, it reflects the scale of the data in this empirical example, where event times are moderate and absolute biases are small in practice. In the simulations, larger absolute biases were observed in Scenario I because the true simulated times included many large values, inflating absolute difference measures even when relative behavior was similar. Therefore, the close alignment in `IR_diabetes` is a dataset-specific outcome, reflecting shorter follow-up periods, and should not be interpreted as evidence that all methods are uniformly unbiased in every no-censoring setting.

6. Conclusions

Interval-censored data pose challenges in survival analysis because the exact event time is unknown and only its occurrence within an interval is observed. This partial information complicates model estimation and validation, making standard approaches developed for right censoring inadequate or potentially biased when applied directly. The main difficulty lies in estimating survival functions and hazard rates without precise event times, which can lead to misleading inferences if the censoring mechanism is not properly accounted for. Consequently, specialized statistical models and imputation techniques are needed to address the uncertainty inherent in interval-censored observations, often at the cost of increased computational effort and model sensitivity.

The treatment of censoring is a central aspect of the analysis of interval-censored survival data. In this study, censoring is handled through a combination of model-based imputation, flexible predictive methods, and a subsequent adjustment step designed to ensure consistency with the observed censoring intervals. Specifically, latent event times are first generated using regression-based and machine learning models that incorporate covariate information (Sections 3.2 and 3.3), and are then rescaled to satisfy the interval constraints (Section 3.4).

This strategy allows the imputation procedure to exploit the predictive ability of modern statistical and machine learning tools, while explicitly accounting for the uncertainty induced by interval censoring. In contrast with simple approaches such as midpoint imputation, which ignore both covariate effects and the underlying distribution of survival times, the proposed framework preserves individual-level variability and maintains the relative ordering of event times across subjects.

The scaled linear redistribution method plays a key role in this framework by enforcing compatibility between model-based predictions and censoring bounds. By mapping unconstrained predictions back into their corresponding intervals $[L, R]$, the method avoids logically inconsistent imputations while retaining sufficient dispersion for reliable survival estimation.

To this end, in this paper, we systematically evaluated several imputation strategies for interval-censored survival data through simulation studies and applications to real datasets. The results show that machine learning methods, particularly XGBoost-based imputation, are robust and flexible, providing survival estimates that closely match those obtained with the nonparametric Turnbull estimator, including moderate to high right censoring.

On the basis of these results, some practical recommendations can be formulated. When the proportion of right censoring is low to moderate (below approximately 20%), XGBoost-based imputation appears particularly suitable, owing to its flexibility and its ability to capture nonlinear relationships between covariates and survival times. As the level of right censoring increases (above roughly 30%), parametric AFT models, such as the log-normal or Gaussian specifications, tend to provide more stable estimates, benefiting from their explicit distributional structure. Irrespective of the predictive model employed, the scaled linear redistribution method remains essential to ensure coherence between the imputed event times and the observed censoring intervals, while preserving the variability of the survival process.

Analyses of real datasets revealed additional nuances. In the `bcos` data, a higher proportion of right censoring led to greater variability among parametric models, while XGBoost, and Weibull estimates remained relatively consistent. In the `mice` data, censoring concentrated at longer event times caused the imputed survival curves to flatten, illustrating the limitations of deterministic

imputations and emphasizing the advantages of approaches that explicitly model interval uncertainty. The `IR_diabetes` dataset confirmed that apparent agreement among methods often reflects data scale, since shorter follow-up periods tend to reduce absolute biases. Taken together, these findings indicate that although no single approach is optimal for all situations, XGBoost offers a practical and reliable alternative across a wide range of censoring patterns.

In summary, the results indicate that no single imputation method is uniformly optimal across all settings. Parametric AFT models, particularly those based on Gaussian and log-normal distributions, tend to exhibit greater stability under heavy right censoring, whereas XGBoost-based imputation achieves higher precision when censoring is mild to moderate and the data are denser. Embedding these approaches within the proposed scaled linear redistribution framework yields a flexible and reliable tool for transforming interval-censored observations into pseudo-complete datasets suitable for standard regression and survival analyses, while respecting the censoring information.

Although the empirical applications in this paper focus on the effect of different imputation strategies on estimated survival functions, an important extension concerns the assessment of predictive performance on independent data. In practice, such validation is difficult for interval-censored datasets due to limited sample sizes and the absence of fully observed event times. Future work may therefore consider resampling or cross-validation schemes to investigate the impact of imputation on out-of-sample prediction, particularly in machine learning contexts.

Further developments will include the implementation of the scaled linear redistribution method in a dedicated R package, integrating parametric, nonparametric, and machine learning–based imputation procedures, and allowing its application under more complex settings, such as time-dependent covariates and alternative censoring mechanisms.

Data availability

All simulation data and R scripts used in this study are publicly available at `https://github.com/gsoutinho/Imputation-strategies-for-interval-censored-data`.

Author contributions

G.S. and L.M-M. equally contributed to all aspects of the work, including the conception, data analysis, writing, and revision of the manuscript. All authors have read and approved the final version of the manuscript for publication.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the Programme Contract UID/05105/2025. `https://doi.org/10.54499/UID/05105/2025`.

Conflict of interest

No conflict of interest is declared.

References

1. J. P. Klein, M. L. Moeschberger, *Survival analysis: techniques for censored and truncated data*, Springer-Verlag, 1997. https://doi.org/10.1007/978-1-4757-2728-9
2. M. Tableman, J. S. Kim, *Survival analysis using S*, Chapman & Hall Ltd, 2003. https://doi.org/10.1201/b16988
3. D. G. Kleinbaum, M. Klein, *Survival analysis: a self-learning text*, Springer-Verlag, 2012. https://doi.org/10.1007/978-1-4419-6646-9
4. M. Abrahamowicz, M. E. Beauchamp, C. S. Moura, S. Bernatsky, S. F. Guerra, C. Danieli, Adapting SIMEX to correct for bias due to interval-censored outcomes in survival analysis with time-varying exposure, *Biometrical J.*, **64** (2022), 1467—1485. https://doi.org/10.1002/bimj.202100013
5. K. Bogaerts, A. Komarek, E. Lesaffre, *Survival analysis with interval-censored data: a practical approach with examples in R, SAS, and BUGS*, Chapman and Hall/CRC, 2017 https://doi.org/10.1201/9781315116945
6. B. W. Turnbull, The empirical distribution function with arbitrarily grouped, censored and truncated data, *J. R. Stat. Soc. Ser. B (Methodol.)*, **38** (1976), 290–295. https://doi.org/10.1111/J.2517-6161.1976.TB01597.X
7. D. R. Cox, Regression models and life-tables, *J. R. Stat. Soc. Ser. B (Methodol.)*, **34** (1972), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x
8. J. F. Lawless, *Statistical models and methods for lifetime data*, John Wiley & Sons, Inc., 2002. https://doi.org/10.1002/9781118033005
9. J. D. Kalbfleisch, R. L. Prentice, *The statistical analysis of failure time data*, John Wiley & Sons, 2002. http://doi.org/10.1002/9781118032985
10. J. W. Bartlett, R. Keogh, E. F. Bonneville, C. T. Ekstrøm, *smcfcs: Substantive model compatible fully conditional specification*, R package, 2024. Available from: `https://cran.r-project.org/package=smcfcs`.
11. P. Wang, Y. Li, C. K. Reddy, Machine learning for survival analysis: a survey, *ACM Comput. Surv.*, **51** (2019), 1–36. https://doi.org/10.1145/3214306
12. H. Kvamme, Ø. Borgan, Continuous and discrete-time survival prediction with neural networks, *Lifetime Data Anal.*, **27** (2021), 710–736. http://doi.org/10.1007/s10985-021-09532-6
13. Y. Deng, T. Lumley, Multiple imputation through XGBoost, *J. Comput. Graph. Stat.*, **33** (2024), 352–363. https://doi.org/10.1080/10618600.2023.2252501
14. Z. Jinbo, L. Yufu, M. Haitao, Handling missing data of using the XGBoost-based multiple imputation by chained equations regression method, *Front Artif. Intell.*, **8** (2025). https://doi.org/10.3389/frai.2025.1553220

15. I. Štajduhar, B. Dalbelo-Bašić, Uncensoring censored data for machine learning: a likelihood-based approach, *Exp. Syst. Appl.*, **39** (2012), 7226–7234. https://doi.org/10.1016/j.eswa.2012.01.054
16. L. P. Chen, B. Qiu, Analysis of length-biased and partly interval-censored survival data with mismeasured covariates, *Biometrics*, **79** (2023), 3929–3940. https://doi.org/10.1111/biom.13898
17. L. P. Chen, B. Qiu, SIMEXBoost: An R package for analysis of high-dimensional error-prone data based on boosting method, *R J.*, **15** (2023), 5–16. https://doi.org/10.32614/RJ-2023-080
18. E. L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, *J. Am. Stat. Assoc.*, **53** (1958), 457–481.
19. L. Meira-machado, The Kaplan-Meier estimator: new insights and applications in multi-state survival analysis, In: *Computational science and its applications – ICCSA 2023 Workshops*, Lecture Notes in Computer Science, Springer, Cham, 2023, 129–139. http://doi.org/10.1007/978-3-031-37129-5_11
20. V. Kariuki, A. Wanjoya, O. Ngesa, M. M. Mansour, E. M. A. Elrazik, A. Z. Afify, The accelerated failure time regression model under the extended-exponential distribution with survival analysis, *AIMS Math.*, **9** (2024), 15610–15638. https://doi.org/10.3934/math.2024754
21. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, 785–794. https://doi.org/10.1145/2939672.2939785
22. K. Borch-Johnsens, P. K. Andersen, T. Decker, The effect of proteinuria on relative mortality in Type I (insulin-dependent) diabetes mellitus, *Diabetologia*, **28** (1985), 590–596. https://doi.org/10.1007/BF00281993
23. D. M. Finkelstein, R. A. Wolfe, A semiparametric model for regression analysis of interval-censored failure time data, *Biometrics*, **41** (1985), 733–945. https://doi.org/10.2307/2530965
24. D. G. Hoel, H. E. Walburg, Statistical analysis of survival experiments, *J. Natl. Cancer Inst.*, **49** (1972), 361–372. https://doi.org/10.1093/JNCI/49.2.361

AIMS Press