Beyond Kaplan-Meier: A Comprehensive R Package for Interval-Censored Survival Analysis Using Turnbull's Approach

Marta Azevedo[1][0009−0003−4412−6991], Gustavo Soutinho[2][0000−1234−5678−9012] and Luís Meira-Machado[1][0000−0002−8577−7665]

[1] Centre of Mathematics, University of Minho
marta.vasconcelos4@gmail.com
[2] Research on Economics, Management and Information Technologies (REMIT) - Portucalense University

Abstract. Interval-censored data frequently arise in survival analysis when the exact time of an event is unknown but is known to occur within a specific time interval. Traditional methods like the Kaplan-Meier estimator are inadequate for such data, necessitating specialized approaches. This paper presents an R library designed to handle interval-censored data, emphasizing the use of Turnbull's estimator for nonparametric survival estimation. The package offers flexible functionalities, including the calculation of survival estimates, the generation of both static and interactive plots, and the construction of bootstrap-based confidence bands. Additionally, the library provides users with detailed outputs such as Turnbull intervals and their corresponding weights, which are instrumental in understanding the survival distribution and serve as an analogue to Kaplan-Meier weights in right-censored contexts. These weights enable the extension of survival analysis methods to more complex models, including multi-state frameworks. The practical utility of the library is demonstrated using real-world datasets, highlighting its potential to support advanced survival analysis and foster the development of new estimators beyond traditional survival probabilities.

Keywords: survivalTB package · Intervalar censoring · Turnbull estimator.

1 Introduction

Survival analysis is a cornerstone of statistical methodology, widely applied in medical research, reliability engineering, and social sciences to model the time until the occurrence of a specific event, such as system failure or disease progression [1,2,3]. While classical survival analysis methods, such as the Kaplan-Meier estimator [4], are tailored for right-censored data, where the exact event time is either observed or known to exceed a certain time point, many real-world scenarios involve more complex censoring mechanisms. One such instance is interval censoring, where the exact time of the event is unknown, but it is known to have

occurred within a specific time interval [5]. This type of censoring is common in longitudinal studies, medical follow-ups, and reliability testing, where the event of interest is only detectable at discrete inspection times.

Traditional methods like the Kaplan-Meier estimator are inadequate for interval-censored data as they cannot account for the uncertainty within the censoring intervals. To address this, Turnbull [6] proposed a nonparametric maximum likelihood estimator (NPMLE) that generalizes the Kaplan-Meier approach to handle interval-censored data. The Turnbull estimator iteratively computes survival probabilities over a set of non-overlapping intervals, known as Turnbull intervals, and assigns corresponding weights to these intervals to construct the survival curve. These weights serve as an analogue to Kaplan-Meier weights in right-censored data, facilitating further statistical analysis and enabling extensions to more complex models, such as multi-state frameworks.

Despite its theoretical robustness, the practical application of the Turnbull estimator remains challenging, particularly when it comes to integrating confidence bands, conducting bootstrap procedures, or visualizing results interactively. Existing R packages provide partial solutions for analyzing interval-censored data. The widely used `survival` package provides foundational tools for survival analysis, including Kaplan-Meier estimation and Cox proportional hazards models. While it does offer some support for interval-censored data, its capabilities remain limited. The `interval` package implements both parametric and nonparametric methods tailored for interval-censored data, while `icenReg` extends this functionality by supporting flexible parametric and semiparametric regression models. Additionally, `intcox` adapts the Cox model for interval-censored data, and `Icens` focuses on nonparametric estimation and allows for multiple imputation strategies to handle censoring uncertainty. Additionally, the `smicd` package offers a suite of statistical methods tailored for interval-censored data, particularly focusing on socioeconomic indicators and regression modeling using multiply imputed datasets. While valuable in applied contexts, it does not implement Turnbull's estimator or provide the same level of integration with visualization and confidence band functionalities as survivalTB.

However, these tools often lack features that facilitate comprehensive survival analysis, such as the ability to compute bootstrap confidence bands and not all directly extract Turnbull intervals and their corresponding weights. These components are crucial for extending survival analysis to more complex contexts, such as multi-state models, and for enabling the development of novel statistical estimators.

To address these gaps, this paper introduces the `survivalTB` package, an R library designed specifically for interval-censored survival data. The package focuses on enhancing user accessibility and analytical flexibility by providing tools for: (i) Estimating survival functions using Turnbull's estimator, (ii) Computing bootstrap-based confidence bands, (iii) Extracting Turnbull intervals and their corresponding weights, (iv) Generating both static and interactive survival plots.

The `survivalTB` package aims to simplify complex survival analysis tasks while maintaining methodological rigor, offering a valuable resource for researchers and practitioners dealing with interval-censored data.

The remainder of this paper is organized as follows. Section 2 describes the methodological framework for survival function estimation, with a focus on interval censoring and Turnbull's estimator. Section 3 introduces the `survivalTB` package, detailing its core functionalities and implementation. Section 4 presents a practical application of the package using a real biomedical dataset, while Section 5 discusses its applications beyond survival curve estimation. Finally, Section 6 provides concluding remarks.

2 Methods

2.1 Survival function estimation

The survival function, denoted as $S(t)$, represents the probability that an individual or object survives beyond a given time t, $S(t) = P(T > t)$ where T is a non-negative random variable representing the event time. Survival analysis plays a crucial role in various fields, including medical research, reliability engineering, and social sciences, as it allows for the estimation of event probabilities over time.

The survival function can be estimated using parametric or nonparametric approaches. Parametric methods assume that the survival times follow a specific distribution, such as exponential or Weibull, allowing for a model-based interpretation. While these methods can provide efficient estimates when the assumed distribution is appropriate, they may lead to biased results if the underlying distributional assumptions do not hold.

In contrast, nonparametric estimators make fewer assumptions about the underlying survival distribution, offering greater flexibility in modeling real-world data. A well-known example is the Kaplan-Meier estimator [4], widely used for right-censored data. This estimator calculates survival probabilities at observed event times, producing a stepwise survival function. It effectively adjusts for right-censored observations by incorporating only the individuals at risk at each event time.

Much of the work in survival analysis has focused on right-censored data. However, in many practical applications, event times are not exactly observed but are only known to lie within a given interval. This scenario, known as interval censoring, occurs frequently in clinical trials, reliability testing, and longitudinal studies where events are only detected at discrete inspection times.

Formally, for an individual i, the event time T_i is known to lie within an observation interval $[L_i, R_i)$, where $L_i \leq T_i < R_i$. If $R_i = \infty$, the observation is right-censored, meaning that the event has not been observed up to the last recorded time. If $L_i = R_i$, the exact event time is known. When $L_i < R_i < \infty$, the event is interval-censored, indicating that the event occurred within a specific time interval but without precise knowledge of the exact time.

A common but suboptimal approach to dealing with interval censoring is to approximate event times using imputation methods such as left, right, or mid-point imputation. These methods assign a single estimated event time within the censoring interval and then apply standard right-censoring techniques, such as the Kaplan-Meier estimator. However, this approach introduces bias and underestimates the variability of survival estimates, as it ignores the full uncertainty associated with the censoring interval.

A more appropriate method for interval-censored data is the Turnbull estimator [6], which generalizes the Kaplan-Meier approach by iteratively computing survival probabilities over non-overlapping intervals. This estimator constructs a nonparametric maximum likelihood estimate (NPMLE) of the survival function while properly accounting for interval censoring.

In this work, we focus on the Turnbull estimator for interval-censored data and its implementation within the `survivalTB` package, which facilitates the estimation of survival probabilities while providing additional analytical tools such as confidence bands and interval weight extraction.

2.2 Turnbull Estimator

The Turnbull estimator [6] extends the Kaplan-Meier estimator to interval-censored data through a nonparametric maximum likelihood estimation (NPMLE) approach. It is specifically designed to account for the uncertainty introduced by interval censoring by incorporating likelihood contributions from exact, right-censored, and interval-censored observations.

Construction of Turnbull Intervals The first step in applying the Turnbull estimator involves determining a set of disjoint intervals, known as Turnbull intervals, over which the survival function will be estimated. These intervals are derived from the union of all observed censoring intervals. Let $\mathcal{D} = \{[L_i, R_i)\}_{i=1}^{n}$ represent the set of all observed intervals. The Turnbull intervals are constructed by identifying the distinct endpoints from $\mathcal{D}$, ordering them, and forming non-overlapping intervals that cover the support of T.

Specifically, these intervals are defined using the set of unique left and right endpoints, denoted as $I_j = [u_j, v_j)$, where u_j corresponds to a left endpoint, v_j to a right endpoint, and no other observed endpoint exists between them.

Likelihood Function and Weights The NPMLE is obtained by maximizing the likelihood function over the discrete distribution of event times across the Turnbull intervals. Let p_j denote the probability that the event occurs in interval I_j.

For each interval $I_j = [u_j, v_j)$, the probability mass p_j, representing the probability that the event time T falls within that interval, is defined as $p_j = P(u_j \leq T < v_j)$. These probabilities must satisfy:

$$\sum_{j=1}^{m} p_j = 1, \quad p_j \geq 0 \tag{1}$$

The likelihood contribution for an observation i with censoring interval $[L_i, R_i)$ is:

$$l_i = \sum_{j:\, I_j \subset [L_i, R_i)} p_j \tag{2}$$

where the summation is taken over all Turnbull intervals $I_j = [u_j, v_j)$ that are fully contained within the observed censoring interval $[L_i, R_i)$ for individual i.

The total log-likelihood for the sample is:

$$\mathcal{L} = \sum_{i=1}^{n} \log \left(\sum_{j: I_j \subset [L_i, R_i)} p_j \right) \tag{3}$$

Maximizing this likelihood is typically achieved using the Expectation-Maximization (EM) algorithm. The iterative steps are as follows:

E-step: Compute the expected number of events in each interval I_j based on the current estimates of p_j:

$$E_j = \sum_{i=1}^{n} \frac{p_j \mathbf{1}_{\{I_j \subset [L_i, R_i)\}}}{\sum_{k: I_k \subset [L_i, R_i)} p_k} \tag{4}$$

M-step: Update the probability estimates:

$$p_j^{(t+1)} = \frac{E_j}{n} \tag{5}$$

These steps are repeated until convergence, typically when the change in the log-likelihood between iterations falls below a predefined threshold.

Estimating the Survival Function Once the weights p_j for each Turnbull interval are obtained, the survival function is estimated as a step function. The survival probability at the right endpoint of each interval t_j is given by:

$$\hat{S}(t_j) = 1 - \sum_{k=1}^{j} p_k \tag{6}$$

This step function decreases at the endpoints of the Turnbull intervals, similar in structure to the Kaplan-Meier curve but adapted for the interval-censored data.

Properties and Interpretation The Turnbull estimator is a nonparametric and consistent estimator for the survival function under interval censoring. The weights p_j represent the estimated proportion of failures within each Turnbull interval and play a role analogous to the Kaplan-Meier jump sizes at exact event times. The flexibility of the estimator allows it to accommodate various forms of censoring, including right-censoring and exact observations.

Confidence intervals for the survival estimates can be obtained using bootstrap resampling techniques, providing additional measures of uncertainty in the survival estimates.

3 The survivalTB package

The survivalTB package provides a comprehensive and flexible workflow for conducting interval-censored survival analysis. It offers functions to preprocess interval-censored data, estimate survival functions using Turnbull's estimator, compute bootstrap-based confidence intervals, and create both static and interactive visualizations. The package is designed to address the complexities inherent in interval-censored data, which frequently arise in clinical trials, epidemiological studies, and reliability engineering. Its implementation in R [7] ensures reproducibility and seamless integration with other statistical packages, promoting efficient data analysis workflows.

The core functionality of the package is structured around three main functions:

- TNBintervals: Estimates survival distributions for interval-censored data using Turnbull's estimator, with optional bootstrap replications for confidence interval estimation.
- TNBsurvival: Provides time-specific survival estimates and allows for the calculation of confidence intervals at user-defined confidence levels.
- plot.TB and plot.TBL: Facilitate the visualization of survival curves, with plot.TB using base R graphics and plot.TBL offering interactive, Plotly-based visualizations.

These functions enable users to handle interval-censored survival data efficiently while providing flexibility in analysis and presentation.

A detailed summary of the arguments for each function is presented in Tables 1, 2, and 3.

Argument	Description
left	Numeric vector of left interval boundaries.
right	Numeric vector of right interval boundaries (can include NA for right-censored data).
nboot	Integer specifying the number of bootstrap replications. Default is 1 (no bootstrap).

Table 1. Summary of the arguments for the function TNBintervals.

4 Example of application

To illustrate the application of survivalTB, we use the bcos dataset from the interval package, which contains interval-censored survival times of breast cancer patients.

Argument	Description
`data`	A list containing survival estimates (output of `TNBintervals`).
`times`	Numeric vector of time points for which survival estimates are required.
`conf`	Logical indicating whether to calculate confidence intervals. Default is `FALSE`.
`conf.level`	Numeric value specifying the confidence level for intervals (e.g., 0.95 for 95%). Default is 0.95.

Table 2. Summary of the arguments for the function `TNBsurvival`.

Argument	Description
`x`	A list containing survival estimates (output of `TNBintervals`).
`conf`	Logical indicating whether to include confidence bands. Default is `FALSE`.
`conf.level`	Numeric value specifying the confidence level for the bands (e.g., 0.95 for 95%). Default is 0.95.
`main`	Character string specifying the title of the plot.
`xlab`	Character string for the x-axis label.
`ylab`	Character string for the y-axis label.
`line.col`	Character string specifying the color of the survival curve. Default is `"blue"`.
`showlegend`	Logical indicating whether to display the legend. Default is `TRUE`.
`main.line`	Character string for the label of the main survival curve in the plot legend. Default is `"survival"`.
`filled`	Character string for the label of the confidence interval ribbon in the legend. Default is `"CI"`.
`line.width`	Numeric value specifying the line width for the survival curve. Default is 2.
`fillcolor`	Character string specifying the color for the confidence band fill (e.g., `"rgba(128, 128, 128, 0.3)"`). Default is grey.
`...`	Additional graphical parameters.

Table 3. Summary of the arguments for the function `plot.TBL`.

4.1 Loading the Package and Exploring the Data

Originally described by Finkelstein and Wolfe (1985) [8], this dataset consists of 94 observations and examines the time to cosmetic deterioration in breast cancer patients. The study includes two treatment groups: 46 patients who received radiotherapy alone and 48 patients who received both radiotherapy and chemotherapy. Since deterioration was only detected during clinical visits, 56 patients were interval-censored, while the remaining 38, who did not experience deterioration, were right-censored.

The dataset includes three variables. The variable `left` represents the lower boundary of the time interval during which the event was observed or censored, while `right` corresponds to the upper boundary. If `right` is `Inf`, the observation is right-censored. The variable `treatment` indicates the type of treatment

received, distinguishing between radiotherapy alone (`Rad`) and a combination of radiotherapy and chemotherapy (`RadChem`).

We begin by loading the `survivalTB` package and importing the `bcos` dataset:

```
library(interval)
head(bcos)
```

```
  left right treatment
1   45   Inf       Rad
2    6    10       Rad
3    0     7       Rad
4   46   Inf       Rad
5   46   Inf       Rad
6    7    16       Rad
```

In this dataset, the `left` and `right` columns define the time interval within which the event was observed or censored. When `right` is `Inf`, the observation is right-censored, meaning the event was not recorded during the study period.

The presence of both interval and right censoring in the dataset makes it particularly valuable for assessing the flexibility and performance of survival analysis techniques, including those implemented in the `survivalTB` package.

4.2 Estimating the Survival Function

The `survivalTB` package enables estimation of survival probabilities from interval-censored data, supports bootstrap replication, and provides visualization tools.

To obtain survival estimates and generate the corresponding plots, it is essential to first use the `TNBintervals` function. This function identifies the set of all Turnbull intervals —the non-overlapping intervals used in the Turnbull estimator for interval-censored data. Additionally, it calculates the weights assigned to each Turnbull interval, which are crucial for deriving survival estimates.

```
library(survivalTB)
tbdata <- TNBintervals(bcos$left, bcos$right, nboot = 500)
tbdata$original
```

```
   left right weight survival
1     4     5 0.0449   0.9551
2     6     7 0.0226   0.9325
3     7     8 0.0560   0.8765
4    11    12 0.0790   0.7975
5    16    17 0.0605   0.7370
6    18    19 0.0216   0.7154
7    19    20 0.1441   0.5713
8    24    25 0.0497   0.5216
9    30    31 0.0911   0.4305
```

```
10   38    39 0.1264   0.3041
11   46    48 0.1869   0.1172
12   48    60 0.1170   0.0002
```

The output consists of three key components: Turnbull intervals, represented by the `left` and `right` columns; weights, which indicate the proportion assigned to each Turnbull interval; and cumulative survival estimates, which reflect the survival probability up to each interval.

We can compute survival probabilities at specific time points using the `TNBsurvival()` function:

```
times <- c(5, 10, 15, 20, 25, 30, 40, 45, 50)
est <- TNBsurvival(data = tbdata, times = times, conf = TRUE,
conf.level = 0.95)
print(est)
```

The resulting survival estimates are:

```
  time survival   ci_lower  ci_upper
1    5   0.9551  0.8951200 1.0000000
2   10   0.8765  0.8073350 0.9423600
3   15   0.7975  0.6939475 0.8776575
4   20   0.5713  0.4509125 0.7043875
5   25   0.5216  0.4234750 0.6416075
6   30   0.5216  0.3765175 0.6279675
7   40   0.3041  0.1888325 0.4364175
8   45   0.3041  0.1888325 0.4320300
9   50   0.0977  0.0000000 0.3350450
```

4.3 Visualizing the Survival Curve

To visualize the survival curve and include confidence bands, we use the `plot()` function:

```
plot(tbdata, conf = TRUE, conf.level = 0.95)
```

The plot displays the estimated survival curve alongside 95% confidence bands, enhancing the interpretability of the results. This visualization allows users to explore survival probabilities over time, providing insights into the progression of the event of interest.

An interactive plot, generated using the Plotly R package, can be obtained by setting the argument *interactive = TRUE*.

5 Applications Beyond Survival Curves

In recent years, several methods have been proposed in the literature for estimating key quantities in multi-state models, including the state occupation

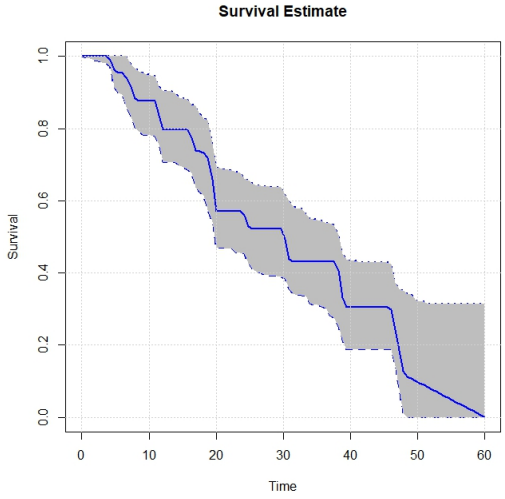


Fig. 1. Estimated survival curve using the Turnbull Estimator with 95% confidence bands.

probabilities and transition probabilities. Notably, the work of Uña-Álvarez and Meira-Machado (2015) [9], published in *Biometrics*, introduced innovative techniques for estimating these probabilities under right-censored data. However, most of these methods are designed specifically for right-censored datasets and do not readily extend to interval-censored scenarios. Furthermore, the existing literature on multi-state models under interval censoring remains sparse, leaving a significant gap in methodological development.

Many of the existing methods for right-censored data rely on the Kaplan-Meier estimator or the associated Kaplan-Meier weights to derive occupation and transition probabilities. While effective in right-censored contexts, these methods cannot be directly applied to interval-censored data due to the inherent uncertainty in event times. However, the principles underlying these approaches can be adapted by substituting the Kaplan-Meier estimator with the Turnbull estimator and using the weights assigned to the Turnbull intervals. This substitution allows for the extension of existing multi-state analysis methods to interval-censored data, enabling more comprehensive and accurate modeling.

The `survivalTB` package provides researchers with the necessary tools to implement these adaptations. By offering access to both the survival estimates and the detailed structure of the Turnbull intervals along with their associated weights, the package equips users with the essential components needed for developing and applying advanced estimation methods in multi-state models under interval censoring.

Moreover, the versatility of the `survivalTB` package extends beyond simple survival analysis. Its functionalities can be applied to more complex frameworks,

including recurrent event models and competing risks, broadening its utility across a range of research fields. Ongoing work is also focused on adapting and extending existing multi-state methods to fully accommodate interval-censored data, aiming to close the current methodological gaps and provide researchers with more robust analytical tools.

The survivalTB library, along with all its functionalities, is freely available at `https://github.com/martaaaa/survivalTB`. This repository provides access to the latest version of the library, documentation, and examples to facilitate its use.

6 Conclusion

The `survivalTB` package provides a robust and user-friendly framework for conducting survival analysis with interval-censored data. Leveraging Turnbull's estimator and bootstrap resampling techniques, it supports accurate survival probability estimation while offering tools for intuitive and interactive data visualization. Its comprehensive feature set makes it a valuable resource for researchers and practitioners in medical, epidemiological, and reliability studies.

Acknowledgments. This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the UID/00013: Centro de Matemática da Universidade do Minho (CMAT/UM) Program Contract, and the project reference 2023.14897.PEX (DOI: 10.54499/2023.14897.PEX).

References

1. Klein, J., Moeschberger, M.: Survival analysis - techniques for censored and truncated data. New York: Springer-Verlag (1997).
2. Tableman, M., Kim, J.: Survival analysis using S. Chapman & Hall Ltd (2003).
3. Kleinbaum, D., Klein, M.: Survival Analysis: A Self-Learning Text. New York: Springer-Verlag (2012).
4. Kaplan, E. L., Meier, P.: Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53, 457-481 (1958).
5. Radke, B. R.: A demonstration of interval-censored survival analysis. Preventive Veterinary Medicine 59, 241–256 (2003).
6. Turnbull, B. W.: The empirical distribution function with arbitrarily grouped, censored and truncated data. J. Roy. Stat. Soc. Ser. B 38, 290-295 (1976).
7. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/` (2024).
8. Finkelstein, D.M., Wolfe, R.A.: A semiparametric model for regression analysis of interval-censored failure time data. Biometrics 41, 731-740 (1985).
9. de Uña-Álvarez, J., Meira-Machado, L.: Nonparametric estimation of transition probabilities in the non-Markov illness-death model: A comparative study. *Biometrics*, 71(2), 364–375 (2015).