Survival Analysis of HIV Seroconversion Using Interval-Censored Data: Nonparametric and Regression Approaches

Rui Alves[1], Filipa Pinto[1], Aurélio Sidumo[1], Carla Moreira[2], Luís Meira-Machado[1(✉)], Paula Meireles[3], Miguel Rocha[3,4], and Maria João Novais[3]

[1] Centre of Mathematics, University of Minho, Campus de Azurém, 4800-058 Guimarães, Portugal
lmachado@math.uminho.pt
[2] RISE-Health, Faculty of Medicine, University of Porto, Porto, Portugal
[3] EPIUnit ITR, Instituto de Saúde Pública da Universidade do Porto, Rua das Taipas 135, 4050-600 Porto, Portugal
[4] GAT – Grupo de Ativistas em Tratamentos, Lisbon, Portugal

Abstract. In prospective HIV cohorts, the exact time of seroconversion is rarely observed. Individuals are tested at irregular follow-up visits, so infection is only known to occur between the last negative and the first positive test. This leads to interval-censored event times, while non-seroconverters are subject to right censoring. Standard survival methods that assume exact or right-censored event times may therefore produce biased estimates.

In this paper, we analyse time to HIV seroconversion using methods specifically designed for interval-censored data. We first describe the Lisbon MSM cohort and its censoring structure. We then estimate the survival function using Turnbull's nonparametric maximum likelihood estimator and compare it with Kaplan–Meier estimators based on left, right and midpoint imputation. Group comparisons are conducted using tests adapted to interval-censored data. Finally, we fit parametric accelerated failure time models and proportional hazards models that explicitly account for interval censoring.

The results show that ad hoc imputation may lead to different survival estimates and that education level is the only covariate showing a statistically significant association with time to seroconversion in this application. In particular, higher educational attainment was associated with longer estimated times to HIV seroconversion under the AFT framework. These findings highlight the importance of using appropriate survival methods in HIV incidence studies based on periodic testing.

Keywords: HIV seroconversion · Interval-censored data · Survival analysis · Turnbull estimator · Imputation methods

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16759, pp. 344–357, 2027.
https://doi.org/10.1007/978-3-032-30518-3_22

1 Introduction

Accurate estimation of the time to HIV seroconversion is a key component in the epidemiological analysis of prospective cohorts and in the assessment of risk factors associated with infection. In many observational studies, individuals are tested for HIV at discrete and often irregular time points. As a result, seroconversion is not observed exactly, but is only known to have occurred between the last negative and the first positive HIV test. This gives rise to interval-censored event times, while individuals who remain HIV-negative throughout follow-up are right-censored.

In practice, survival analyses in HIV research often rely on methods developed for exact or right-censored event times, such as the Kaplan–Meier estimator and the Cox proportional hazards model. These methods are sometimes combined with ad hoc imputation of the event time, for example by taking the midpoint between the last negative and the first positive test. Although simple to implement, such approaches ignore the uncertainty about the true infection time and may affect the estimation of survival probabilities and covariate effects, particularly when the intervals between visits are wide or highly variable.

The statistical literature on interval-censored data provides a rigorous framework to address this problem. The nonparametric maximum likelihood estimator of the survival function for interval-censored data was introduced by Turnbull [1] and remains the standard reference in this setting. Extensions of classical inference procedures, including comparison of survival curves and regression modelling, have also been developed, covering both proportional hazards models [2,3] and accelerated failure time models [4,5]. Detailed discussions of these methods can be found in [6,7].

From an epidemiological perspective, the Lisbon cohort of men who have sex with men (MSM), established at a community-based voluntary counselling and testing centre, provides a relevant example of a prospective study based on repeated and irregular HIV testing [8,9]. This setting is particularly suited to illustrating the practical consequences of analysing interval-censored event times using either appropriate likelihood-based methods or simpler imputation-based alternatives.

The main objectives of this paper are as follows. First, we describe the data structure and the construction of the censoring intervals in the Lisbon MSM cohort. Second, we estimate the survival function using Turnbull's estimator and compare it with Kaplan–Meier estimators obtained after left, right and midpoint imputation. Third, we compare survival curves across levels of selected categorical covariates using methods adapted to interval-censored data. Finally, we assess the effect of covariates through regression models that explicitly account for interval censoring.

2 Data and Censoring Structure

2.1 Study Cohort and Analytical Sample

The data arise from the Lisbon cohort of men who have sex with men (MSM), a prospective observational study conducted at a community-based voluntary HIV counselling and testing centre in Lisbon, Portugal [8,9]. The cohort includes individuals who were initially HIV-negative and were subsequently followed through repeated, irregularly spaced testing visits.

The initial dataset comprised 11,389 individuals, among whom 521 had at least one positive HIV test result. Of these, 346 individuals were already HIV-positive at their first visit and were therefore excluded from the survival analysis, since the time to seroconversion could not be determined. The remaining 175 individuals had a positive HIV test after baseline.

Among the 11,043 individuals who were HIV-negative at the first visit, 5,433 had only a baseline observation and no subsequent HIV test. These individuals were excluded from the survival analysis, since no follow-up interval could be defined and they would otherwise correspond to observations of the form $(0, \infty)$, providing no information on the timing of seroconversion beyond baseline negativity.

After excluding these individuals, 5,610 participants had at least two testing visits. Eligibility criteria were then applied, restricting the analysis to participants aged 18 years or older, of male gender, with a valid testing record, and HIV-negative at the first visit. The final analytical sample included 5,106 participants, among whom 171 experienced HIV seroconversion during follow-up.

Figure 1 summarizes the selection process of the analytical sample.

2.2 Censoring Structure

Let T_i denote the unobserved time to HIV seroconversion for individual i, measured from the date of the first negative test. For individuals who seroconverted during follow-up, T_i is not observed exactly. Instead, it is only known to belong to an interval $(L_i, R_i]$, where L_i is the time of the last negative test and R_i is the time of the first positive test, with $0 \leq L_i < R_i < \infty$.

For individuals who remained HIV-negative throughout the observation period, the event time is right-censored. In this case, L_i is the time of the last negative test and $R_i = \infty$, so that $T_i > L_i$. Thus, the observed information for each participant can be represented by the interval $(L_i, R_i]$, with R_i possibly infinite.

Participants with only a baseline negative test and no subsequent follow-up were not included in the survival analysis, since no censoring interval could be defined beyond the baseline observation.

The visit process was irregular, reflecting the voluntary nature of testing and follow-up. Consequently, the lengths of the censoring intervals varied across participants. This feature is central to the analysis, since replacing each interval by a single imputed time ignores part of the uncertainty about the true seroconversion time.

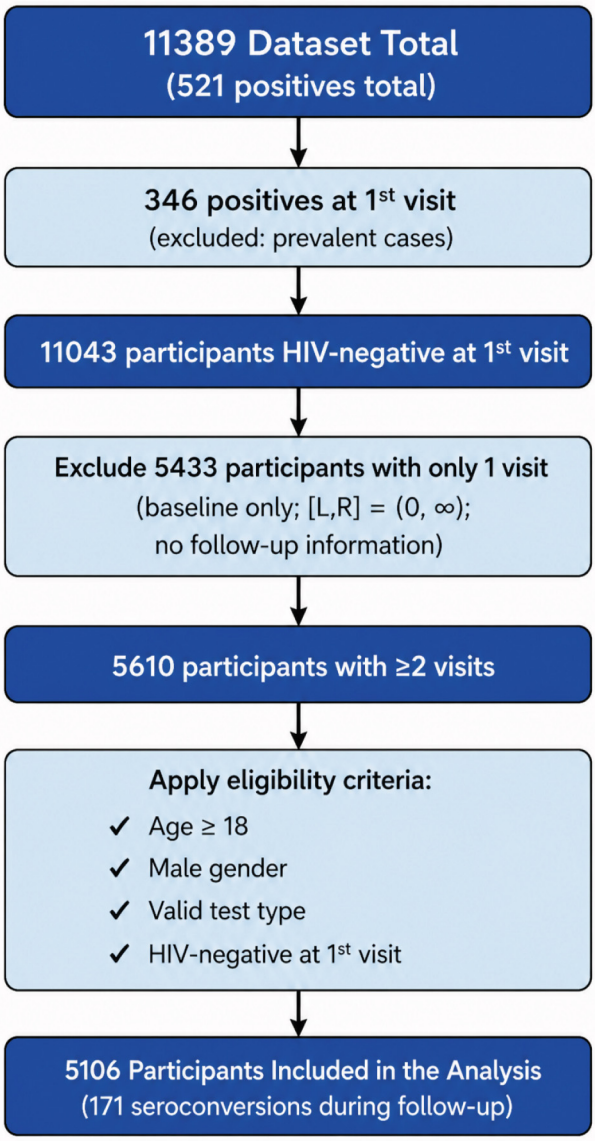

Fig. 1. Flowchart of the selection of the analytical sample for the analysis of interval-censored HIV seroconversion times.

The analysis assumes that the observation process, namely the timing of HIV testing visits, is independent of the underlying time to seroconversion, conditional on the observed covariates. This is a standard working assumption in analyses of interval-censored data collected through periodic observation, but its plausibility is revisited in the Discussion.

2.3 Covariates and Missing Data

The available covariates include demographic and behavioural variables collected through questionnaires at baseline and, when appropriate, during follow-up. These include age, sex assigned at birth, gender identity, previous HIV testing, substance use, sex with casual partners, condom use during anal sex, education level and country of birth. The dataset also includes the dates of the different HIV testing visits, which were used to construct the intervals $(L_i, R_i]$.

Table 1 summarizes the main sociodemographic and behavioural characteristics of the analytical sample. For each variable, the table reports the number and percentage of participants in each observed category, together with the number and percentage of seroconversion events. Missing or unanswered responses are shown explicitly whenever present, since the corresponding denominators differ across variables. Percentages for the categories are calculated using the total analytical sample as denominator, unless otherwise stated.

Some variables contain a substantial proportion of missing or unanswered responses, mainly because not all questions were available or completed for all participants. This is particularly evident for variables such as sex assigned at birth, substance use and country of birth. For this reason, the descriptive table should be interpreted as a summary of the available questionnaire information rather than as a complete profile for all variables.

Although the inclusion criteria restricted the analysis to individuals of male sex, the variable "sex assigned at birth" was not available for all participants in the questionnaire data. Similarly, gender identity was recorded as "man" for almost all participants with available information, but a small number of missing responses remained. These missing values explain the apparent discrepancy between the eligibility criteria and the totals displayed in Table 1.

3 Methods

3.1 Nonparametric Estimation

Let T denote the time to HIV seroconversion, with survival function $S(t) = P(T > t)$. In the presence of interval censoring, each observation is of the form $(L_i, R_i]$, indicating that the event occurred between two observation times.

The nonparametric maximum likelihood estimator (NPMLE) of S in this setting is given by Turnbull's estimator [1]. This estimator generalizes the Kaplan–Meier estimator to interval-censored data and is obtained by maximizing the likelihood

$$\prod_{i=1}^{n} \left[F(R_i) - F(L_i) \right], \tag{1}$$

for interval-censored observations, with contribution $S(L_i)$ for right-censored individuals, where $F(t) = 1 - S(t)$.

The resulting estimator is a step function with jumps at a subset of the interval endpoints and is typically computed using the self-consistency algorithm.

Table 1. Sociodemographic and behavioural characteristics of the analytical sample

Variable	n (%)	n Events (%)
Sex assigned at birth		
Male	603 (11.8)	5 (0.8)
Missing/unanswered	4503 (88.2)	166 (3.7)
Gender identity		
Man	5071 (99.3)	170 (3.4)
Missing/unanswered	35 (0.7)	1 (2.9)
Previous HIV test		
Yes	4148 (81.2)	138 (3.3)
No	889 (17.4)	33 (3.7)
Missing/unanswered	45 (0.9)	0 (0.0)
Substance use		
No	1709 (33.5)	107 (6.3)
Yes	23 (0.5)	1 (4.3)
Does not know	176 (3.4)	19 (10.8)
Missing/unanswered	3193 (62.5)	43 (1.3)
Sex with casual partners		
Yes	4358 (85.4)	153 (3.5)
No	686 (13.4)	16 (2.3)
Missing/unanswered	4 (0.1)	0 (0.0)
Condom use (anal sex)		
Always	737 (14.4)	27 (3.7)
Almost always	681 (13.3)	18 (2.6)
Never	728 (14.3)	28 (3.8)
Rarely	449 (8.8)	15 (3.3)
Sometimes	439 (8.6)	13 (3.0)
Missing/unanswered	2064 (40.4)	69 (3.3)
Education level		
Bachelor's degree	2028 (39.7)	66 (3.3)
Secondary school	1424 (27.9)	66 (4.6)
Master's/PhD	1071 (21.0)	19 (1.8)
Vocational training	337 (6.6)	15 (4.5)
$\leq$ Basic school	165 (3.2)	4 (2.4)
Missing/unanswered	81 (1.6)	1 (1.2)
Country of birth		
Portugal	1494 (29.3)	95 (6.4)
Brazil	187 (3.7)	15 (8.0)
Other European countries	127 (2.5)	7 (5.5)
African countries	71 (1.4)	8 (11.3)
Other countries	37 (0.7)	0 (0.0)
Missing/unanswered	3190 (62.5)	46 (1.4)

3.2 Kaplan–Meier Estimators with Imputation

For comparison purposes, we also consider Kaplan–Meier estimators obtained after imputing a single event time within each censoring interval. Three imputation strategies are used: left imputation ($T_i = L_i$), right imputation ($T_i = R_i$) and midpoint imputation ($T_i = (L_i + R_i)/2$).

These approaches transform interval-censored observations into pseudo-exact event times, allowing the use of standard Kaplan–Meier estimation. However, they ignore the uncertainty about the true event time and may introduce bias in the estimation of survival probabilities.

3.3 Comparison of Survival Curves

To assess differences in survival distributions across levels of categorical covariates, we employ log-rank type tests adapted to interval-censored data. These tests are based on score statistics derived from the proportional hazards model or on permutation procedures using the NPMLE of the survival function [3,6].

The null hypothesis corresponds to equality of survival functions across groups. These methods extend the classical log-rank test by incorporating interval-censored likelihood contributions instead of exact event times.

3.4 Regression Models

To evaluate the effect of covariates on the time to HIV seroconversion, we considered both accelerated failure time (AFT) and proportional hazards (PH) models adapted to interval-censored data, with particular emphasis on the parametric AFT approach.

The main regression analysis was based on parametric accelerated failure time (AFT) models fitted using the `ic_par()` function from the `icenReg` package. A semiparametric proportional hazards model for interval-censored data was also considered using the `ic_sp()` function as a complementary analysis.

Under the AFT parametrization used by `icenReg`, the survival function is expressed as

$$S(t \mid \mathbf{x}_i) = S_0\left(t \exp(\mathbf{x}_i^\top \boldsymbol{\beta})\right), \tag{2}$$

where S_0 is the baseline survival function and $\mathbf{x}_i$ is the vector of covariates. With this parametrization, negative coefficients correspond to longer survival times, whereas positive coefficients correspond to shorter survival times. Equivalently, the time ratio associated with a covariate is given by $\exp(-\beta)$.

Likelihood-based inference is performed by integrating the corresponding density over the interval $(L_i, R_i]$, thus explicitly accounting for interval and right censoring.

In the PH framework, the hazard function is specified as

$$\lambda(t \mid \mathbf{x}_i) = \lambda_0(t) \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \tag{3}$$

where $\lambda_0(t)$ is an unspecified baseline hazard. Estimation is carried out using likelihood-based methods adapted to interval-censored data [2,3].

3.5 Implementation Details

All analyses were performed in the R statistical environment [10]. Nonparametric estimation and hypothesis testing for interval-censored data were carried out using the `interval` package, while regression models were implemented using the `icenReg` package [11]. Standard survival methods were obtained using the `survival` package.

Several parametric distributions were explored for the AFT models, including log-normal, log-logistic, Weibull and exponential distributions. The log-normal specification provided the best fit according to the Akaike Information Criterion (AIC) and was therefore retained for the main regression analysis.

To avoid degenerate intervals when $L_i = 0$, a small positive constant ($\varepsilon = 10^{-4}$) was added when necessary to ensure valid interval limits. For categorical variables, categories corresponding to missing responses were excluded from regression analyses and appropriate reference categories were defined.

4 Results

4.1 Descriptive Characteristics

The main sociodemographic and behavioural characteristics of the analytical sample are summarized in Table 1. The sample includes 5,106 participants, among whom 171 HIV seroconversion events were observed during follow-up.

Overall, most participants reported previous HIV testing and a high proportion indicated having sex with casual partners. Condom use during anal sex showed considerable variability across individuals. Education levels were heterogeneous, with a large proportion of participants reporting secondary or higher education. Most participants were born in Portugal, although a non-negligible proportion originated from other countries.

As discussed in Sect. 2, several variables include missing or unanswered responses, which should be taken into account when interpreting the distributions presented in Table 1.

4.2 Nonparametric Survival Estimates

Figure 2 presents the estimated survival function for time to HIV seroconversion obtained using Turnbull's estimator. The survival probability remains close to one over most of the follow-up period, reflecting the relatively low incidence of seroconversion in the cohort. A gradual decline is observed as follow-up time increases, corresponding to the occurrence of new infections.

For comparison purposes, Fig. 3 displays the Kaplan–Meier survival curves obtained under the three imputation strategies in a single graph. Systematic differences are observed across the three approaches. As expected, left imputation produces lower survival estimates, reflecting the fact that events are assigned to the earliest possible time within each interval.

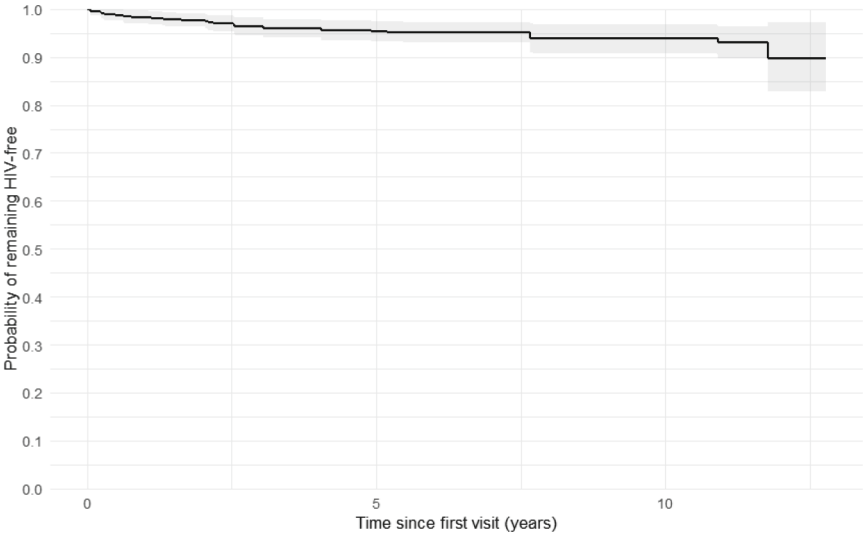

Fig. 2. Estimated survival function for time to HIV seroconversion using Turnbull's estimator.

In this dataset, differences between the imputation strategies reflect the distribution of censoring intervals. When the last negative test occurs at baseline, left imputation assigns the event time close to zero, whereas midpoint and right imputation place the event later within the interval.

In contrast, midpoint and right imputation yield very similar survival estimates. This reflects the fact that, for a large proportion of individuals, the censoring intervals are relatively short or dominated by the timing of the first positive test, so that the midpoint and right endpoints are close. For clarity, the figure is displayed using a restricted vertical scale (0.75–1), which allows these small differences to be more easily visualised.

These differences illustrate the impact of replacing each censoring interval by a single imputed value. In the present data, the effect is not uniform across imputation strategies: left imputation leads to a noticeable downward shift of the survival curve, while midpoint and right imputation produce very similar estimates. This pattern reflects the distribution of censoring intervals in the cohort and highlights how imputation-based approaches may distort the timing of events.

4.3 Comparison Across Covariates

Differences in survival distributions across levels of categorical covariates were assessed using log-rank type tests adapted to interval-censored data.

Among the variables considered, education level was the only covariate for which statistically significant differences were detected (test statistic = 16.37,

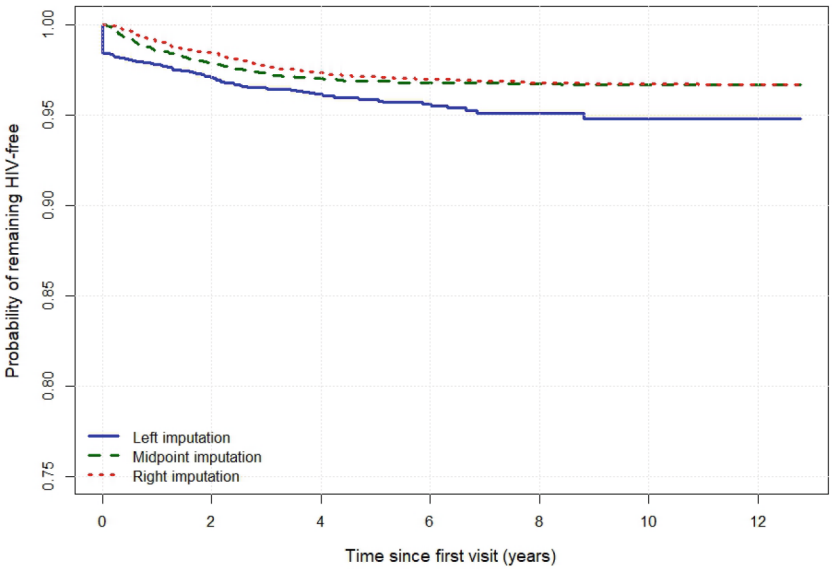

Fig. 3. Kaplan–Meier survival curves obtained under left, midpoint and right imputation. The vertical axis is restricted to the interval (0.75–1) to improve visualisation of the differences.

$p = 0.0026$). No significant differences were observed for condom use, sex with casual partners, substance use or country of birth.

Figures 4 and 5 show the estimated survival curves by education level obtained using Turnbull's nonparametric estimator. The first figure presents the full probability scale, while the second provides a zoomed view to better highlight the differences between groups. Individuals with higher educational attainment tend to present slightly higher probabilities of remaining HIV-negative over time, whereas lower education levels are associated with a somewhat faster decline in the survival function.

Although the absolute differences between curves are relatively small, the statistical test suggests that they are unlikely to be due to random variation.

4.4 Regression Analysis

The results of the univariable log-normal accelerated failure time (AFT) models are presented in Table 2. Most covariates did not show statistically significant associations with time to HIV seroconversion.

Age, condom use, substance use, previous HIV testing and country of birth were not significantly associated with time to HIV seroconversion.

In contrast, education level showed a statistically significant association with time to HIV seroconversion. Under the parametrization used in the AFT model, time ratios greater than one correspond to longer times to seroconversion. Individuals with a bachelor's degree and those with a master's or doctoral degree

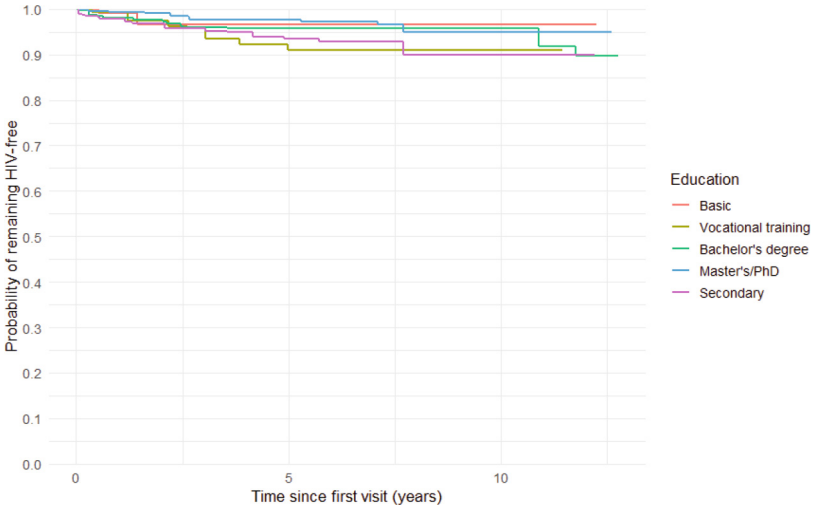

Fig. 4. Estimated survival curves for time to HIV seroconversion according to education level on the full probability scale (0–1).

presented longer estimated times to HIV seroconversion compared with individuals with secondary education.

In particular, participants with a master's or doctoral degree showed an estimated time ratio of 2.56 ($p = 0.0003$), suggesting substantially longer times to HIV seroconversion relative to the reference group. Similarly, individuals with a bachelor's degree presented a time ratio of 1.48 ($p = 0.0246$).

Among the parametric models considered, the log-normal distribution provided the best fit according to the AIC criterion, although differences with other distributions were relatively small. The corresponding hazard ratios obtained from the proportional hazards model were consistent with the direction of the effects observed under the AFT specification.

These findings are consistent with the results obtained from the comparison of survival curves.

5 Discussion

This study examined time to HIV seroconversion in a prospective cohort of men who have sex with men, explicitly accounting for the interval-censored nature of the data arising from periodic and irregular HIV testing. In this setting, the exact infection time is not observed and is only known to lie between consecutive visits, making standard survival methods based on exact or right-censored times potentially inadequate.

The comparison between Turnbull's nonparametric maximum likelihood estimator and Kaplan–Meier estimators based on imputed event times highlights the impact of ignoring interval censoring. Although the differences between

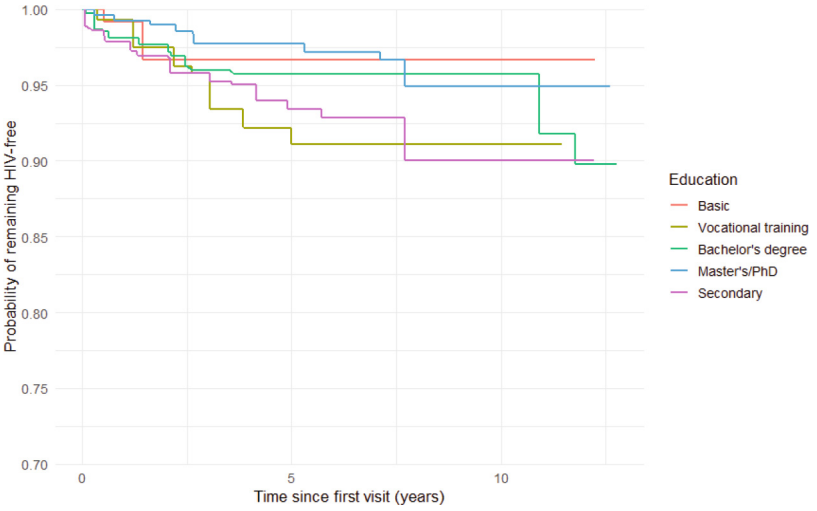

Fig. 5. Estimated survival curves for time to HIV seroconversion according to education level with a restricted vertical scale (0.7–1) to highlight differences between groups.

approaches are moderate, they are systematic: left imputation tends to underestimate survival probabilities, while right imputation may overestimate them. These results illustrate how simplifying assumptions about event timing may affect survival estimates.

The comparison of survival curves across covariates indicated that education level was the only variable showing statistically significant differences. Individuals with higher educational attainment tended to present higher probabilities of remaining HIV-negative over time. This finding was supported by the AFT regression analysis, where higher education levels were associated with longer estimated times to HIV seroconversion.

This association should, however, be interpreted with caution. Education level may act as a proxy for other factors not fully captured in the available data, such as risk perception, access to prevention strategies, testing behaviour, or healthcare access. As a result, the observed relationship does not necessarily imply a direct causal effect.

Participants with only a baseline HIV-negative test and no subsequent follow-up were excluded from the survival analysis. Although this is necessary because no follow-up interval can be defined for these individuals, it may affect the generalizability of the results if participants who did not return for testing differ systematically from those with repeated visits.

Several limitations should be considered. First, the analysis assumes that the timing of HIV testing visits is independent of the underlying time to seroconversion, conditional on observed covariates. In a voluntary testing setting, this assumption may be violated, since individuals may seek testing in response to

Table 2. Univariable log-normal AFT models for time to HIV seroconversion

Variable	Category	Estimate	TR	p-value
Age	—	-0.0086	1.009	0.3335
Condom use	Sometimes	−0.0875	1.091	0.7962
	Never	0.1294	0.879	0.6324
	Almost always	−0.2493	1.283	0.4136
	Rarely	−0.0252	1.026	0.9377
Casual partners	Yes	0.3803	0.684	0.1478
Substance use	Yes	−0.2925	1.340	0.7695
Previous HIV test	Yes	−0.0825	1.086	0.6703
Education level	≤ Primary education	−0.5906	1.805	0.2496
	Vocational training	−0.0083	1.008	0.9767
	Bachelor's degree	−0.3906	1.478	**0.0246**
	Master's/PhD	−0.9410	2.563	**0.0003**
Country of birth	Brazil	0.4297	0.651	0.1220
	Other European countries	−0.0122	1.012	0.9751
	African countries	0.6223	0.537	0.0919
	Other countries	−6.1730	479.6	0.2488

TR = time ratio = exp(−Estimate)
Reference categories: condom use = Always; casual partners = No; substance use = No; previous HIV test = No; education level = Secondary education;
country of birth = Portugal. TR > 1 indicates longer times to seroconversion.

perceived exposure or symptoms. Second, some variables contain a substantial proportion of missing or unanswered responses, which may affect the interpretation of descriptive and regression results. The small adjustment applied to zero left limits ($\varepsilon = 10^{-4}$) was only a technical device to avoid degenerate interval limits. Given its very small magnitude relative to the follow-up times, it is not expected to affect the substantive results. Finally, the relatively low number of seroconversion events limits the power to detect associations for some covariates.

Despite these limitations, the results illustrate the importance of using statistical methods that properly account for interval censoring in the analysis of HIV incidence data.

6 Conclusion

This work illustrates the practical relevance of interval-censored survival methods in the analysis of HIV seroconversion data collected through irregular testing schedules. Methods such as Turnbull's estimator and regression models for interval-censored data provide a principled alternative to approaches based on ad hoc imputation of event times.

In the Lisbon MSM cohort, education level was the only covariate showing a statistically significant association with time to seroconversion. In particular, individuals with higher educational attainment presented longer estimated times to HIV seroconversion under the AFT modelling framework, while most behavioural variables considered did not show statistically significant effects. More broadly, the results support the use of statistical methods that explicitly account for interval censoring in epidemiological studies where event times are only observed between visits.

Acknowledgements. 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. Turnbull, B.W.: The empirical distribution for interval-censored data. J. Am. Stat. Assoc. **71**, 253–264 (1976)
2. Finkelstein, D.M.: A proportional hazards model for interval-censored failure time data. Biometrics **42**, 845–854 (1986)
3. Pan, W.: A multiple imputation approach to Cox regression with interval-censored data. Biometrics **56**, 199–203 (2000)
4. Lindsey, J.C., Ryan, L.M.: Tutorial in biostatistics: methods for interval-censored data. Stat. Med. **17**, 219–238 (1998)
5. Bogaerts, K., Komarek, A., Lesaffre, E.: Survival analysis with interval-censored data: a practical approach with examples in R. CRC Press, SAS and BUGS (2017)
6. Sun, J.: The Statistical Analysis of Interval-Censored Failure Time Data. Springer (2006)
7. Kleinbaum, D.G., Klein, M.: Survival Analysis: A Self-Learning Text, 3rd edn. Springer (2012)
8. Meireles, P., Lucas, R., Martins, A., et al.: The Lisbon cohort of men who have sex with men. BMJ Open **5**, e007220 (2015)
9. Meireles, P., Lucas, R., Carvalho, C., et al.: Incident risk factors as predictors of HIV seroconversion in the Lisbon cohort of men who have sex with men: first results, 2011–2014. Eurosurveillance **20**(14), 21091 (2015)
10. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2024). https://www.R-project.org
11. Anderson-Bergman, C.: icenReg: regression models for interval censored data in R. J. Stat. Softw. **81**(12), 1–23 (2017)