AIMS Mathematics, 11(4): 9191–9209.
DOI: 10.3934/math.2026379
Received: 14 February 2026
Revised: 26 March 2026
Accepted: 31 March 2026
Published: 03 April 2026

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

Research article

On the reproducibility of survival quantile decisions beyond Greenwood-based precision

Norah D. Alshahrani*

Department of Mathematics, College of Science, University of Bisha, P.O. Box 551, Bisha 61922, Bisha, Saudi Arabia

* **Correspondence:** Email: ndmuflih@ub.edu.sa.

Abstract: Greenwood-based confidence intervals are widely used to quantify uncertainty in survival quantile estimation based on the Kaplan-Meier estimator, and narrow intervals are often interpreted as evidence of stable and reliable inference. However, such numerical precision does not directly address the reproducibility of inferential conclusions under repeated sampling. The relationship between Greenwood-based confidence-interval precision and reproducibility in survival quantile inference is investigated. Reproducibility is quantified using reproducibility probability (RP), defined as the probability that a survival quantile estimate is reproduced within a specified tolerance under repeated sampling, along with its decision-based analogue RP(D) for two-group survival comparisons. Both measures are estimated via a nonparametric bootstrap framework under a fixed study design and sample size. Extensive simulations are conducted for single-group and two-group settings under exponential, Weibull, and lognormal survival distributions, with independent and dependent right-censoring. The results show that Greenwood-based confidence interval width is not a reliable indicator of reproducibility: Narrow intervals may coexist with low RP, whereas wider intervals may be associated with high RP, depending on the distribution, censoring mechanism, and inferential target. In two-group comparisons, decision reproducibility is driven primarily by the stability of the ordering between group-specific quantiles rather than by the numerical precision of individual estimates, and under dependent censoring, decision reproducibility can be high even when confidence intervals are wide. These findings highlight a fundamental distinction between numerical precision and inferential reproducibility in survival analysis and underscore the need to assess reproducibility alongside conventional confidence-interval reporting.

Keywords: survival analysis; Kaplan-Meier estimator; Greenwood variance; reproducibility probability; survival quantiles; decision reproducibility; censoring mechanisms
Mathematics Subject Classification: 62F03, 62F12, 62G10, 62G20, 62N01

1. Introduction

Time-to-event data arise naturally in a wide range of scientific fields, including medicine, engineering, economics, and reliability analysis. A defining feature of such data is the presence of censoring, whereby the event of interest is not fully observed for all study units. Nonparametric methods play a central role in the analysis of censored survival data, with the Kaplan-Meier estimator [1] remaining the most widely used tool for estimating survival functions.

Inference based on the Kaplan-Meier estimator [1] is commonly accompanied by variance estimates and confidence intervals derived from Greenwood's formula [2]. Greenwood-based confidence intervals are routinely used to quantify estimation uncertainty for survival probabilities and related quantities, such as survival quantiles. In practice, narrow confidence intervals are often interpreted as evidence of reliable inference and stable conclusions. This interpretation implicitly assumes that estimation precision adequately reflects the robustness of inferential results under repeated sampling.

However, recent discussions in the statistical literature have emphasized that numerical precision and reproducibility are distinct concepts, with precision describing the variability of an estimator around its target and reproducibility reflecting the likelihood that similar inferential conclusions would be obtained under repeated sampling or comparable conditions. This distinction has been articulated in the broader reproducibility literature and formalized in statistical frameworks that clarify differences between reproducibility, replicability, and related notions such as precision and robustness [3–6]. A growing body of work has demonstrated that conventional measures of uncertainty, such as standard errors or confidence-interval widths, may fail to capture the stability of hypothesis-testing decisions. Despite this increasing interest in reproducibility, analogous questions have received limited attention in the context of survival analysis.

In particular, the relationship between Greenwood-based confidence interval precision and the reproducibility of survival quantile inference has not been formally investigated. Survival quantiles play an important role in applied time-to-event studies, as they provide interpretable summaries of the survival distribution and are frequently used for comparison between treatment groups. Yet, the stability of survival quantile estimates and associated comparison decisions under repeated sampling remains largely unexplored.

The issue becomes even more pronounced in the presence of dependent (informative) censoring. While the Kaplan-Meier estimator relies on an assumption of independent censoring, practical data-generating mechanisms may violate this assumption [1, 7]. Under informative censoring, standard inferential procedures may not only lose efficiency but may also fail to identify target survival quantities altogether [8, 9]. The extent to which Greenwood-based confidence intervals reflect inferential reliability under such conditions is unclear.

Beyond estimation precision, increasing attention has been directed toward the concept of reproducibility in statistical inference [3, 4]. Broadly, reproducibility refers to the probability that a similar inferential conclusion would be obtained if a study were repeated under the same design and sample size [4, 5]. While reproducibility has been extensively discussed in the context of hypothesis testing and p-values [3, 6], its role in time-to-event analysis has received comparatively little attention. In particular, it remains unclear whether commonly reported measures of precision, such as Greenwood-based confidence intervals, adequately reflect the reproducibility of survival quantile

estimates and related inferential conclusions.

This paper investigates the reproducibility of survival quantile inference beyond Greenwood-based precision. Reproducibility probability (RP) measures are introduced for both single-group survival quantile estimation and binary comparison decisions between two groups. Using nonparametric bootstrap resampling, the probability that a given inferential conclusion is reproduced under repeated sampling from the observed data is quantified. Through extensive simulation studies, the relationship between Greenwood-based confidence interval width and reproducibility is examined under both independent and dependent censoring mechanisms.

The remainder of the paper is organized as follows. Section 2 introduces RP for survival quantiles and comparison decisions. Section 3 describes the simulation design under independent and dependent censoring for single- and two-group settings. Section 4 presents and discusses the simulation results. Section 5 summarizes the findings and conclusions.

2. Reproducibility probability for survival quantile inference

2.1. Survival quantiles

Let T denote a nonnegative survival time with survival function

$$S(t) = \Pr(T > t), \quad t \geq 0.$$

For a given probability level $p_0 \in (0, 1)$, the corresponding survival quantile is defined as

$$\tau(p_0) = \inf\{t : S(t) \leq p_0\}.$$

The quantity $\tau(p_0)$ provides an interpretable summary of the survival distribution and is commonly reported in applied time-to-event studies.

Given right-censored observations (X_i, δ_i), $i = 1, \ldots, n$, where $X_i = \min(T_i, C_i)$ denotes the observed time and $\delta_i = \mathbb{I}(T_i \leq C_i)$ is the event indicator, the survival function $S(t)$ is typically estimated using the Kaplan-Meier estimator [1]. Let $t_{(1)} < \cdots < t_{(k)}$ denote the ordered distinct event times. The Kaplan-Meier estimator is given by

$$\widehat{S}(t) = \prod_{t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right),$$

where d_j is the number of events at time $t_{(j)}$ and n_j is the number of individuals at risk immediately prior to $t_{(j)}$.

The corresponding estimator of the survival quantile is defined by

$$\widehat{\tau}(p_0) = \inf\{t : \widehat{S}(t) \leq p_0\}.$$

2.2. Greenwood-based confidence intervals

Uncertainty in the Kaplan-Meier estimator $\widehat{S}(t)$ is commonly quantified using Greenwood's variance estimator, which provides pointwise variance estimates for the estimated survival probabilities [2, 7]. Based on this variance, pointwise confidence intervals for the survival function, denoted by $[L(t), U(t)]$, can be constructed using standard large-sample approximations.

Confidence intervals for the survival quantile $\tau(p_0)$ are then obtained by inverting these pointwise confidence intervals, that is, by identifying the time points at which the lower and upper confidence limits of the survival function cross the probability level p_0. The resulting Greenwood-based confidence interval for $\tau(p_0)$ is denoted by

$$[\widehat{\tau}_{\mathrm{L}}(p_0), \widehat{\tau}_{\mathrm{U}}(p_0)],$$

with width

$$W_{\mathrm{G}}(p_0) = \widehat{\tau}_{\mathrm{U}}(p_0) - \widehat{\tau}_{\mathrm{L}}(p_0).$$

In applied analyses, the interval width $W_{\mathrm{G}}(p_0)$ is frequently interpreted as a measure of estimation precision, with narrower intervals suggesting more reliable inference. However, as demonstrated in this paper, such numerical precision does not necessarily reflect the reproducibility or stability of inferential conclusions under repeated sampling.

Due to the step-function nature of the Kaplan-Meier estimator, the set

$$\mathcal{T} = \{t : L(t) \leq p_0 \leq U(t)\}$$

may be empty or consist of multiple disjoint intervals, particularly under heavy censoring or when the target quantile is not identifiable.

When $\mathcal{T}$ is empty, the quantile interval is defined to collapse to the point estimate, i.e.,

$$[\widehat{\tau}_L, \widehat{\tau}_U] = [\widehat{\tau}(p_0), \widehat{\tau}(p_0)].$$

When $\mathcal{T}$ contains multiple disjoint intervals, the quantile confidence interval is defined as $[\min(\mathcal{T}), \max(\mathcal{T})]$, which may result in widened intervals.

In this study, Greenwood confidence intervals are constructed using the plain (untransformed) variance estimator, consistent with standard Kaplan-Meier implementations. While this choice may lead to conservative quantile intervals under sparse data or heavy censoring, such behavior reflects intrinsic limitations of quantile estimation under censoring rather than a deficiency of the interval construction.

2.3. Reproducibility probability for single-group inference

While confidence interval width reflects estimation variability, it does not directly capture the stability of an inferential result under repeated sampling. In particular, a numerically precise estimate may still be unstable if the study were repeated under the same design and sample size. To address this limitation, RP is introduced as an explicit measure of inferential stability.

The RP is defined as the probability that a given inferential result would be replicated under repeated sampling from the same population, conditional on the observed data and the resampling scheme. In the context of this study, RP measures the likelihood that the estimated survival quantile remains within a predefined tolerance level ε across bootstrap samples. Unlike confidence intervals, which primarily quantify estimation uncertainty, RP captures the stability of inferential conclusions. A high RP indicates that the result is reproducible under resampling, whereas a low RP suggests that the conclusion is sensitive to sampling variability.

Let $\widehat{\tau}(p_0)$ denote the estimated survival quantile obtained from the observed dataset. Consider a hypothetical repetition of the study under the same experimental design, sample size, and censoring

mechanism. Since the underlying data-generating distribution is unknown, repeated sampling is approximated using nonparametric bootstrap resampling from the empirical distribution of the observed data.

Let $\widehat{\tau}^*(p_0)$ denote the survival quantile estimate obtained from a bootstrap sample. For a pre-specified tolerance $\varepsilon > 0$, the RP is defined as

$$\mathrm{RP} = \Pr(|\widehat{\tau}^*(p_0) - \widehat{\tau}(p_0)| \le \varepsilon \,\big|\, \text{data}),$$

where this probability is approximated by the empirical proportion of bootstrap replicates for which the inequality holds.

The tolerance ε is chosen as a fixed fraction of $\widehat{\tau}(p_0)$, ensuring scale invariance across different survival time distributions and facilitating interpretation across scenarios. Under this definition, RP represents a data-conditional measure of reproducibility and approximates the probability of obtaining a practically equivalent survival quantile estimate if the study were repeated under the same conditions.

Importantly, RP is conceptually distinct from coverage probability or confidence-interval width. Whereas Greenwood-based confidence intervals quantify estimation uncertainty for a single realization of the data, RP directly assesses the stability of the estimated survival quantile under repeated sampling. Consequently, high numerical precision, as indicated by a narrow confidence interval, does not necessarily imply high reproducibility, a phenomenon investigated in detail in subsequent sections.

2.4. Decision reproducibility in two-group comparisons

In comparative survival studies, interest often lies not only in estimating survival quantiles but also in comparing them across groups. Consider two groups, denoted by A and B, with corresponding survival quantiles $\tau_A(p_0)$ and $\tau_B(p_0)$. A binary decision is defined as

$$D = \mathbb{I}\{\tau_A(p_0) < \tau_B(p_0)\},$$

which indicates whether group A exhibits shorter survival at level p_0 than group B.

Let $\widehat{D}$ denote the observed decision based on the Kaplan-Meier quantile estimates. Using bootstrap resampling performed independently within each group, bootstrap decisions $\widehat{D}^*$ are obtained. The decision RP is then defined as

$$\mathrm{RP}(D) = \Pr(\widehat{D}^* = \widehat{D} \,\big|\, \text{data}),$$

which is approximated by the proportion of bootstrap samples yielding the same decision as the original dataset.

The quantity $\mathrm{RP}(D)$ directly quantifies the stability of the inferential conclusion, rather than the variability of individual parameter estimates. This distinction is particularly relevant when narrow confidence intervals coexist with unstable comparison outcomes.

Precision versus reproducibility: Under asymptotic normality $\widehat{\theta} \sim N(\theta, \sigma^2/n)$, one expects $\Pr(|\widehat{\theta}^* - \widehat{\theta}| \le \varepsilon) \approx 2\Phi(\varepsilon/\hat{\sigma}) - 1$. When $\varepsilon = k \cdot \text{CI width}$, RP should increase monotonically with precision.

This relationship breaks down for Kaplan-Meier quantiles due to three factors:

(1) Censoring, which creates heterogeneous information density along the survival curve;
(2) The step-function form of the Kaplan-Meier estimator, which induces discrete jumps and may produce artificial clustering of quantile estimates;
(3) Quantile inversion, which applies a non-linear transformation to the sampling behavior of the survival estimator.

These irregularities explain the simulation results where narrow confidence intervals coexist with low reproducibility, particularly for upper-tail quantiles under heavy censoring.

2.5. Bootstrap approximation of reproducibility

The RPs considered in this study are estimated using a nonparametric bootstrap resampling scheme. Since the true data-generating distribution is unknown, repeated sampling under the same experimental conditions is approximated by resampling from the empirical distribution of the observed data [10, 11].

Specifically, bootstrap samples are generated by resampling individual observations with replacement, where each observation consists of an observed time and censoring indicator pair (X_i, δ_i). This pairs bootstrap preserves the joint structure of survival times and censoring indicators and is consistent with the nonparametric nature of the Kaplan-Meier estimator [7, 12]. For each bootstrap sample, the Kaplan-Meier estimator and the corresponding survival quantile estimate are recomputed, yielding an empirical approximation to the sampling distribution of the survival quantile under repeated sampling.

RPs are then obtained by evaluating the proportion of bootstrap replicates that yield inferential results consistent with those obtained from the original dataset, according to the definitions given in the preceding subsections. All reproducibility measures reported in this study are based on a fixed number of bootstrap replicates, and the same resampling scheme is applied across all simulation scenarios to ensure comparability.

Bootstrap approximation and its limitations: The RP proposed in this study is estimated using a nonparametric bootstrap procedure, which approximates repeated sampling from the underlying population by resampling from the empirical distribution of the observed data. This approach implicitly assumes that the empirical distribution provides a reasonable approximation to the joint data-generating mechanism governing both survival times and censoring.

While this assumption is standard in nonparametric inference, it may be violated in settings with dependent censoring, where the censoring mechanism is related to the underlying survival process. In such cases, the empirical distribution may fail to adequately capture the dependence structure between event and censoring times. Consequently, the bootstrap-based RP should be interpreted as a data-conditional approximation to repeated-sampling reproducibility rather than an exact probability under the true data-generating process.

Despite this limitation, the bootstrap framework provides a practical and widely used tool for assessing inferential stability, particularly in complex settings where analytic characterization of the sampling distribution is not available.

2.6. Algorithm for RP estimation

The complete procedure for estimating RP for survival quantiles under right censoring is summarized in Algorithm 1.

Algorithm 1 Simulation procedure for RP of $\tau(p_0)$ under right censoring

Require: n, $p_0 \in (0,1)$, target censoring c, event-time model, B, α, N_{sim}
Ensure: $\widehat{\tau}(p_0)$, CI width W, and $\widehat{\text{RP}}$ for each replicate

for $r = 1, \ldots, N_{\text{sim}}$ **do**
  Generate $T_1, \ldots, T_n$ from the chosen event-time distribution
  Generate censoring times $C_1, \ldots, C_n$ independently targeting c
  Set $X_i = \min(T_i, C_i)$ and $\delta_i = \mathbb{I}(T_i \le C_i)$
  Fit Kaplan-Meier $\widehat{S}(t)$ using $\{(X_i, \delta_i)\}_{i=1}^n$
  Compute $\widehat{\tau}(p_0) = \inf\{t : \widehat{S}(t) \le p_0\}$
  Obtain Greenwood limits $(L(t), U(t))$ for $\widehat{S}(t)$
  $\mathcal{T} \leftarrow \{t : L(t) \le p_0 \le U(t)\}$
  $[\widehat{\tau}_L(p_0), \widehat{\tau}_U(p_0)] \leftarrow [\min(\mathcal{T}), \max(\mathcal{T})]$
  $W \leftarrow \widehat{\tau}_U(p_0) - \widehat{\tau}_L(p_0)$
  for $b = 1, \ldots, B$ **do**
    Resample $\{(X_i, \delta_i)\}_{i=1}^n$ with replacement
    Fit Kaplan-Meier $\widehat{S}^{*(b)}(t)$ and compute $\widehat{\tau}^{*(b)}(p_0)$
  end for
  $\varepsilon \leftarrow \alpha\, \widehat{\tau}(p_0)$
  $\widehat{\text{RP}} \leftarrow \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\left(\left|\widehat{\tau}^{*(b)}(p_0) - \widehat{\tau}(p_0)\right| \le \varepsilon\right)$
  Store $(\widehat{\tau}(p_0), W, \widehat{\text{RP}})$
end for

This algorithm is applied in the simulation study described in the following section to assess the relationship between Greenwood-based confidence interval precision and RP.

3. Simulation study

A simulation study is conducted to investigate the relationship between Greenwood-based confidence-interval precision and RP for survival quantile inference. The main goal is to assess whether narrow Greenwood-based confidence intervals reliably indicate stable and reproducible conclusions under repeated sampling, across a range of data-generating mechanisms and censoring structures.

3.1. Data-generating mechanisms

To examine reproducibility under diverse survival and censoring structures, seven data-generating scenarios are considered, combining different survival distributions, censoring mechanisms, and study designs.

Survival times are generated from several parametric distributions commonly used in time-to-event analysis. In the baseline setting, survival times T follow an exponential distribution with constant hazard rate $\lambda = 0.5$, providing a benchmark scenario with relatively simple and well-understood survival dynamics.

To examine robustness beyond the exponential case, additional simulations use Weibull and lognormal distributions. Specifically, the Weibull distribution with shape parameter 1.5 and scale parameter 2.0 allows for monotone hazard functions, extending the exponential model to settings with time-varying risk. The lognormal distribution, specified with meanlog = 0 and sdlog = 0.8, induces a non-monotone hazard function and represents more complex survival dynamics. Together, these models enable a systematic evaluation of reproducibility across survival distributions of increasing structural complexity.

Two censoring mechanisms are considered. Under independent censoring, censoring times C are generated independently of survival times T, satisfying the standard assumptions required for valid Kaplan-Meier inference. The censoring distribution is calibrated to achieve approximately 10%, 30%, and 50% censoring, representing low, moderate, and high levels commonly observed in practice.

To investigate the impact of informative censoring, a dependent censoring mechanism is considered. Survival times T are generated from the specified distributions, and censoring times C are generated conditionally on T according to

$$C = T \cdot U^{1/\gamma}, \quad U \sim \text{Uniform}(0, 1),$$

where $\gamma > 0$ controls the strength of dependence between T and C.

This construction induces informative censoring, as the censoring time C is directly linked to the underlying survival time T. Larger values of γ produce censoring that is closer to independence, while smaller values induce stronger dependence, leading to earlier censoring for individuals with shorter survival times.

The parameter γ is calibrated to achieve the target realized censoring levels:

- Low: $\gamma = 0.5$ (approximately 15% censoring)
- Moderate: $\gamma = 0.3$ (approximately 30% censoring)
- High: $\gamma = 0.2$ (approximately 50% censoring)
- Very High: $\gamma = 0.1$ (approximately 70% censoring)

This formulation provides a transparent way to control both the level of censoring and the strength of dependence, allowing systematic evaluation of informative censoring effects on reproducibility.

Simulations cover both single-group (reproducibility of $\tau(p_0)$) and two-group settings. In the two-group case, independent samples represent treatment/control groups with distinct survival distributions. Decision reproducibility RP(D) is computed as defined in Section 2, using group-specific Kaplan-Meier quantiles $\widehat{\tau}_A(p_0)$ and $\widehat{\tau}_B(p_0)$.

Greenwood-based confidence intervals are constructed using the log-log transformation and obtained by inverting Kaplan-Meier confidence bands. Their widths are used as measures of numerical precision and compared with RP(D) to assess whether Greenwood-based precision reflects the reproducibility of comparative survival conclusions.

3.2. Simulation design

The simulation study examines sample sizes $n \in \{50, 100, 200\}$, survival quantile levels $p_0 \in \{0.5, 0.8\}$ (representing central and upper-tail behavior), and target censoring proportions c.

For each scenario (n, p_0, c) and for each specified survival distribution and censoring mechanism, $N_{\text{sim}} = 150$ independent datasets are generated. Within each simulated dataset, the Kaplan-Meier estimator is computed, the survival quantile estimate $\widehat{\tau}(p_0)$ is obtained, and Greenwood-based confidence intervals are constructed by inversion of the Kaplan-Meier confidence bands. RPs are estimated using a nonparametric bootstrap procedure with $B = 300$ resamples, as described in the previous subsection.

The simulation study is designed to address the following questions:

- Does Greenwood-based confidence-interval precision reliably reflect reproducibility across different survival distributions?
- How do independent and dependent censoring mechanisms affect reproducibility?
- To what extent can high numerical precision coexist with low reproducibility in both single-group and two-group survival inference?

4. Simulation results and discussion

This section presents the results of the simulation study designed to investigate the relationship between Greenwood-based confidence interval (CI) precision and RP in survival analysis.

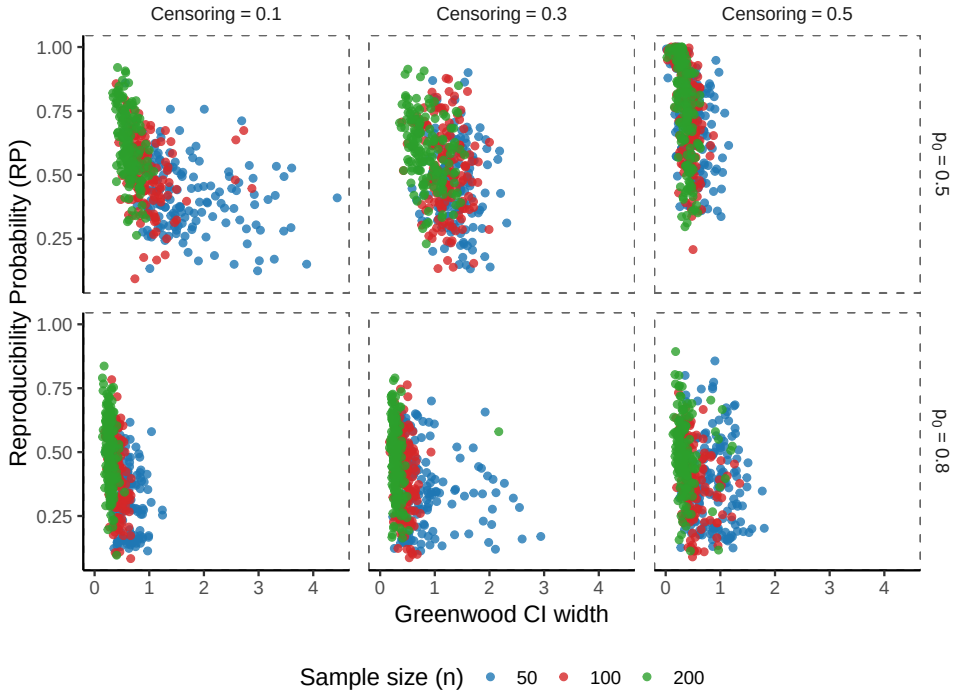

Figure 1. Relationship between Greenwood-based precision and RP for survival quantiles under an exponential model.

Figure 1 illustrates the relationship between RP and the width of the Greenwood confidence interval for the survival quantile $\tau(p_0)$ in a single-group under an exponential model. The results are shown for different sample sizes ($n = 50, 100, 200$), censoring levels ($0.1, 0.3, 0.5$), and quantile levels ($p_0 = 0.5, 0.8$). Although Greenwood-based confidence intervals generally become narrower as the sample size increases, the RP remains heterogeneous, suggesting that estimation precision alone is insufficient to ensure reproducibility of survival quantile inference. For $p = 0.5$, RP increases with sample size and approaches one as censoring increases.

The effect of censoring on Greenwood-based confidence interval width depends strongly on the location of the target survival quantile. For $p_0 = 0.5$, the quantile lies in a region of the survival curve that is typically well supported by observed events. As a result, increasing censoring primarily truncates the upper tail of the survival distribution while leaving substantial information around the median, which may lead to stable or even narrower Greenwood-based confidence intervals. In contrast, for $p_0 = 0.8$, the survival quantile lies in a region with fewer observed events and a rapidly diminishing risk set. Under increased censoring, estimation uncertainty in this region grows substantially, resulting in wider Greenwood-based confidence intervals. This contrast highlights the local nature of Greenwood-based precision and its dependence on information availability around the target quantile.

Figure 2 displays the relationship between RP and Greenwood-based confidence-interval width for the survival quantile $\tau(p_0)$ in a single group using a Weibull model. The results are shown for sample sizes $n = 50, 100, 200$, censoring levels $0.1, 0.3, 0.5$, and quantile levels $p_0 = 0.5, 0.8$.

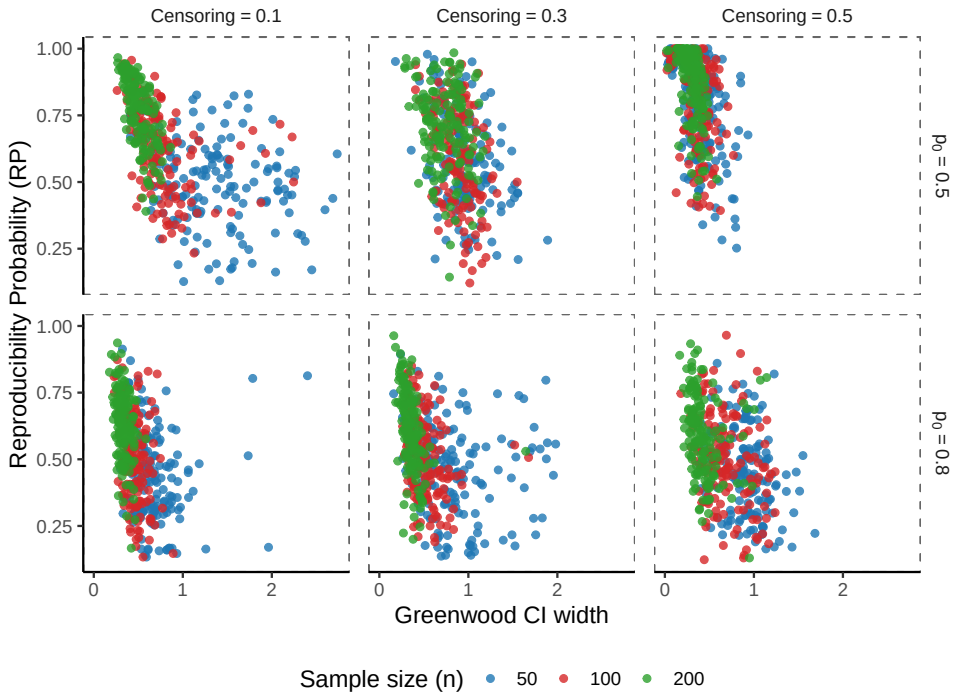

Figure 2. Relationship between Greenwood-based precision and RP for survival quantiles under a Weibull model.

Compared with the exponential model, the Weibull distribution yields somewhat wider Greenwood-based confidence intervals for the survival quantiles. This reflects the increased flexibility of the Weibull hazard function, which induces additional variability in the estimation of the Kaplan-Meier survival curve and its associated quantiles.

Despite this increase in interval width, RPs under the Weibull model are consistently slightly higher than those observed under the exponential model. Thus, although numerical precision is reduced, the estimated survival quantiles exhibit greater stability under repeated sampling.

The contrast between Greenwood-based precision and reproducibility under the Weibull and exponential models highlights the distinct roles of distributional structure in survival inference. While the Weibull model introduces additional variability due to its non-constant hazard rate, it also imposes a smoother and more structured evolution of risk over time. This structure can enhance the stability of survival quantile estimates across repeated samples, leading to higher RPs despite wider confidence intervals.

These findings further emphasize that reproducibility is not solely determined by estimation precision and that wider confidence intervals do not necessarily imply less reproducible inference.

Figure 3 displays the relationship between RP and Greenwood-based confidence-interval width for the survival quantile $\tau(p_0)$ in a single group under a lognormal model. The results are shown for sample sizes $n = 50, 100, 200$, censoring levels $0.1, 0.3, 0.5$, and quantile levels $p_0 = 0.5, 0.8$.

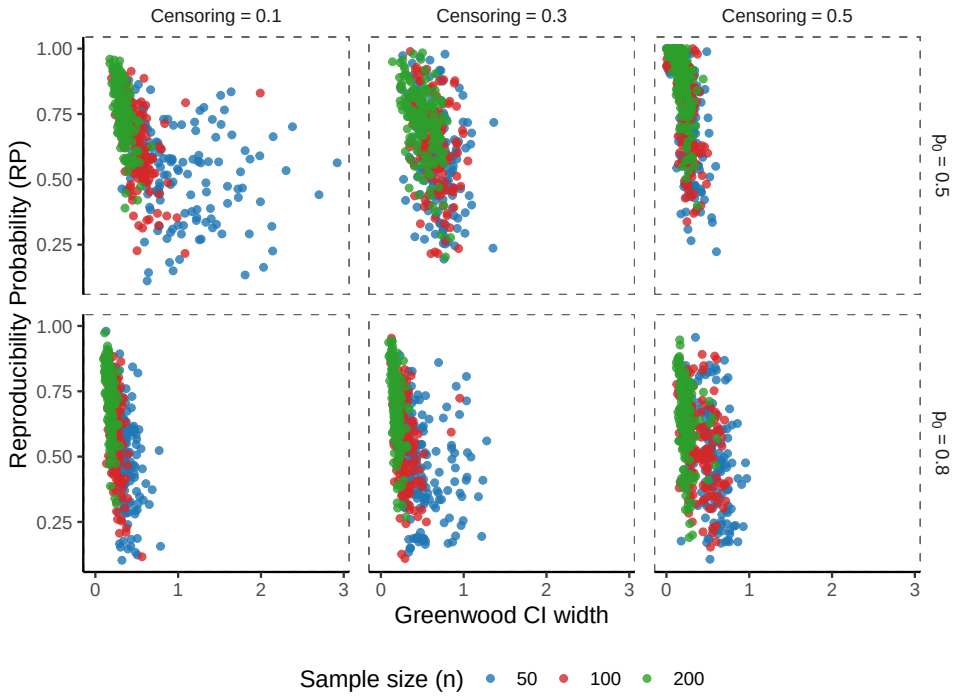

Figure 3. Relationship between Greenwood-based precision and RP for survival quantiles under a lognormal model.

Under the lognormal model, Greenwood-based confidence intervals for the survival quantiles are generally narrower than those obtained under the exponential model. This reflects the concentration of event times in the early-to-moderate portion of the survival distribution, where the Kaplan-Meier

estimator is supported by a relatively large number of observed events.

In addition, RPs under the lognormal model are consistently slightly higher than those observed under the exponential model, indicating greater stability of the estimated survival quantiles under repeated sampling. The improved Greenwood-based precision and reproducibility observed under the lognormal model can be attributed to the strong clustering of event times in regions where the target survival quantiles are evaluated. Although the lognormal distribution exhibits heavier tails and a non-monotone hazard function, the availability of substantial local information enhances both numerical precision and inferential stability.

Figure 4 illustrates the relationship between Greenwood-based confidence interval width and RP under independent censoring. For the central survival quantile ($p_0 = 0.5$), a clear positive association between numerical precision and reproducibility is observed. As censoring increases, confidence intervals tend to become narrower, and RP correspondingly increases. Wider intervals are generally associated with lower reproducibility, whereas narrower intervals coincide with greater inferential stability.

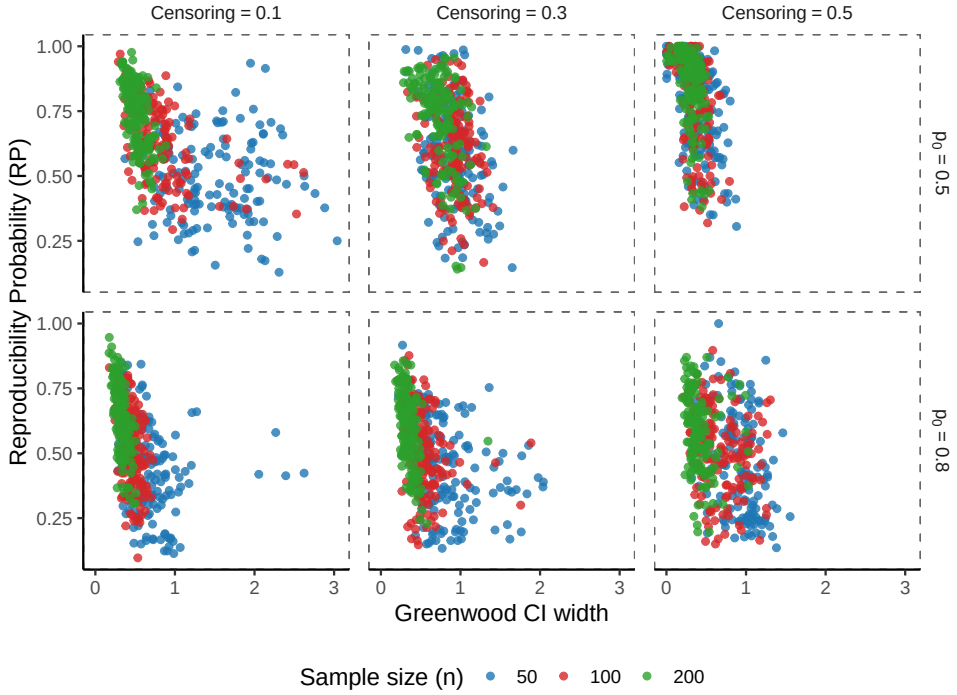

Figure 4. Relationship between Greenwood-based precision and RP for survival quantiles under independent censoring.

For the upper survival quantile ($p_0 = 0.8$), a similar but more moderate pattern is observed. For larger sample sizes ($n = 100, 200$), confidence intervals are slightly narrower, and RPs are correspondingly higher. As censoring increases, interval width expands modestly, and RP decreases slightly across sample sizes.

Generally, under independent censoring, Greenwood-based interval width and RP exhibit a broadly consistent relationship, particularly for central quantiles. These results indicate that when the non-informative censoring assumption holds, numerical precision provides a reasonable, though not perfect,

indication of inferential stability.

Figure 5 illustrates how the relationship between Greenwood-based confidence interval width and RP depends both on the target quantile and on the level of realized dependent censoring.

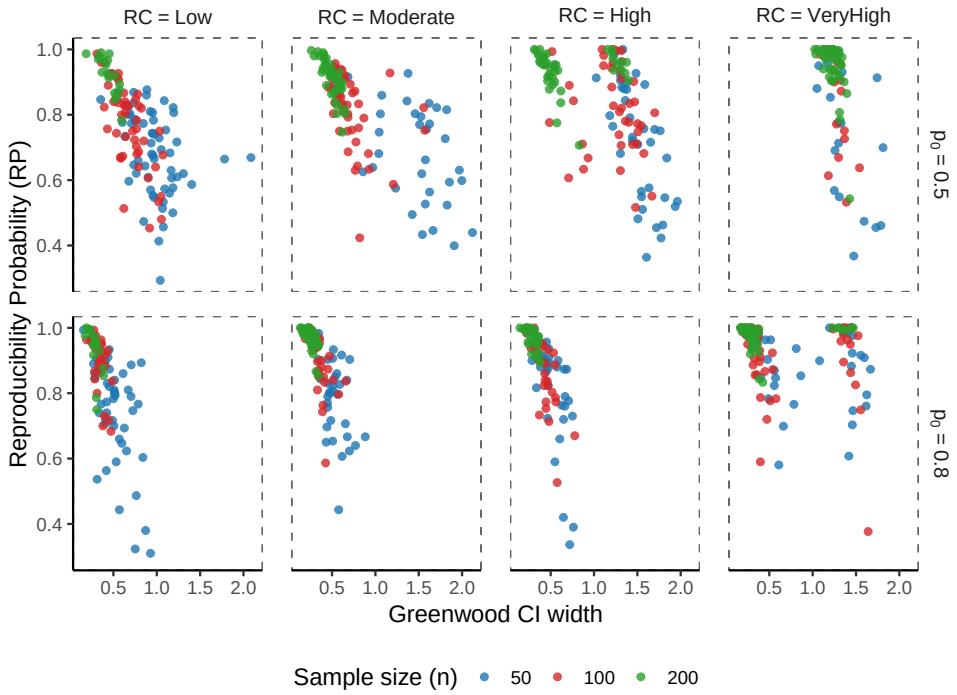

Figure 5. Relationship between Greenwood-based precision and RP for survival quantiles under dependent censoring.

For the median ($p_0 = 0.5$), confidence interval widths are relatively wide under low to high censoring, and RP displays noticeable heterogeneity, with some low values despite moderate interval widths. Interestingly, under very high censoring, confidence interval widths become narrower particularly for larger sample sizes, and RP frequently approaches one. This behavior suggests that the median is primarily determined by earlier event times, which may remain informative even when late observations are heavily censored. Hence, severe censoring does not necessarily destabilize median-based decisions.

In contrast, for the upper quantile ($p_0 = 0.8$), confidence interval widths are relatively narrow, and RP is generally high under low to high censoring. When censoring becomes very high, confidence interval widths increase substantially, reflecting weaker identification of the upper-tail survival quantile due to limited late-event information. Nevertheless, RP often remains high. This indicates that reproducibility of the decision may persist even when the quantile estimate itself is imprecisely identified. In such cases, the stepwise structure of the Kaplan-Meier estimator can produce repeated crossing at similar event times across replications, yielding stable decisions despite widened confidence intervals.

Overall, these results demonstrate that interval precision and decision reproducibility respond differently to dependent censoring, and that Greenwood-based confidence interval width alone is insufficient to characterize inferential reliability.

Figure 6 illustrates the relationship between the combined Greenwood-based confidence interval width and the decision RP, RP(D), in a two-group survival comparison under independent censoring. The horizontal axis represents the total Greenwood-based interval width for the two groups, while the vertical axis reports the probability of reproducing the binary decision that group A experiences events faster than group B.

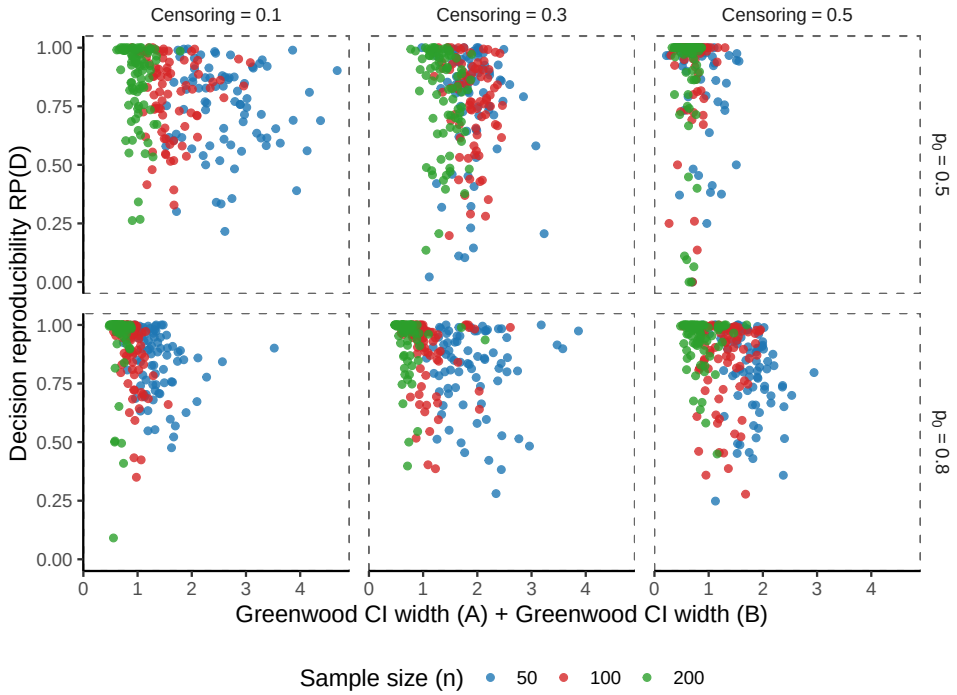

Figure 6. Decision reproducibility under independent censoring in a two-group survival comparison.

Across all censoring levels, decision reproducibility is generally high, even when the combined confidence interval width is relatively large. This indicates that numerical imprecision in the estimation of group-specific survival quantiles does not necessarily translate into instability of the comparative decision. In particular, the ordering of the estimated survival quantiles is frequently preserved under bootstrap resampling.

For $p_0 = 0.5$, confidence intervals tend to be wider under lower censoring levels, while RP(D) remains slightly high. As censoring increases, interval widths become moderately narrower, and RP(D) increases for most sample sizes, suggesting that the comparative decision becomes more stable despite changes in interval precision.

For $p_0 = 0.8$, the primary effect of censoring is reflected in increased dispersion of RP(D) values, particularly for smaller sample sizes ($n = 50, 100$). As censoring increases from 0.1 to 0.5, greater variability in RP(D) is observed for small and moderate samples, whereas larger sample sizes maintain consistently high decision reproducibility. Overall, RP(D) at $p_0 = 0.8$ is slightly higher than that observed for $p_0 = 0.5$, further illustrating that decision stability in group comparisons is not solely determined by interval width.

In the two-group setting under dependent censoring (Figure 7), results are reported only for a low nominal censoring level ($c = 0.1$), as higher censoring levels frequently prevent identification of the target survival quantiles and render the comparison decision ill-defined. Even at this low censoring level, Greenwood-based confidence intervals for the group-specific survival quantiles are relatively wide, reflecting the violation of the non-informative censoring assumption underlying Greenwood's variance estimator.

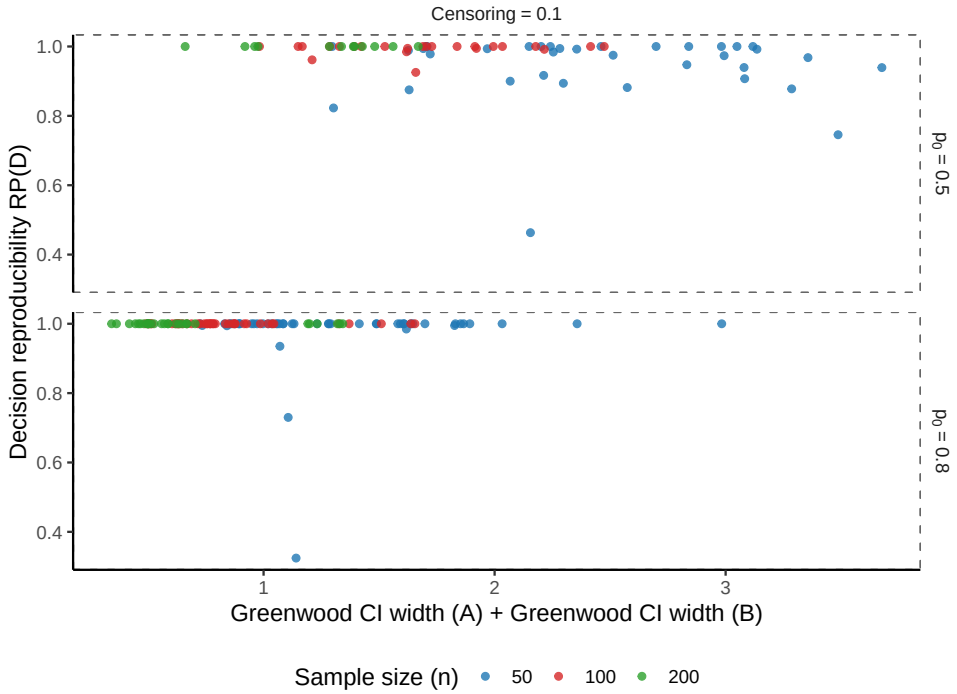

Figure 7. Decision reproducibility under dependent censoring in a two-group survival comparison.

Despite these wide confidence intervals, the decision RP, RP(D), is frequently very high and often equal to one. This apparent stability arises because the bootstrap resampling procedure repeatedly preserves the same ordering of the estimated survival quantiles across groups. Consequently, the binary comparison decision remains unchanged across bootstrap replicates, even though the individual quantile estimates themselves exhibit substantial numerical uncertainty.

Importantly, high decision reproducibility in this setting should not be interpreted as evidence of valid or unbiased inference. Rather, it reflects the structural stability of the group ordering induced by the dependent censoring mechanism, which is propagated through bootstrap resampling. Thus, RP(D) may remain close to one even when the underlying interval estimation is theoretically invalid due to violation of the non-informative censoring assumption.

It is important to distinguish between valid and reproducible inference. Validity refers to whether the assumptions underlying a statistical method are satisfied, such as the assumption of independent censoring in survival analysis. Reproducibility, on the other hand, refers to the stability of results under repeated sampling. A key insight of this work is that reproducibility does not guarantee validity. For example, an estimator may yield highly reproducible results under repeated sampling but still be biased

or invalid if model assumptions are violated. Conversely, a valid method may produce results with low reproducibility due to inherent variability in the data. Therefore, both validity and reproducibility should be considered jointly when evaluating statistical inference.

It is also important to note that the bootstrap procedure used in this study relies on the assumption that the empirical distribution adequately represents the underlying data-generating mechanism. This assumption may be violated under dependent censoring, where the censoring mechanism is related to the survival process. In such settings, the standard nonparametric bootstrap may fail to capture the true dependence structure, potentially leading to biased or misleading estimates of RP. Consequently, RP values should be interpreted with caution in the presence of dependent censoring, and developing methods that more appropriately address reproducibility under dependent censoring remains an important direction for future research.

The choice of ε plays a key role in defining the RP. In this study, ε is selected as a fixed proportion of the estimated quantile to ensure scale invariance; however, this choice is somewhat arbitrary. To assess the robustness of the proposed measure, a sensitivity analysis is conducted over multiple values of ε. The results indicate that, although the numerical values of RP change with ε, the overall qualitative conclusions remain consistent: Scenarios with low reproducibility under one choice of ε tend to remain low under alternative choices. These findings suggest that the proposed framework is robust to moderate variations in ε, although careful selection of ε may be warranted depending on the application context.

The reproducibility measures proposed in this work are defined operationally through the bootstrap and are studied here primarily via simulation. A formal large-sample theoretical framework (e.g., consistency or asymptotic distributions) for the estimators of RP and RP(D) is not developed in this study. Establishing such properties would require additional technical conditions on the data-generating mechanism and the bootstrap scheme, and lies beyond the scope of the present paper. Nonetheless, simulation results across a wide range of designs provide empirical support for the stability and interpretability of the proposed measures, and a rigorous asymptotic treatment represents an important avenue for future research.

4.1. Real data illustration

To illustrate the practical relevance of the proposed RP, the methodology is applied to a real-world survival dataset. Specifically, the well-known lung cancer dataset from the `survival` package in R is considered, which has been widely used in the survival analysis literature.

The dataset contains survival times and censoring indicators for patients with advanced lung cancer. The Kaplan-Meier estimator is used to estimate the survival function, and survival quantiles are obtained for $p_0 = 0.5$ and $p_0 = 0.8$.

To assess reproducibility, a nonparametric bootstrap procedure with $B = 300$ resamples is implemented. For each bootstrap sample, the survival quantile $\widehat{\tau}(p_0)$ is re-estimated, and the reproducibility probability is computed as

$$\mathrm{RP} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left(\left|\widehat{\tau}^{*(b)}(p_0) - \widehat{\tau}(p_0)\right| \le \varepsilon\right),$$

where ε is taken as a fixed proportion of the estimated quantile.

Table 1 presents the estimated survival quantiles, Greenwood-based confidence interval widths, and RPs for the lung dataset. For the median survival ($p_0 = 0.5$), the confidence interval is relatively wider, and the corresponding RP is higher. In contrast, for the upper quantile ($p_0 = 0.8$), the confidence interval appears slightly narrower, while the RP is lower.

Table 1. Greenwood-based confidence interval width and RP for survival quantiles in the lung dataset.

p_0	$\widehat{\tau}(p_0)$	CI width	RP
0.5	310	0.139	0.773
0.8	145	0.105	0.663

This pattern indicates that narrower confidence intervals do not necessarily imply greater reproducibility. Although Greenwood-based precision suggests improved numerical accuracy for the upper quantile, the corresponding RP reflects reduced stability of the estimated quantile under resampling. This empirical finding is consistent with the simulation results and highlights the importance of complementing classical precision measures with reproducibility-based assessments.

For the two-group analysis, patients are stratified by sex, and survival quantiles are estimated separately for each group. The decision of whether one group experiences events faster than the other is evaluated, and its reproducibility is assessed via bootstrap resampling.

Table 2 presents the estimated survival quantiles for the two groups, along with the corresponding decision and its RP. The results indicate that group A exhibits shorter survival times than group B at the median level, leading to the decision $\widehat{\tau}_A(p_0) < \widehat{\tau}_B(p_0)$.

Table 2. Decision reproducibility for survival quantiles in the lung dataset (grouped by sex).

p_0	$\widehat{\tau}_A$	$\widehat{\tau}_B$	Decision	RP(D)
0.5	269	371	A < B	1

The RP of the decision is equal to one, indicating that the ordering between the two groups is perfectly stable under bootstrap resampling. In this case, the inferential conclusion is highly reproducible.

However, this result also illustrates that reproducibility reflects the stability of the decision rather than its statistical precision. Even when confidence intervals for the group-specific quantiles may be relatively wide, the decision itself can remain highly stable if the separation between the groups is sufficiently large. This further emphasizes the distinction between numerical precision and decision reproducibility highlighted throughout this study.

5. Conclusions

This study examines the relationship between Greenwood-based confidence interval precision and reproducibility in survival quantile inference. By introducing RP and its decision-based analogue, RP(D), as explicit measures of inferential stability, numerical precision and reproducibility are shown to capture fundamentally distinct aspects of uncertainty in survival analysis.

Across a wide range of simulation scenarios, including multiple survival distributions, censoring mechanisms, and study designs, Greenwood-based confidence interval width was shown to be an unreliable proxy for reproducibility. In single-group settings, narrow confidence intervals did not necessarily correspond to stable survival quantile estimates, while wide intervals could coexist with high reproducibility. These phenomena were particularly pronounced under dependent censoring, where violations of the non-informative censoring assumption undermined both the interpretation of Greenwood-based precision and the validity of inference.

In two-group survival comparisons, decision reproducibility was found to depend primarily on the stability of the ordering between group-specific survival quantiles rather than on the numerical precision of the individual estimates. Under independent censoring, narrower confidence intervals were often associated with higher decision reproducibility, particularly for central quantiles. In contrast, under dependent censoring, decision reproducibility could reach one even when confidence intervals were wide, reflecting deterministic stability induced by the data-generating mechanism rather than genuine inferential reliability.

These results emphasize that reproducibility should not be inferred from confidence interval width alone. Greenwood-based confidence intervals remain useful for quantifying local estimation uncertainty, but they do not capture the stability of inferential conclusions under repeated sampling. Incorporating reproducibility measures alongside conventional precision metrics can provide a more complete and transparent assessment of reliability in survival analysis.

Future work may extend this framework to real-world survival data, alternative resampling schemes, and population-level reproducibility measures, further clarifying the role of reproducibility in time-to-event inference.

Use of Generative-AI tools declaration

The author declares he has not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The author is thankful to the Deanship of Graduate Studies and Scientific Research at University of Bisha for supporting this work through the Fast-Track Research Support Program.

Conflict of interest

The author declares no conflicts of interest in this paper.

References

1. E. L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, *J. Am. Stat. Assoc.*, **53** (1958), 457–481. https://doi.org/10.1080/01621459.1958.10501452
2. M. Greenwood, The natural duration of cancer, In: *Reports on Public Health and Medical Subjects*, Ministry of Health, 1926.

3. S. N. Goodman, A comment on replication, *p*-values and evidence, *Stat. Med.*, **11** (1992), 875–879. https://doi.org/10.1002/sim.4780110705
4. J. Shao, S. C. Chow, Reproducibility probability in clinical trials, *Stat. Med.*, **21** (2002), 1727–1742. https://doi.org/10.1002/sim.1177
5. D. De Martini, Reproducibility probability estimation for testing statistics, *Stat. Probab. Lett.*, **78** (2008), 1056–1061. https://doi.org/10.1016/j.spl.2007.09.064
6. D. D. Boos, L. A. Stefanski, P-value precision and reproducibility, *Am. Stat.*, **65** (2011), 213–221. https://doi.org/10.1198/tas.2011.10129
7. T. R. Fleming, D. P. Harrington, *Counting processes and survival analysis*, New York: John Wiley & Sons, 1991. https://doi.org/10.1002/9781118150672
8. P. K. Andersen, Ø. Borgan, R. D. Gill, N. Keiding, *Statistical models based on counting processes*, New York: Springer-Verlag, 1993. https://doi.org/10.1007/978-1-4612-4348-9
9. M. A. Hernán, J. M. Robins, *Causal Inference: What If*.
10. B. Efron, Bootstrap methods: Another look at the jackknife, *Ann. Statist.*, **7** (1979), 1–26. https://doi.org/10.1214/aos/1176344552
11. B. Efron, R. J. Tibshirani, *An introduction to the bootstrap*, New York: Chapman & Hall, 1994. https://doi.org/10.1201/9780429246593
12. A. C. Davison, D. V. Hinkley, *Bootstrap methods and their application*, Cambridge University Press, 1997. https://doi.org/10.1017/CBO9780511802843

AIMS Press