Received: 21 October 2019 | Revised: 22 June 2020 | Accepted: 28 June 2020

DOI: 10.1002/sim.8699

Check for updates

RESEARCH ARTICLE

Statistics in Medicine WILEY

Restricted mean survival time for interval-censored data

Chenyang Zhang[1] | **Yuanshan Wu**[2] | **Guosheng Yin**[1]

[1]Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China

[2]School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, China

Correspondence
Guosheng Yin, Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China.
Email: gyin@hku.hk

Restricted mean survival time (RMST) evaluates the mean event-free survival time up to a prespecified time point. It has been used as an alternative measure of treatment effect owing to its model-free structure and clinically meaningful interpretation of treatment benefit for right-censored data. In clinical trials, another type of censoring called interval censoring may occur if subjects are examined at several discrete time points and the survival time falls into an interval rather than being exactly observed. The missingness of exact observations under interval-censored cases makes the nonparametric measure of treatment effect more challenging. Employing the linear smoothing technique to overcome the ambiguity, we propose a new model-free measure for the interval-censored RMST. As an alternative to the commonly used log-rank test, we further construct a hypothesis testing procedure to assess the survival difference between two groups. Simulation studies show that the bias of our proposed interval-censored RMST estimator is negligible and the testing procedure delivers promising performance in detecting between-group difference with regard to size and power under various configurations of survival curves. The proposed method is illustrated by reanalyzing two real datasets containing interval-censored observations.

KEYWORDS

interval censoring, nonparametric estimator, perturbation resampling, restricted mean survival time, two-sample test

1 | INTRODUCTION

In randomized clinical trials with time-to-event observations, one primary aim is to evaluate the benefit of new medical treatment in comparison with the benchmark, that is, assessing the survival difference between two groups. The hazard ratio (HR) is frequently used to quantify the between-group survival difference with right-censored data under the proportional hazards (PH) assumption, that is, the ratio of two hazard functions is constant over time. However, the HR is no longer an accurate and interpretable measure if the PH assumption does not hold. Distribution-free hypothesis testing procedures for detecting the difference among survival curves have also been extensively studied for right-censored data, such as the log-rank test and the generalized Wilcoxon-Mann-Whitney (WMW) test.[1] Although these tests can be used to assess whether there is any difference among survival curves, they cannot quantify the treatment effects. Recently, a model-free and clinically meaningful measure, the so-called restricted mean survival time (RMST), has attracted much attention in both the statistical and medical communities.[2-4] Evaluating the mean event-free survival time up to a prespecified time point, the RMST provides a global summary of survival information. Furthermore, the RMST difference

 © 2020 John Wiley & Sons, Ltd.

between two groups can be interpreted as survival gain on average during the follow-up period up to the prespecified time point, which is explicit and valuable for assessing the between-group difference.

As a linear functional of the survival function, the estimate of the RMST with right-censored survival time can be directly obtained by plugging in the Kaplan-Meier (KM) estimator.[5] Nevertheless, another type of censoring named interval censoring appears often when subjects cannot be continuously examined, leading to observations of an interval instead of an exact time point. In such cases, subjects are observed at several discrete time points rather than being monitored continuously during the follow-up period. Thus, the exact event time T of the specified survival endpoint is not available, and all we know is that the event has not happened up to time L, but has occurred by time R, that is, T lies in the interval $(L,R]$ between two consecutive examinations. The right-censored and exactly observed samples can be treated as special cases of interval expression $(L,R]$: $R=\infty$ indicates a right-censored observation; the event time is exactly observed when $L=R-$, that is, the interval $(L,R]$ shrinks to point R. A common case of interval censoring in clinical trials is that patients are examined periodically. For example, in the Breast Cosmesis Study (BCOS),[6] patients were supposed to visit the clinic every 4 to 6 months, and thus the survival endpoint can only be detected discretely at several follow-up examination times. Model-based approaches are routinely adopted to analyze interval-censored survival data,[6-8] which, however, often rely upon certain model assumptions, for example, the PHs model[6] and the proportional odds model.[8] There are few studies on nonparametric and model-free metrics for summarizing interval-censored survival information and quantifying the between-group survival difference.[9] Given n interval-censored observations $\{(L_i,R_i]\}_{i=1}^n$, Peto[10] divided the time axis into a set of disjoint intervals $\{(s_{j-1},s_j]\}_{j=1}^m$, where the points $\{s_j\}_{j=0}^m$ are $m+1$ unique ordered elements from $\{(L_i)_{i=1}^n,(R_i)_{i=1}^n\}$. Each observed interval $(L_i,R_i]$ can be formulated as a finite union of these disjoint intervals. Unlike the KM estimator for right-censored data, the nonparametric maximum likelihood estimator (NPMLE) of the survival function with interval-censored data can only be estimated up to the boundary points of each interval, and its behaviors within these intervals $\{(s_{j-1},s_j)\}_{j=1}^m$ can be versatile. Treating interval-censored observations as right-censored is a naive but commonly used approach to circumventing ambiguity in estimation of the survival function. One recent example is a study on laser peripheral iridotomy,[11] which examined patients on the follow-up visits at 2 weeks, 6, 18, 36, 54, and 72 months, leading to interval-censored observations. Their statistical analysis was conducted on the right-censored approximation by assuming that events would only happen at these discrete visits. However, naively treating interval-censored observations as right-censored may introduce bias and underestimate the variance, which potentially results in false positive findings. Employing the linear smoothing technique to overcome the ambiguity caused by interval censoring, we develop an estimator for the survival function and propose a new model-free measure for the interval-censored RMST. We further construct a hypothesis testing procedure to evaluate the survival difference between two groups based on the interval-censored RMST estimator. Compared with existing testing procedures with interval-censored data, our procedure is not restricted by any model assumption, and can explicitly quantify the treatment benefit as well as the between-group difference with direct and meaningful interpretations. The rest of this article is organized as follows. In Section 2, we introduce the literature on interval-censored data and propose a nonparametric estimator for the RMST with interval-censored observations as well as a hypothesis testing procedure for assessing survival difference. Section 3 presents the results of simulation studies. We illustrate our interval-censored RMST with two real datasets in Section 4 and conclude the article in Section 5.

2 | INTERVAL-CENSORED RMST

2.1 | Interval censoring

Interval-censored survival data occur naturally in medical studies with periodic follow-ups and examinations. Let $T_i, i=1,\ \dots\ ,n$, be the event times with survival function $S(\cdot)$, the interval-censored observations of $\{T_i\}_{i=1}^n$ are collections of intervals $\{(L_i,R_i]\}_{i=1}^n$ for which the likelihood function has the form

$$L(S)\propto\prod_{i=1}^{n}\Pr\left(T_i\in(L_i,R_i]\right)=\prod_{i=1}^{n}\{S(L_i)-S(R_i)\}.$$

For right-censored survival data, the NPMLE has a product-limit closed form.[5] However, there is no explicit formula for the NPMLE in interval-censored cases and thus several iterative algorithms have been proposed for deriving the NPMLE. Following the formulation of the likelihood function in Peto,[10] Turnbull[12] developed a self-consistency approach as a

special case of the EM algorithm to calculate the NPMLE. The self-consistency algorithm is built upon the condition that the behaviors of S in each disjoint interval would not influence the likelihood function $L(S)$, and solving the NPMLE is equivalent to computing the optimal probability assigned to each disjoint interval. Groeneboom and Wellner[13] proposed an iterative convex minorant (ICM) algorithm to speed up the convergence of Turnbull's method. Wellner and Zhan[14] introduced EM-ICM, a hybrid algorithm combining the self-consistency and ICM methods.

On the other hand, comparison of survival curves in the presence of interval-censored observations has also been investigated, mainly relying on certain model assumptions. Specifically, Finkelstein[6] formulated a score test for interval-censored data based on a continuous grouped PHs model. Sun[15] derived a nonparametric test mimicking the usual log-rank test when the observed failure times are on a discrete scale. Fay[8] showed that the generalized WMW test with right-censored data proposed by Peto and Peto[1] can be applied to interval-censoring cases under the proportional odds model. Rather than directly conducting tests on interval-censored data, Pan[16] and Huang et al[17] treated the unobserved exact failure time as missing data and generated right-censored data using a multiple imputation approach. Fang et al[9] extended the weighted KM test introduced by Pepe and Fleming[18] to the continuous interval-censoring scenarios and derived the asymptotic properties of the proposed statistic under the null hypothesis.

2.2 | RMST with interval-censored data

The RMST of survival time T is defined as the mean of the restricted survival time $X = \min(T, \tau)$ limited to some specified time point τ,[19] which is equivalent to the area under the survival curve S from 0 to τ,

$$\text{RMST}(\tau) = E(T \wedge \tau) = \int_0^\tau S(t)dt.$$

In fact, the mean survival time $E(T)$ is an essential metric for treatment benefit, which, however, is not estimable due to the existence of censoring. As a remedy, RMST has been proposed as an alternative measure for quantifying the mean survival time and assessing the between-group treatment difference.[3,4,19,20]

It is known that the log-rank test is the locally most powerful test under the PH assumption while it incurs power loss when the PH assumption is violated. Moreover, under violation of the PH assumption, the HR derived from the Cox model is not a meaningful quantity and its clinical interpretation is questionable. Owing to the nonparametric and model-free structure, hypothesis testing based on RMST maintains its power regardless of the validity of model assumptions[21] and thus provides a robust and interpretable measure for assessing treatment effects.

The estimate of RMST can be obtained by plugging in the NPMLE $\tilde{S}(\cdot)$ of the survival function $S(\cdot)$,

$$\widetilde{\text{RMST}}(\tau) = \int_0^\tau \tilde{S}(t)dt. \tag{1}$$

However, unlike the KM estimator which provides a clear and consistent estimate of the survival function for right-censored data, information contained in the interval-censored data is not adequate to derive a unique NPMLE of the survival function. When interval censoring occurs, the NPMLE $\tilde{S}$ is a probability vector $\boldsymbol{p} = (p_j)_{j=1}^m$ representing the assigned probabilities on the disjoint intervals $\{(s_{j-1}, s_j]\}_{j=1}^m$. As a result, $\tilde{S}(\cdot)$ can be determined at the boundary points $\{s_j\}_{j=0}^m$, while its behaviors within the intervals $(s_{j-1}, s_j]$ cannot be identified. Therefore, in contrast to the KM estimator for right-censored data, the NPMLE $\tilde{S}$ is no longer a nonincreasing step function, but displays rectangles in disjoint intervals. Figure 1 shows the NPMLEs of survival curves for two treatment groups in the BCOS data.[6] The NPMLE $\tilde{S}$ only provides the estimated lower and upper bound of S in the shape of rectangles, and uncertainties within these rectangles cause ambiguity in the estimation of RMST(τ).

To deal with the ambiguities in the NPMLE $\tilde{S}$, we consider a linear smooth $\hat{S}$ by connecting the diagonal lines of all rectangles (eg, the dashed line in Figure 1), that is, we assume the event happens with equal probability within each rectangle, mimicking Pan[16] and Geskus.[22] Thus, we can replace $\tilde{S}$ by the smoothed piecewise linear estimator $\hat{S}$ in (1) to obtain the estimate of RMST as the area under the survival curve by connecting diagonal lines of all rectangles,

$$\widehat{\text{RMST}}(\tau) = \int_0^\tau \hat{S}(t)dt, \tag{2}$$

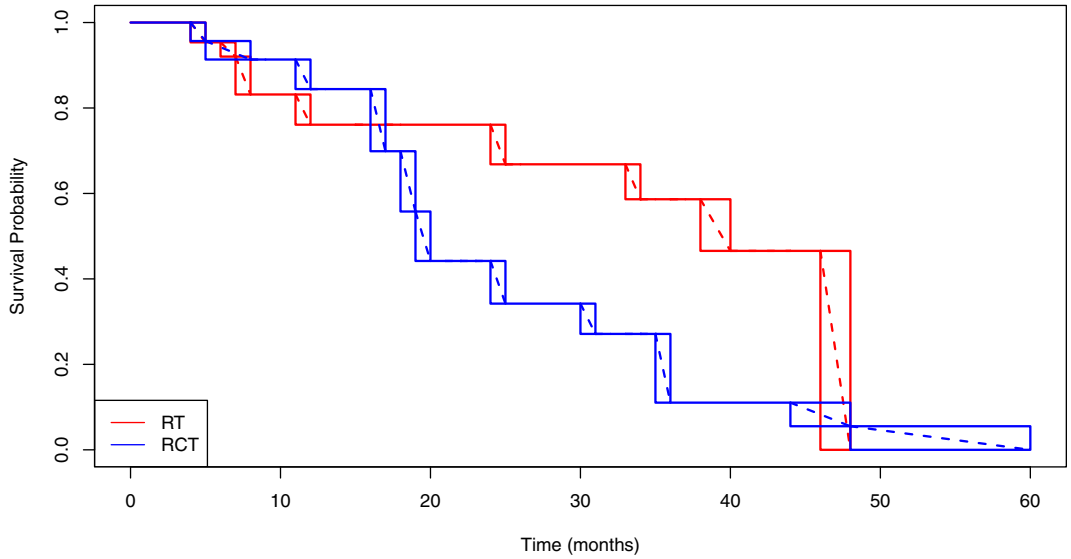

FIGURE 1 NPMLEs of survival curves for the radiation therapy only group (RT in red) and the radiation therapy plus adjuvant chemotherapy group (RCT in blue) from the Breast Cosmesis Study with dashed diagonal lines in the rectangles. NPMLE, nonparametric maximum likelihood estimator [Color figure can be viewed at wileyonlinelibrary.com]

where τ is the specified time point. Groeneboom and Wellner[13] restricted the NPMLE $\tilde{S}$ to be piecewise constant with jumps only at the points $\{s_j\}_{j=0}^{m}$ of the disjoint intervals and proved the uniform consistency of $\tilde{S}$ under some mild assumptions. The smoothed estimator $\hat{S}$ is also consistent since the total length of the intervals on which $\hat{S}$ and $\tilde{S}$ have different values shrinks to zero as the sample size $n \to \infty$.[23] However, unlike the $n^{1/2}$-convergence rate of the KM estimator for right-censored data, the NPMLE $\tilde{S}$ of interval-censored observations only has the $n^{1/3}$-convergence rate in ℓ_2-measure.[23] Therefore, the weak convergence of $\sqrt{n}\{\widehat{\text{RMST}}(\cdot) - \text{RMST}(\cdot)\}$ to a mean-zero Gaussian process cannot be derived as that for right-censored data.[24]

On the other hand, consider the linear functionals of S,

$$\eta(S) = \int_0^\infty w(t)dS(t), \tag{3}$$

where $w(t)$ is a given weight function. The interval-censored RMST has the form of a linear functional of S if we take $w(t) = -(t \wedge \tau)$. Geskus and Groeneboom[23,25] showed that although $\tilde{S}$ can only achieve a $n^{1/3}$-convergence rate, its linear functionals (3) can retain the usual $n^{1/2}$-convergence rate and the asymptotic normality under mild conditions,

$$\sqrt{n}\{\eta(\tilde{S}) - \eta(S)\} \xrightarrow{D} N(0, \sigma^2), \tag{4}$$

where $\xrightarrow{D}$ represents convergence in distribution and σ^2 is the asymptotic variance. Due to the consistency of both $\hat{S}$ and $\tilde{S}$, we can replace $\tilde{S}$ by $\hat{S}$ in (4), and for interval-censored data we set $w(t) = -(t \wedge \tau)$, which leads to

$$\sqrt{n}\{\widehat{\text{RMST}}(\tau) - \text{RMST}(\tau)\} = -\sqrt{n}\int_0^\infty (t \wedge \tau)d\{\hat{S}(t) - S(t)\} \xrightarrow{D} \mathcal{N}(0, \sigma^2(\tau)).$$

However, there is no explicit formula for the asymptotic variance $\sigma^2(\tau)$, which is typically estimated by a perturbation-resampling method.[2,26]

Following the testing procedure of the RMST under right-censored cases,[2] we can construct a counterpart for interval-censored data. Let $S_0(\cdot)$ and $S_1(\cdot)$ denote the survival functions of groups 0 and 1, respectively, and we consider the hypothesis test for the equivalence of two survival curves in the interval $(0, \tau]$. Define the RMST difference by

$$D(\tau) = \text{RMST}_1(\tau) - \text{RMST}_0(\tau) = \int_0^\tau \{S_1(t) - S_0(t)\}dt,$$

which can be estimated by

$$\hat{D}(\tau) = \widehat{\mathrm{RMST}}_1(\tau) - \widehat{\mathrm{RMST}}_0(\tau) = \int_0^\tau \{\hat{S}_1(t) - \hat{S}_0(t)\}dt.$$

It holds that

$$\sqrt{n}\{\hat{D}(\tau) - D(\tau)\} \xrightarrow{D} \mathcal{N}(0, \sigma_D^2(\tau))$$

where $\sigma_D^2(\tau) = \sigma_0^2(\tau)/\rho_0 + \sigma_1^2(\tau)/\rho_1$, $\rho_k > 0$ is the limit of n_k/n with $n = n_0 + n_1$ and n_k being the size of group k, and $\sigma_k^2(\tau)$ is the asymptotic variance of $\sqrt{n}\widehat{\mathrm{RMST}}_k$, $k = 0, 1$. Under the significance level α, the $100(1-\alpha)\%$ confidence interval (CI) of $\hat{D}(\tau)$ is

$$[\hat{D}(\tau) - z_{1-\alpha/2}n^{-1/2}\hat{\sigma}_D(\tau), \hat{D}(\tau) + z_{1-\alpha/2}n^{-1/2}\hat{\sigma}_D(\tau)], \tag{5}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution. We can obtain $\hat{\sigma}_D(\tau)$ via the perturbation-resampling method, for which the details are presented in Algorithm 1. The null hypothesis H_0 would be rejected if point zero is not included in the CI (5) of $D(\tau)$. The corresponding P-value takes the form of

$$P\text{-value} = 2\max\left\{\Phi\left(\frac{\hat{D}(\tau)}{\hat{\sigma}_D(\tau)/\sqrt{n}}\right), 1 - \Phi\left(\frac{\hat{D}(\tau)}{\hat{\sigma}_D(\tau)/\sqrt{n}}\right)\right\},$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

Our test statistic $\hat{D}(\tau)$ shares the same form of the integrated weighted difference of the estimated survival functions in Fang et al[9] when the weight function is a constant. Fang et al[9] constructed a two-sample test for the equivalence of two survival functions, and proved that under the null hypothesis their test statistic asymptotically follows a normal distribution and a bootstrap procedure can be used to obtain the P-value. Our goal is to first estimate the interval-censored RMST and derive its asymptotical distribution, and then for two-sample comparison the difference in RMSTs can be used for hypothesis testing.

The hypothesis testing procedure for the survival difference using the interval-censored RMST is described in Algorithm 1.

Algorithm 1. Hypothesis test using interval-censored RMST

Input: Interval-censored observations $\boldsymbol{Y}_k = \left\{\left(L_j^{(k)}, R_j^{(k)}\right]\right\}_{j=1}^{n_k}$ $(k = 0, 1)$;and the specified time point τ.

Step 1: Calculate the NPMLEs $\tilde{S}_0$ and $\tilde{S}_1$ using the EM-ICM algorithm and their linear smooth $\left(\hat{S}_0, \hat{S}_1\right)$;

Step 2: Compute the interval-censored RMST difference $\hat{D}(\tau)$;

Step 3: Approximate the asymptotic standard deviation $\hat{\sigma}_D(\tau)$ of $\hat{D}(\tau)$ by a perturbation-resampling method:

(a) For $b = 1, \ldots, B$, draw the weights $\{\omega_{kj}^{(b)}\}_{j=1}^{n_k} \overset{i.i.d.}{\sim}$ Exp(1) and reweight the original data $\boldsymbol{Y}_k$ to obtain the perturbed sample $\boldsymbol{Y}_k^{(b)}(k = 0, 1)$;

(b) Based on each $\boldsymbol{Y}_k^{(b)}(k = 0, 1)$, calculate the NPMLEs $\left(\hat{S}_0^{(b)}, \hat{S}_1^{(b)}\right)$ and their linear smooth $\left(\tilde{S}_0^{(b)}, \tilde{S}_1^{(b)}\right)$; and then obtain theRMST difference $\hat{D}^{(b)}(\tau)$;

(c) Obtain $\hat{\sigma}_D(\tau)$ as the standard deviation of $\{\hat{D}^{(b)}(\tau)\}_{b=1}^B$.

Step 4: Derive the $100(1-\alpha)\%$ confidence interval in (5).

Output: Reject the null hypothesis if the confidence interval does not contain zero.

2.3 | Choice of the time point τ

The choice of τ is an important issue for RMST as an inappropriately specified τ may lead to invalid inference. For right-censored data, Tian et al[21] pointed out that the RMST estimator using the KM curve is valid if $P(X > \tau) > 0$, where X denotes the observed time and thus τ should be smaller than the largest observed time. Under some mild conditions

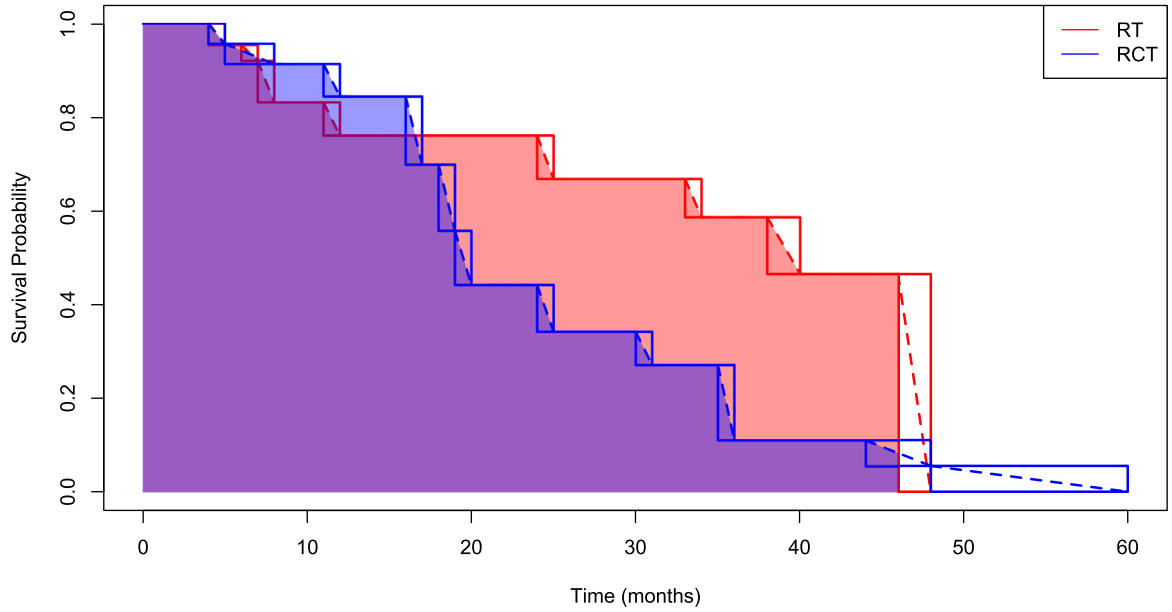

FIGURE 2 NPMLEs of survival curves for the radiation therapy only group (RT in red) and radiation therapy plus adjuvant chemotherapy group (RCT in blue) in the BCOS data with the colored areas as the estimated interval-censored RMSTs. BCOS, Breast Cosmesis Study; NPMLE, nonparametric maximum likelihood estimator; RMST, restricted mean survival time [Color figure can be viewed at wileyonlinelibrary.com]

on the distribution of censoring time, Tian et al[27] proved that the RMST estimator can still be asymptotically valid with τ equal to the largest observed time.

For interval-censored data, the asymptotic normality of the proposed RMST at time τ holds if in the interval $[0, \tau]$, the event time and censoring time distributions satisfy the conditions listed as (M1) to (M3) and (D1) to (D3) in Geskus and Groeneboom.[25] Given the interval-censored observation $(L,R]$, consider a new variable U where $U = L$ if $R = \infty$ and $U = R$ if $R < \infty$. Let h_L,h_U denote the density functions of L,U, respectively, and let f denote the density function of the event time T. To satisfy the conditions, one need to choose τ such that (i) $h_L(t) + h_U(t) > 0$; (ii) $f(t) \geq c$ for some $c > 0$ and all $t \in [0, \tau]$. A suitable choice of τ should be the one not greater than the maximum left endpoint $\max\{L_i\}_{i=1}^n$ of the observed intervals $\{(L_i, R_i]\}_{i=1}^n$.

We use the BCOS data to illustrate the empirical choice of τ in the two-sample comparison cases. Following the criterion described above, we choose $\tau_{\text{RT}} = 46$ months for the RT group and $\tau_{\text{RCT}} = 48$ months for the RCT group as the maximum left time point of the observations in RT and RCT groups, respectively. We take $\tau = \min(\tau_{\text{RT}}, \tau_{\text{RCT}}) = 46$ months such that the asymptotic properties of the proposed interval-censored RMST hold for both treatments. We calculate the RMSTs as the areas under the survival curves with diagonally connected rectangles up to 46 months, which are shown as the colored areas in Figure 2. Note that for the RCT group, $\tau = 46$ months falls in the rectangle with X-axis [44,48], while only the area under the linear smoothed survival curve in the interval [44,46] months is counted for the RMST calculation.

3 | NUMERICAL STUDIES

3.1 | Simulation settings

We first describe the data generation procedure for interval-censored cases. For each subject, the baseline examination is conducted at τ_0, and then K follow-up examinations are taken at $\{\tau_k = \tau_0 + k \times l,\ k = 1, \ldots, K\}$, where l is the time gap between two adjacent examinations and K is the number of follow-up visits. The time interval is chosen as a constant to fit the pattern of many real clinical trials, in which patients visit clinics periodically, for example, weekly or monthly. A dropout probability vector $\boldsymbol{p}_{\text{dropout}}$ of length K is included to account for the missingness of visits at each follow-up time point, and we assume that all patients have the baseline examination. To investigate how the proportion of exactly

observed data influences the performance of our interval-censored RMST, we introduce a parameter p_{exact} to control the proportion of exact observations under the partially interval-censored cases.[28] In each simulation, the study ends at time t_{end}, and the randomly interval-censored outcomes $\{(L_i,R_i), i=1, \dots, n\}$ are generated as follows. We first specify n, K, l, $\boldsymbol{p}_{\text{dropout}}$, p_{exact}, and t_{end}, and for $i=1, \dots, n$, repeat steps 1 to 4.

Step 1. Generate the baseline examination time τ_{0i} from a specified distribution, and construct the examination time grid $G_i = \{0, \tau_{0i}, \tau_{1i}, \dots, \tau_{Ki}, \infty\}$.

Step 2. Generate event time T_i from a specified distribution and the missingness of follow-ups at $\tau_{ji}, j=1, \dots, K$, according to the dropout probabilities $\boldsymbol{p}_{\text{dropout}}$. Sample the exact observation indicator ξ_i from Bernoulli(p_{exact}).

Step 3. If $\xi_i = 1$, let $L_i = T_i - \epsilon$ and $R_i = T_i$, where ϵ is a small positive number close to 0; otherwise, find $\tau_{i_1}, \tau_{i_2} \in G_i$ such that $(\tau_{i_1}, \tau_{i_2}]$ is the shortest interval covering T_i and subject i does not miss the visits at both τ_{i_1} and τ_{i_2}. The interval expression for patient i is then $(L_i = \tau_{i_1}, R_i = \tau_{i_2}]$.

Step 4. Adjust the observed interval $(L_i,R_i]$ according to the end time t_{end} of the study, such that L_i cannot exceed t_{end} and we set $R_i = \infty$ if $R_i > t_{\text{end}}$.

For simplicity, the baseline examination time τ_{0i} is sampled from Unif(0,l) in all simulations. We set $t_{\text{end}} = 1$, $\epsilon = 10^{-6}$. In practice, the interval-censored survival data are often naively treated as right-censored observations $\{X_i, \Delta_i\}$. If the event of subject i is observed within a finite interval, that is, $R_i < \infty$, its right-censored approximation would be equal to the first time detecting the occurrence of the event, that is, $X_i = R_i,\ \Delta_i = 1$. If subject i is originally right-censored with $R_i = \infty$, then we have $X_i = L_i,\ \Delta_i = 0$.

3.2 | One-sample simulation study

For one-sample analysis, we explore different parameter settings to evaluate the performance of the interval-censored RMST estimator. The probabilities of missing any visit are set to be the same for all follow-up examinations except the last one, for which the probability is doubled. Four dropout scenarios are considered,

1. None: $p_{\text{dropout},k} = 0$, for $k=1, \dots, K$;
2. Low: $p_{\text{dropout},k} = 0.1$, for $k=1, \dots, K-1$, $p_{\text{dropout},K} = 0.2$;
3. Medium: $p_{\text{dropout},k} = 0.2$, for $k=1, \dots, K-1$, $p_{\text{dropout},K} = 0.4$;
4. High : $p_{\text{dropout},k} = 0.3$, for $k=1, \dots, K-1$, $p_{\text{dropout},K} = 0.6$.

With regard to the event time distribution, we consider the Weibull distributions with various combinations of shape ξ and scale λ, and six distributions with piecewise-linear hazard functions:

(i) $h(t)=1, t\in[0,0.5); h(t)=2, t\in[0.5,1]$;
(ii) $h(t)=2, t\in[0,0.5); h(t)=1, t\in[0.5,1]$;
(iii) $h(t)=2t+1, t\in[0,0.5); h(t)=3-2t, t\in[0.5,1]$;
(iv) $h(t)=-2t+2, t\in[0,0.5); h(t)=2t, t\in[0.5,1]$;
(v) $h(t)=1, t\in[0,0.5); h(t)=2t, t\in[0.5,1]$;
(vi) $h(t)=2, t\in[0,0.5); h(t)=3-2t, t\in[0.5,1]$;

where $h(\cdot)$ denotes the hazard function. Figure 3 displays the hazard functions of six piecewise-linear hazard cases (i) to (vi). The default simulation setting in Tables 1 and 2 is $n=100, K=5, p_{\text{exact}}=0$, with a medium dropout rate and $T \sim$ Weibull(1,1). We change one configuration at a time with the other settings fixed to examine the effects of sample size n, the number of follow-up visits K, the proportion of exactly observed samples p_{exact}, the dropout rate and distributions of event time T, respectively.

We replicate 5000 simulations for each configuration and present simulation results with $\tau = 1$ and $\tau = 0.8$ in Tables 1 and 2, respectively. The column "SD" represents the sample standard deviation of $\widehat{\text{RMST}}(\tau)$, the column "ESE" is the average of estimated standard errors using the perturbation-resampling method, the column "CP" is the empirical coverage probability of the proposed CI. Our interval-censored RMST estimator is close to the ground truth, and the sample standard deviations and estimated standard errors match well under all simulation settings, which indicates the accuracy and robustness of our estimator. The empirical coverage probabilities under different simulation settings and specified time

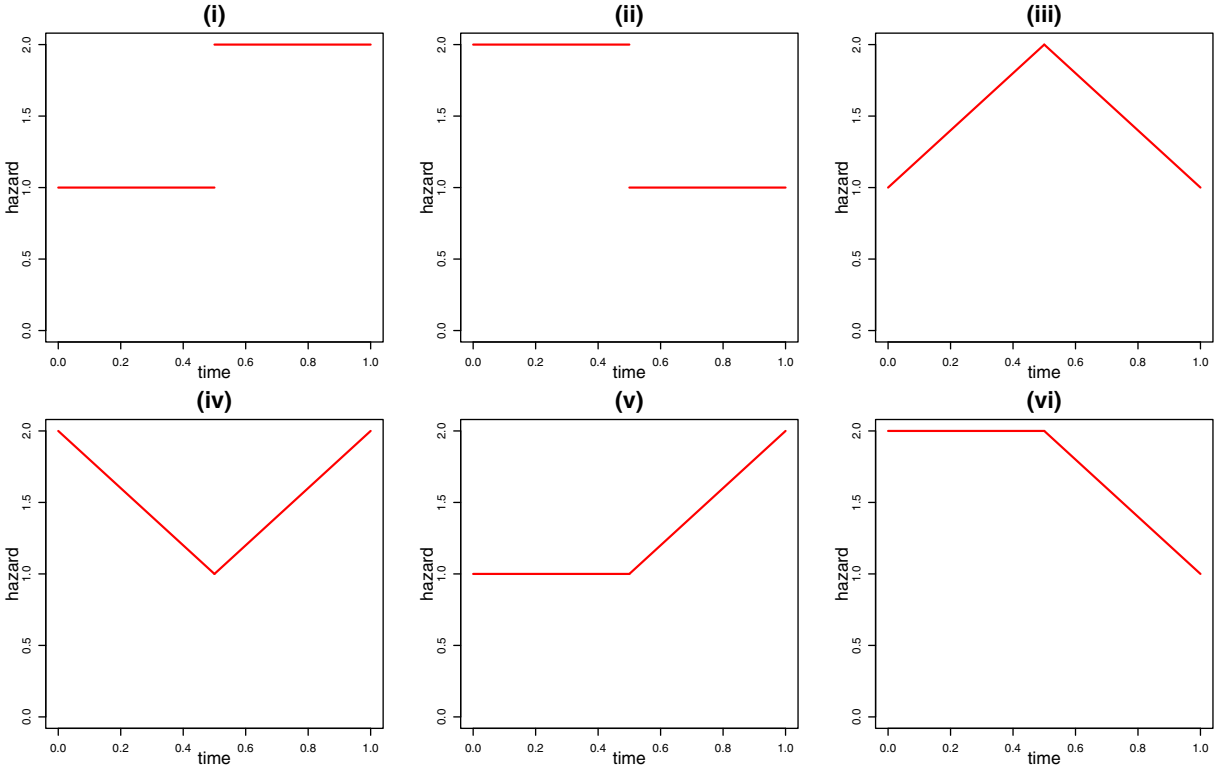

FIGURE 3 Hazard functions of six piecewise-linear hazard cases for one-sample analysis [Color figure can be viewed at wileyonlinelibrary.com]

points τ range from 0.92 to 0.95, which further confirms the consistency and asymptotic normality of the interval-censored RMST estimator.

On the other hand, if naively treated as right-censored observations, the RMST estimation is biased, especially with a small number of observed time points, severe dropout and a low proportion of fully observed samples. The bias of the RMST estimator based on naive right-censored adjustment and underestimation of standard errors result in low coverage probabilities under almost all configurations. Therefore, it is inappropriate to naively treat interval-censored observations as right-censored in the estimation of RMST.

3.3 | Two-sample simulation study

In two-sample cases, we first evaluate the performance of the estimates for RMST difference under null and alternative cases,

- $T_0, T_1 \sim$Weibull(1,1);
- $T_0 \sim$Weibull(1,1), $T_1 \sim$Weibull(0.5,1).

As shown in Table 3, both the estimates of the RMST difference using linear smoothing and the logarithm of HR are close to the ground truth, and the sample standard deviations and estimated standard errors match well. The coverage probabilities of the RMST difference and log(HR) are close to the nominal level 95% under all settings. For both the RMST difference and log(HR), their precisions exhibit insensitivity to the simulation settings, although slight reductions in the SD and ESE are observed with a larger value of K, a higher proportion of exact observations and a lower dropout rate.

Simulation studies for the two-sample RMST test are conducted under both the PH and non-PH assumptions as the settings in Pan.[16] We use the Weibull distributions with the same shape parameter ξ but distinct scale parameters λ_0, λ_1

TABLE 1 Simulation results for the interval-censored RMST estimation with $\tau = 1$ using the proposed linear smoothing method and the naive method which treats interval-censored data as right-censored based on 5000 replications

		Linear smooth				Naive right-censored			
Parameter setting	**True RMST**	$\widehat{\text{RMST}}$	**SD**	**ESE**	**CP**	$\widehat{\text{RMST}}$	**SD**	**ESE**	**CP**
n									
100*	0.63	0.63	0.037	0.036	0.9430	0.71	0.032	0.032	0.3208
200	0.63	0.63	0.026	0.026	0.9446	0.71	0.023	0.022	0.0752
400	0.63	0.63	0.018	0.018	0.9496	0.71	0.016	0.016	0.0012
K†									
3	0.63	0.63	0.037	0.036	0.9336	0.73	0.030	0.030	0.1080
5	0.63	0.63	0.037	0.036	0.9374	0.69	0.032	0.032	0.5364
10	0.63	0.63	0.036	0.036	0.9424	0.66	0.034	0.034	0.8336
20	0.63	0.63	0.035	0.036	0.9478	0.65	0.034	0.035	0.9206
p_{exact}									
0.2	0.63	0.63	0.037	0.036	0.9404	0.69	0.033	0.033	0.5078
0.5	0.63	0.63	0.036	0.036	0.9476	0.67	0.034	0.034	0.7716
1	0.63	0.63	0.036	0.036	0.9416	0.63	0.036	0.036	0.9460
Dropout rate									
None	0.63	0.63	0.037	0.036	0.9374	0.69	0.032	0.032	0.5364
Low	0.63	0.63	0.036	0.036	0.9452	0.70	0.032	0.032	0.4082
High	0.63	0.63	0.038	0.036	0.9394	0.72	0.032	0.031	0.2370
Distribution of T									
Weibull(1,0.5)	0.53	0.53	0.043	0.042	0.9418	0.60	0.038	0.038	0.4740
Weibull(1,2)	0.75	0.75	0.029	0.028	0.9298	0.82	0.023	0.023	0.1418
Weibull(0.5,1)	0.43	0.43	0.034	0.033	0.9398	0.54	0.032	0.031	0.0576
Weibull(2,1)	0.79	0.79	0.033	0.032	0.9354	0.83	0.027	0.027	0.5682
Piecewise-linear hazard (I)	0.59	0.59	0.034	0.033	0.9370	0.68	0.031	0.030	0.1328
Piecewise-linear hazard (II)	0.46	0.46	0.037	0.036	0.9396	0.56	0.033	0.033	0.1388
Piecewise-linear hazard (III)	0.53	0.53	0.035	0.034	0.9382	0.63	0.031	0.031	0.0984
Piecewise-linear hazard (IV)	0.51	0.51	0.038	0.037	0.9438	0.61	0.034	0.034	0.1936
Piecewise-linear hazard (V)	0.62	0.62	0.036	0.035	0.9414	0.70	0.032	0.031	0.2288
Piecewise-linear hazard (VI)	0.44	0.44	0.035	0.034	0.9428	0.55	0.032	0.032	0.0798

Abbreviation: RMST, restricted mean survival time.

*The default simulation setting is $n = 100, K = 5, p_{\text{exact}} = 0$, medium dropout rate and $T \sim$ Weibull(1,1).

†We set no dropout in the simulations with various K.

in the PH scenarios to fulfill the assumption of a constant HR between two treatment groups, where the HR has the form $(\lambda_0/\lambda_1)^{\xi}$. For piecewise-linear hazard functions, we consider both PH and non-PH cases, and hazard functions under non-PH cases are classified as early difference, late difference, crossing hazards, and crossing survivals:

Null cases

Weibull

(i) $T_0, T_1 \sim$ Weibull(1,0.5).
(ii) $T_0, T_1 \sim$ Weibull(1,1).
(iii) $T_0, T_1 \sim$ Weibull(1,2).

TABLE 2 Simulation results for the interval-censored RMST estimation with $\tau = 0.8$ using the proposed linear smoothing method and the naive method which treats interval-censored data as right-censored based on 5000 replications

		Linear smooth				Naive right-censored			
Parameter setting	**True RMST**	$\widehat{\text{RMST}}$	**SD**	**ESE**	**CP**	$\widehat{\text{RMST}}$	**SD**	**ESE**	**CP**
n									
100*	0.55	0.55	0.029	0.028	0.9358	0.62	0.024	0.024	0.2032
200	0.55	0.55	0.021	0.020	0.9376	0.62	0.017	0.017	0.0304
400	0.55	0.55	0.015	0.014	0.9438	0.62	0.012	0.012	0.0006
K†									
3	0.55	0.55	0.030	0.028	0.9344	0.63	0.022	0.022	0.0516
5	0.55	0.55	0.029	0.028	0.9418	0.60	0.025	0.024	0.4242
10	0.55	0.55	0.028	0.028	0.9436	0.58	0.026	0.026	0.7908
20	0.55	0.55	0.028	0.028	0.9452	0.57	0.027	0.027	0.9030
p_{exact}									
0.2	0.55	0.55	0.029	0.028	0.9382	0.60	0.025	0.025	0.4172
0.5	0.55	0.55	0.028	0.028	0.9370	0.58	0.026	0.026	0.7288
1	0.55	0.55	0.028	0.028	0.9388	0.55	0.028	0.028	0.9388
Dropout rate									
None	0.55	0.55	0.029	0.028	0.9418	0.60	0.025	0.024	0.4242
Low	0.55	0.55	0.029	0.028	0.9346	0.61	0.024	0.024	0.3046
High	0.55	0.55	0.029	0.029	0.9338	0.63	0.024	0.024	0.1326
Distribution of T									
Weibull(1,0.5)	0.45	0.45	0.034	0.034	0.9380	0.52	0.030	0.029	0.3370
Weibull(1,2)	0.66	0.66	0.021	0.020	0.9262	0.71	0.016	0.016	0.1126
Weibull(0.5,1)	0.40	0.40	0.029	0.028	0.9380	0.50	0.025	0.025	0.0292
Weibull(2,1)	0.66	0.66	0.025	0.024	0.9280	0.70	0.020	0.020	0.4600
Piecewise-linear hazard (i)	0.53	0.53	0.028	0.027	0.9356	0.61	0.023	0.023	0.0966
Piecewise-linear hazard (ii)	0.41	0.41	0.030	0.029	0.9424	0.50	0.026	0.026	0.0562
Piecewise-linear hazard (iii)	0.48	0.48	0.028	0.028	0.9414	0.57	0.024	0.024	0.0406
Piecewise-linear hazard (iv)	0.46	0.46	0.031	0.031	0.9398	0.54	0.026	0.027	0.1270
Piecewise-linear hazard (v)	0.55	0.55	0.029	0.028	0.9352	0.62	0.024	0.024	0.1724
Piecewise-linear hazard (vi)	0.40	0.40	0.029	0.029	0.9390	0.50	0.025	0.026	0.0326

Abbreviation: RMST, restricted mean survival time.

*The default simulation setting is $n = 100, K = 5, p_{\text{exact}} = 0$, medium dropout rate and $T \sim \text{Weibull}(1,1)$.

†We set no dropout in the simulations with various K.

Piecewise-linear hazard

(iv) $h_0(t) = h_1(t) = 2, t \in [0,0.5]; h_0(t) = h_1(t) = 1, t \in (0.5,1]$.

(v) $h_0(t) = h_1(t) = 1, t \in [0,0.5]; h_0(t) = h_1(t) = 2t, t \in (0.5,1]$.

Alternative cases

Proportional hazards

Weibull

(i) $T_0 \sim \text{Weibull}(1,0.5)$, $T_1 \sim \text{Weibull}(0.5,0.5)$.

(ii) $T_0 \sim \text{Weibull}(1,1)$, $T_1 \sim \text{Weibull}(0.5,1)$.

(iii) $T_0 \sim \text{Weibull}(1,2)$, $T_1 \sim \text{Weibull}(0.75,2)$.

TABLE 3 Simulation results for the two-sample RMST difference with $\tau = 1$ using the proposed linear smoothing method and the logarithm of hazard ratio (HR) with interval-censored data

		RMST difference (linear smoothing)					log(HR)			
Parameter setting	**RMST difference**	**Est.**	**SD**	**SE**	**CP**	**True log(HR)**	**Est.**	**SD**	**ESE**	**CP**
		$T_0, T_1 \sim$Weibull (1,1) (under H_0)								
Default*	0.000	0.000	0.053	0.051	0.940	0.000	0.001	0.185	0.186	0.951
K †										
3	0.000	0.000	0.053	0.051	0.938	0.000	0.003	0.184	0.182	0.950
5	0.000	0.000	0.051	0.050	0.944	0.000	0.000	0.179	0.181	0.954
10	0.000	−0.001	0.050	0.050	0.944	0.000	−0.001	0.179	0.181	0.954
20	0.000	−0.001	0.050	0.050	0.943	0.000	−0.001	0.179	0.181	0.953
p_{exact}										
0.2	0.000	0.000	0.052	0.051	0.942	0.000	0.001	0.183	0.185	0.953
0.5	0.000	−0.001	0.052	0.051	0.946	0.000	−0.005	0.181	0.183	0.954
1	0.000	−0.001	0.051	0.050	0.949	0.000	−0.005	0.178	0.179	0.945
Dropout rate										
None	0.000	0.000	0.051	0.050	0.944	0.000	0.000	0.179	0.181	0.954
Low	0.000	0.000	0.052	0.051	0.939	0.000	0.001	0.182	0.183	0.951
High	0.000	0.000	0.053	0.051	0.940	0.000	0.001	0.190	0.191	0.953
		$T_0 \sim$Weibull(1,1), $T_1 \sim$Weibull(0.5,1) (under H_1)								
Default*	0.200	0.199	0.049	0.051	0.940	0.693	0.698	0.176	0.173	0.958
K †										
3	0.200	0.200	0.049	0.050	0.939	0.693	0.705	0.173	0.169	0.957
5	0.200	0.199	0.048	0.050	0.941	0.693	0.703	0.172	0.169	0.950
10	0.200	0.199	0.048	0.049	0.941	0.693	0.702	0.171	0.168	0.950
20	0.200	0.199	0.048	0.049	0.943	0.693	0.701	0.170	0.168	0.950
p_{exact}										
0.2	0.200	0.198	0.049	0.050	0.942	0.693	0.695	0.174	0.171	0.957
0.5	0.200	0.198	0.049	0.050	0.949	0.693	0.694	0.172	0.168	0.958
1	0.200	0.198	0.048	0.049	0.944	0.693	0.690	0.168	0.167	0.950
Dropout rate										
None	0.200	0.199	0.048	0.050	0.941	0.693	0.703	0.172	0.169	0.950
Low	0.200	0.198	0.049	0.050	0.939	0.693	0.697	0.173	0.170	0.956
High	0.200	0.199	0.050	0.051	0.941	0.693	0.699	0.180	0.178	0.954

Abbreviation: RMST, restricted mean survival time.

*The default simulation setting is $n = 100, K = 5, p_{\text{exact}} = 0$, a medium dropout rate.

†We set no dropout in the simulations with various K.

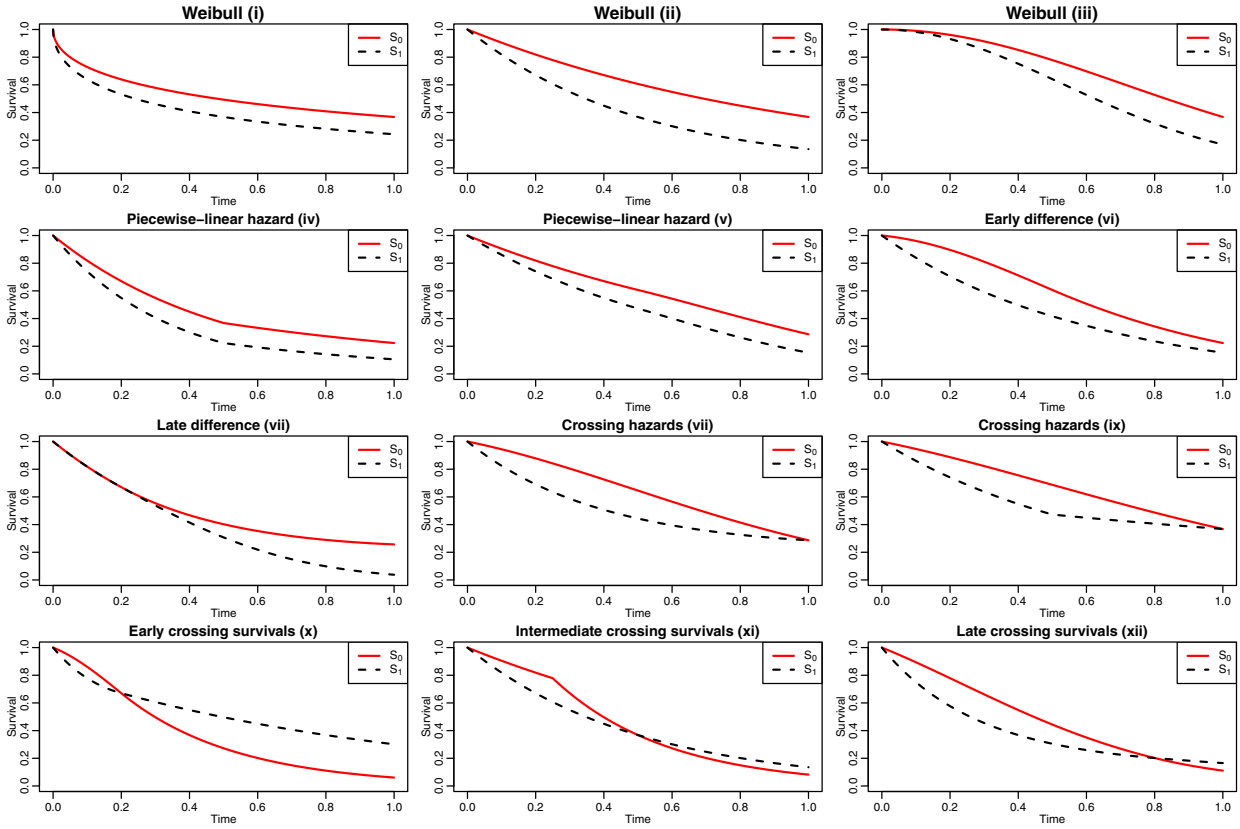

FIGURE 4 Patterns of survival functions under H_1 for two-sample tests [Color figure can be viewed at wileyonlinelibrary.com]

Piecewise-linear hazard

(iv) $h_0(t)=2$, $h_1(t)=3$, $t\in[0,0.5]$; $h_0(t)=1$, $h_1(t)=1.5$, $t\in(0.5,1]$.
(v) $h_0(t)=1$, $h_1(t)=1.5$, $t\in[0,0.5]$; $h_0(t)=2t$, $h_1(t)=3t$, $t\in(0.5,1]$.

Non-proportional hazards

Early difference

(vi) $h_0(t)=3t+0.25, h_1(t)=1.75, t\in[0,0.5]; h_0(t)=h_1(t)=t+1.25, t\in(0.5,1]$.

Late difference

(vii) $h_0(t)=h_1(t)=2, t\in[0,0.2]$; $h_0(t)=-2t+2.4$, $h_1(t)=4t+1.2$, $t\in(0.2,1]$.

Crossing hazards

(viii) $h_0(t)=1.5t+0.5, h_1(t)=-1.5t+2, t\in[0,1]$.
(ix) $h_0(t)=t+0.5, h_1(t)=1.5, t\in[0,0.5]; h_0(t)=t+0.5, h_1(t)=0.5, t\in(0.5,1]$.

Crossing survivals

(x) Early: $h_0(t)=10t+1$, $h_1(t)=-10t+3$, $t\in[0,0.2]$; $h_0(t)=3, h_1(t)=1, t\in(0.2,1]$.
(xi) Intermediate: $h_0(t)=1, h_1(t)=2, t\in[0,0.25]$; $h_0(t)=3, h_1(t)=2, t\in(0.25,1]$.
(xii) Late: $h_0(t)=2.5t+1, h_1(t)=-2.5t+3, t\in[0,0.8]$; $h_0(t)=3, h_1(t)=1, t\in(0.8,1]$.

The patterns of survival curves under alternative hypothesis H_1 are shown in Figure 4.

TABLE 4 Size and power of the interval-censored RMST test with $\tau = 1$ and 0.8, respectively, in comparison with five existing methods under the significance level $\alpha = 0.05$ and $n = 100$ for each arm based on 5000 replications

	RMST difference				Size/power						
	$\tau = 1$		$\tau = 0.8$		RMST		Log-rank-type			WMW-type	
Survival curves	True	Est.	True	Est.	$\tau = 1$	$\tau = 0.8$	Sun	Finkelstein	Log-rank	Fay	WMW
	Under H_0 (size)										
Weibull (H_0) (i)	0.00	0.00	0.00	0.00	0.0536	0.0586	0.0498	0.0500	0.0502	0.0508	0.0506
Weibull (H_0) (ii)	0.00	0.00	0.00	0.00	0.0526	0.0584	0.0488	0.0490	0.0508	0.0494	0.0500
Weibull (H_0) (iii)	0.00	0.00	0.00	0.00	0.0586	0.0598	0.0536	0.0534	0.0546	0.0498	0.0506
Piecewise-linear hazard (H_0) (iv)	0.00	0.00	0.00	0.00	0.0546	0.0580	0.0516	0.0510	0.0528	0.0510	0.0518
Piecewise-linear hazard (H_0) (v)	0.00	0.00	0.00	0.00	0.0554	0.0584	0.0532	0.0532	0.0552	0.0506	0.0510
	Under H_1 (power)										
Proportional hazards											
Weibull (H_1) (i)	0.12	0.12	0.09	0.09	0.5166	0.4798	0.5120	0.5126	0.5126	0.4890	0.4864
Weibull (H_1) (ii)	0.20	0.20	0.15	0.15	0.9792	0.9598	0.9846	0.9846	0.9848	0.9776	0.9772
Weibull (H_1) (iii)	0.12	0.12	0.08	0.08	0.8668	0.7386	0.9132	0.9130	0.9144	0.8886	0.8906
Piecewise-linear hazard (H_1) (iv)	0.12	0.12	0.10	0.10	0.7302	0.6906	0.7302	0.7326	0.7344	0.6718	0.6710
Piecewise-linear hazard (H_1) (v)	0.11	0.11	0.08	0.08	0.6510	0.5566	0.6884	0.6892	0.6914	0.6470	0.6480
Nonproportional hazards											
Early difference (vi)	0.15	0.15	0.13	0.13	0.8928	0.9192	0.6916	0.6708	0.6966	0.8828	0.8868
Late difference (vii)	0.10	0.10	0.06	0.05	0.5566	0.2872	0.8368	0.8550	0.8394	0.3526	0.3542
Crossing hazards (viii)	0.14	0.14	0.13	0.13	0.8060	0.8982	0.3598	0.3428	0.3654	0.7144	0.7148
Crossing hazards (ix)	0.13	0.13	0.12	0.12	0.7078	0.8378	0.2494	0.2404	0.2546	0.5384	0.5402
Early crossing survivals (x)	−0.15	−0.16	−0.10	−0.11	0.8870	0.7334	0.9672	0.9720	0.9678	0.7096	0.7092
Intermediate crossing survivals (xi)	0.02	0.02	0.03	0.03	0.0880	0.1516	0.0482	0.0520	0.0512	0.1348	0.1372
Late crossing survivals (xii)	0.10	0.10	0.10	0.10	0.5484	0.7388	0.1998	0.1730	0.2078	0.6492	0.6538

Abbreviations: RMST, restricted mean survival time; WMW, Wilcoxon-Mann-Whitney.

For two-sample simulation studies, we follow the data generation procedure in the former subsection and set $n_0 = n_1 = 100, k = 5, l = 0.2, p_{\text{exact}} = 0, t_{\text{end}} = 1$ with a medium dropout rate for each arm. In terms of size and power, we compare the proposed interval-censored RMST test with $\tau = 1$ and 0.8, respectively, against five existing methods: three score tests proposed by Finkelstein,[6] Sun,[15] Fay,[8] respectively, and two commonly used approaches for right-censored data (the log-rank test and generalized WMW test) after imputing the exact failure time for interval-censored observations via the multiple imputation method.[17] Note that Sun's and Finkelstein's score tests would reduce to two forms of log-rank scores in Peto and Peto[1] for right-censored data; Fay's score test and the WMW test with multiple imputation are variants of the generalized WMW test for right-censored data. We can group the five existing tests into two types: (i) log-rank-type tests including Sun's, Finkelstein's score tests and the log-rank test with multiple imputation; (ii) WMW-type tests including Fay's score test and the WMW test with multiple imputation. The size and power are reported in Table 4 based on 5000 replications.

Under H_0, the sizes of all testing procedures are close to the significance level $\alpha = 0.05$, indicating that they perform well and successfully control the type I error. For the interval-censored RMST test with $\tau = 0.8$, only part of

survival information is used, and there appears to be slight inflation in the type I error. In terms of power under H_1, the three log-rank-type tests perform slightly better than the others under the PH cases, which tallies with the results for right-censored data, as the log-rank test is the locally most powerful test when the PH assumption is satisfied. For the WMW-type tests and RMST test with $\tau = 1$, their performances are comparable, while the RMST test with $\tau = 0.8$ suffers from power loss due to partial use of survival information.

In Lee and Wang,[29] it is shown that when the PH assumption is violated, for right-censored data the log-rank test is usually more powerful in detecting late survival differences, while the generalized WMW test is more sensitive to early differences between survival curves. Our simulation results with interval-censored data display similar patterns. Three log-rank-type tests, Sun's, Finkelstein's score tests, and the log-rank test with multiple imputation, substantially outperform others in the late difference scenario.

Under the early difference and crossing survivals scenarios, the interval-censored RMST test shows great advantages over all five existing approaches in terms of power. Although the two WMW-type tests are as powerful as our interval-censored RMST test in the early difference case, their relatively poor performances in the crossing hazards scenarios indicate that the WMW-type tests might not detect the treatment difference efficiently when the two hazard functions cross. For the late difference case, our interval-censored RMST test is outperformed by the log-rank-type tests, while it can still produce much higher power than the WMW-type ones. The power of our RMST test with $\tau = 0.8$ is higher than that with $\tau = 1$ under the early difference and crossing hazards cases since the ratio between the estimated RMST difference and τ, that is, $\hat{D}(\tau)/\tau$, is larger for $\tau = 0.8$. As the difference between two survival curves in [0.8,1] is ignored, power loss incurs for the late difference case.

Moreover, we consider the cases where the two survival curves cross, respectively, at the early ($t = 0.2$), intermediate ($t = 0.5$) and late ($t = 0.8$) parts of the follow-up period [0,1]. When the survival curves cross at the intermediate follow-up period, it might be difficult to detect survival difference and all methods yield low power (less than 0.2). The WMW-type tests and the RMST test with $\tau = 0.8$ perform relatively better. The early crossing survival curves is similar to the late difference case, and the log-rank-type tests show advantages over other methods in terms of power. When the survival curves cross at the late part of the study ($t = 0.8$), the two groups also display the crossing hazards pattern in [0,0.8] and, as a result, the negative RMST difference between the two groups in [0.8,1] offsets the positive one in [0,0.8] to some extent. Thus, the proposed interval-censored RMST test with $\tau = 1$ is not as powerful as the WMW-type tests, although the RMST test with $\tau = 0.8$ still performs the best. For the three crossing survivals cases, the interval-censored RMST tests also perform reasonably well under each scenario.

Overall, our interval-censored RMST test provides an accurate measure for the RMST difference which can explicitly assess the between-group survival difference. For the PH cases, our RMST test produces slightly lower power than the log-rank-type tests which are known to be the most powerful tests under the PH assumption. For the non-PH cases, the interval-censored RMST test performs the best in the early difference and crossing hazards scenarios with regard to power, and can maintain high power under the late difference and crossing survivals case compared with the other tests.

4 | EXAMPLES

4.1 | Analysis of BCOS dataset

For illustration, we apply the proposed RMST method to two real datasets, the BCOS[6] and the HIV-1 infection dataset studied by Goedert et al.[30] The BCOS dataset recorded interval-censored observations of 94 patients who were treated by either radiation therapy plus adjuvant chemotherapy (RCT, 48 patients) or only radiation therapy (RT, 46 patients) with breast retraction as the survival endpoint. The missingness of exact event times occurred since the patients were supposed to visit the clinic every 4 to 6 months and the actual visit time varied among patients. The goal of this study was to compare the treatment effects of two therapies, and their nonparametric survival curves are displayed in Figure 1.

The P-values of the five existing hypothesis tests discussed in Section 3 are .008 (Sun), .007 (Finkelstein), .030 (Fay), .008 (the log-rank test with imputation), and .031 (the WMW test with imputation).

As all P-values are smaller than the significance level $\alpha = .05$, we can draw the conclusion that the adjuvant chemotherapy significantly increased the breast retraction rate and thus radiation therapy alone achieved long-term benefit.

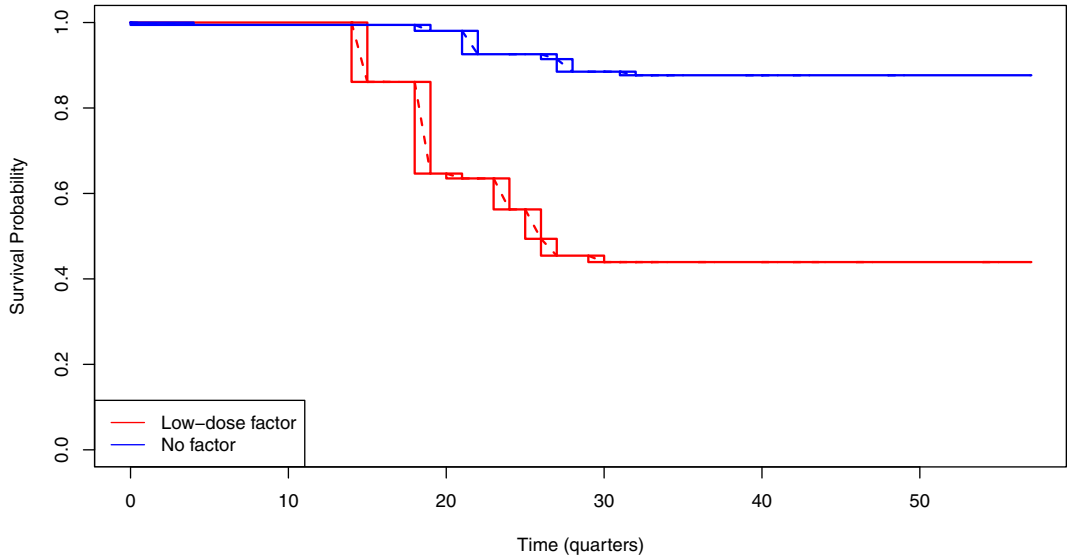

FIGURE 5 NPMLEs of survival curves for the no factor (blue) and low-dose factor (red) groups from the HIV-1 infection dataset with dashed diagonal lines in the rectangles. NPMLE, nonparametric maximum likelihood estimator [Color figure can be viewed at wileyonlinelibrary.com]

We take $\tau = 46$ months as the specified time point for the calculation of RMST, which is the minimum of the largest left endpoints of the interval-censored observations for each arm. The other five tests can only provide *P*-values, while the proposed interval-censored RMST estimator delivers a clinically interpretable quantity on the survival difference. The RMST of the RT and RCT groups are 33.04 (95% CI [28.50, 37.58]) and 23.91 (95% CI [20.33, 27.49]) months, respectively. It implies that for patients treated by RT (or RCT) the expected event-free time is 33.04 (or 23.91) months with a follow-up period of 46 months. Furthermore, the estimated interval-censored RMST difference between the RT and RCT groups is 9.14 (95% CI [3.35, 14.92]) months, which indicates that during the 46-month follow-up, on average patients in the RT group would gain 9.14 months more event-free time than those in the RCT group.

4.2 | Analysis of HIV-1 infection dataset

The second dataset arises from a multicenter trial which enrolled patients with hemophilia to investigate HIV-1 infection risk.[30] A total of 368 patients were randomized to receive no or low-dose factor VIII concentrate with respective sample sizes of 236 and 132. Figure 5 shows the two estimated survival curves for interval-censored data.

The *P*-values from five existing hypothesis testing procedures are all smaller than .001, indicating that there exists significant survival differences between patients with no or low-dose factor VIII concentrate.

We specify $\tau = 57$ quarters and for the no factor and low-dose factor groups, the estimated RMSTs are 52.79 (95% CI [51.32, 54.26]) and 36.21 (95% CI [33.05, 39.36]) quarters, respectively. The estimated RMST difference is 16.58 quarters (95% CI [13.10, 20.06]), which indicates that with a 57-quarter follow-up period, on average patients who received no factor VIII concentrate would have 16.58 quarters more infection-free time than those receiving low-dose factor VIII concentrate. Apparently, this clinically meaningful interpretation is not implied by the other five testing procedures which can only produce a *P*-value.

We observe that after 32 quarters the two survival curves are flat without any jumps. Thus, treatment benefit during a relatively short follow-up period for each group might be of interest. For the low-dose factor arm, the maximum left endpoint of the observed finite intervals is 31 quarters, and that of the no factor arm is 29 quarters. We then choose the minimum, 29 quarters, as the prespecified time τ for the interval-censored RMST analysis. For the no factor arm, the estimated RMST is 28.22 (95% CI [27.82, 28.63]) quarters, and for the low-dose factor arm the RMST is 23.90 (95% CI [22.84, 24.96]) quarters. The significant RMST difference (4.32, 95% CI [3.19, 5.46]) between the two groups demonstrates the superiority of the no factor treatment, which is consistent with the result using $\tau = 57$ quarters.

5 | CONCLUSION

The RMST, as a nonparametric and model-free estimator, provides an interpretable and global summary of treatment benefit for survival data. We have developed an estimation method for the RMST (difference) with interval-censored data and established the asymptotic properties. It is known that when the PH assumption is not satisfied, the HR, a commonly used measure for assessing the difference of treatment effects, is difficult to interpret, while the RMST difference remains clinically meaningful and interpretable regardless of any model assumptions.[3,31]

We focus on the RMST estimation with case II interval-censored survival data, while the proposed method can be adapted to estimate the RMST with case I interval-censored survival data (also known as current status data). For the current status data, a closed form is available for the NPMLE of the survival function, calculated by the max-min formula for an isotonic regression.[32,33] Huang and Wellner[34] established ℓ_2-consistency and the $n^{1/3}$-convergence rate for the NPMLE of current status data, while its linear functionals maintain asymptotic normality at the $n^{1/2}$ order with an explicit formula for the asymptotic variance. Thus, the estimation and hypothesis testing procedures for the RMST with current status data can be developed along the lines.

Rather than using the linear smoothing technique as a solution to circumvent ambiguity in the survival curve, Turnbull[12] and Groeneboom and Wellner[13] assumed a discrete scale of event time T, that is, events would only happen on the points $\{s_j\}_{j=1}^m$. The survival curve then has a shape of a nonincreasing step function. Apparently, this strategy would result in an over-optimistic estimate of the survival curve, and the discrete event times may not be in accordance with real cases where events occur continuously over time.

The selection of the time point τ is an important issue for RMST-based methods. Different time points may lead to different hypothesis testing results due to the varying patterns of survival difference over time and the amount of survival information ignored after the specified time point. For right-censored data, to obtain a robust and flexible estimator which can adapt to various event time distributions, Horiguchi et al[35] calculated the RMST estimates at a set of τ's and utilized the maximum of Z-statistic at each time point for hypothesis testing. A similar procedure can be developed for interval-censored data based on the RMST estimator to solve the inconsistency of hypothesis testing results with different time points.

ACKNOWLEDGEMENTS

We thank the two referees, the Associate Editor and Editor for their many constructive and insightful comments that have led to significant improvements in the article. The research was supported by a grant (grant number 17307218) from the Research Grants Council of Hong Kong.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

AUTHOR CONTRIBUTIONS

All authors contributed to the development of the concept, design, methodology, and writing of the article. Zhang conducted all numerical studies, and Yin oversaw the entire project.

ORCID

Chenyang Zhang https://orcid.org/0000-0003-1628-8638
Guosheng Yin https://orcid.org/0000-0003-3276-1392

REFERENCES

1. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. *J Royal Stat Soc Ser A (General)*. 1972;135:185-198.
2. Zhao L, Tian L, Uno H, et al. Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clin Trials*. 2012;9:570-577.
3. Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol*. 2014;32:2380-2385.

4. Yin G. *Clinical Trial Design: Bayesian and Frequentist Adaptive Methods*. New York, NY: John Wiley & Sons; 2012.
5. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53:457-481.
6. Finkelstein DM. A proportional hazards model for interval-censored failure time data. *Biometrics*. 1986;42:845-854.
7. Rabinowitz D, Tsiatis A, Aragon J. Regression with interval-censored data. *Biometrika*. 1995;82:501-513.
8. Fay MP. Rank invariant tests for interval censored data under the grouped continuous model. *Biometrics*. 1996;52(3):811-822.
9. Fang H-B, Sun J, Lee M-LT. Nonparametric survival comparisons for interval-censored continuous data. *Statistica Sinica*. 2002;12:1073-1083.
10. Peto R. Experimental survival curves for interval-censored data. *J Royal Stat Soc Ser C (Appl Stat)*. 1973;22:86-91.
11. He M, Jiang Y, Huang S, et al. Laser peripheral iridotomy for the prevention of angle closure: a single-centre, randomised controlled trial. *Lancet*. 2019;393:1609-1618.
12. Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. *J Royal Stat Soc Ser B (Methodol)*. 1976;38:290-295.
13. Groeneboom P, Wellner JA. *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Basel: Birkhäuser Verlag; 1992.
14. Wellner JA, Zhan Y. A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *J Am Stat Assoc*. 1997;92:945-959.
15. Sun J. A non-parametric test for interval-censored failure time data with application to AIDS studies. *Stat Med*. 1996;15:1387-1395.
16. Pan W. A two-sample test with interval censored data via multiple imputation. *Stat Med*. 2000;19:1-11.
17. Huang J, Lee C, Yu Q. A generalized log-rank test for interval-censored failure time data via multiple imputation. *Stat Med*. 2008;27:3217-3226.
18. Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics*. 1989;45:497-507.
19. Royston P, Parmar MKB. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol*. 2013;13:152.
20. Yuan Y, Yin G. Bayesian dose finding by jointly modelling toxicity and efficacy as time-to-event outcomes. *J Royal Stat Soc Ser C (Appl Stat)*. 2009;58:719-736.
21. Tian L, Fu H, Ruberg SJ, Uno H, Wei LJ. Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations. *Biometrics*. 2018;74:694-702.
22. Geskus RB. Methods for estimating the AIDS incubation time distribution when date of seroconversion is censored. *Stat Med*. 2001;20:795-812.
23. Geskus RB, Groeneboom P. Asymptotically optimal estimation of smooth functionals for interval censoring, Part 2. *Statistica Neerlandica*. 1997;51:201-219.
24. Zhao L, Claggett B, Tian L, et al. On the restricted mean survival time curve in survival analysis. *Biometrics*. 2016;72:215-221.
25. Geskus R, Groeneboom P. Asymptotically optimal estimation of smooth functionals for interval censoring, case 2. *Ann Stat*. 1999;27:627-674.
26. Lin D, Wei L-J, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*. 1993;80:557-572.
27. Tian L, Jin H, Uno H, et al. On the empirical choice of the time window for restricted mean survival time. *Biometrics*. 2020. https://doi.org/10.1111/biom.13237.
28. Huang J. Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Stat Sin*. 1999;9:501-519.
29. Lee ET, Wang J. *Statistical Methods for Survival Data Analysis*. New York, NY: John Wiley & Sons; 2003.
30. Goedert JJ, Kessler CM, Aledort LM, et al. A prospective study of human immunodeficiency virus type 1 infection and the development of AIDS in subjects with hemophilia. *N Engl J Med*. 1989;321:1141-1148.
31. Uno H, Wittes J, Fu H, et al. Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies. *Ann Intern Med*. 2015;163:127-134.
32. Ayer M, Brunk HD, Ewing GM, Reid WT, Silverman E. An empirical distribution function for sampling with incomplete information. *Ann Math Stat*. 1955;26:641-647.
33. Sun J. *The Statistical Analysis of Interval-censored Failure Time Data*. New York, NY: Springer Science & Business Media; 2007.
34. Huang J, Wellner JA. Asymptotic normality of the NPMLE of linear functionals for interval censored data, Case 1. *Statistica Neerlandica*. 1995;49:153-163.
35. Horiguchi M, Cronin AM, Takeuchi M, Uno H. A flexible and coherent test/estimation procedure based on restricted mean survival times for censored time-to-event data in randomized clinical trials. *Stat Med*. 2018;37:2307-2320.

How to cite this article: Zhang C, Wu Y, Yin G. Restricted mean survival time for interval-censored data. *Statistics in Medicine*. 2020;1–17. https://doi.org/10.1002/sim.8699