Received: 16 September 2018 | Revised: 24 January 2020 | Accepted: 27 January 2020

DOI: 10.1111/biom.13237

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

On the empirical choice of the time window for restricted mean survival time

Lu Tian[1] | Hua Jin[2] | Hajime Uno[3] | Ying Lu[1] | Bo Huang[4] | Keaven M. Anderson[5] | LJ Wei[6]

[1]Department of Biomedical Data Science, Stanford University, Stanford, California

[2]School of Mathematical Sciences, South China Normal University, Guangzhou, P. R. China

[3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts

[4]Pfizer Inc, Groton, Connecticut

[5]Merck & Co., Inc., North Wales, Pennsylvania

[6]Department of Biostatistics, Harvard University, Boston, Massachusetts

Correspondence
Lu Tian, Department of Biomedical Data Science, Stanford University, Stanford, CA, 94305.
Email: lutian@stanford.edu

Funding information
National Institutes of Health, Grant/Award Numbers: R01 HL089778, R00 HS022193, R21 AG049385; Natural Science Foundation of Guangdong Province, Grant/Award Numbers: 2019A1515011717, 2017A030313018

Abstract

The t-year mean survival or restricted mean survival time (RMST) has been used as an appealing summary of the survival distribution within a time window $[0, t]$. RMST is the patient's life expectancy until time t and can be estimated nonparametrically by the area under the Kaplan-Meier curve up to t. In a comparative study, the difference or ratio of two RMSTs has been utilized to quantify the between-group-difference as a clinically interpretable alternative summary to the hazard ratio. The choice of the time window $[0, t]$ may be prespecified at the design stage of the study based on clinical considerations. On the other hand, after the survival data have been collected, the choice of time point t could be data-dependent. The standard inferential procedures for the corresponding RMST, which is also data-dependent, ignore this subtle yet important issue. In this paper, we clarify how to make inference about a random "parameter." Moreover, we demonstrate that under a rather mild condition on the censoring distribution, one can make inference about the RMST up to t, where t is less than or even equal to the largest follow-up time (either observed or censored) in the study. This finding reduces the subjectivity of the choice of t empirically. The proposal is illustrated with the survival data from a primary biliary cirrhosis study, and its finite sample properties are investigated via an extensive simulation study.

KEYWORDS

hazard ratio, Kaplan-Meier estimator, logrank test, RMST

1 | INTRODUCTION

The Kaplan-Meier curve provides estimated survival probabilities over the entire study duration (Kaplan and Meier, 1958). On the other hand, summary measures for such a curve are essential for decision making. The mean survival time might be the first summary to consider, but it is not used because it often cannot be estimated from right-censored data. Most commonly used measures are the median survival time and the t-year survival probability. However, the median survival time also may not be estimable under heavy censoring. The survival probability at a specific time point does not capture temporal survival profile information before or after the time point. Recently, the t-year mean survival time or restricted mean survival time (RMST) has been proposed as an alternative summary for the survival curve (Irwin, 1949; Uno *et al.*, 2014, 2015; Trinquart *et al.*, 2016). The RMST is the expected value of survival time up to a fixed time point t. Graphically, the t-year RMST estimate is the area under the Kaplan-Meier curve up to t-year. Inferences about the RMST have been discussed extensively in the literature (Karrison, 1987; Zucker, 1998; Royston and Parmar, 2011; Tian *et al.*, 2014). In comparing two survival distributions, the difference or ratio of two RMSTs has been demonstrated to be better than the conventional hazard ratio in term of clinical interpretability, especially when the proportional hazards assumption is violated (Uno *et al*., 2014, 2015; Pak *et al*., 2017). Recently, as a test statistic for testing the equality of two survival curves, the RMST-based test was shown to be more powerful than the logrank test when the proportional hazards model is not valid, and almost as powerful as the logrank test even when this model is plausible (Zhao *et al*., 2012; Huang and Kuan, 2018; Tian *et al*., 2018).

As with the t-year survival probability, an important question for the RMST is how to choose the t-year time point, which is constrained due to the duration of study follow-up and other censoring. One may even summarize the survival distribution via the RMST up to a sequence of t's in an interval (Zhao *et al*., 2016). Ideally these choices should be prespecified in the design stage of the study based on clinical and study feasibility consideration. On the other hand, after data are collected, it is interesting and important to know what possible time window one can choose for computing RMST estimates. It is known that the large sample Gaussian approximation to the distribution of the Kaplan-Meier estimator is valid up to a time point τ, at which the proportion of the patients at risk is greater than a prespecified level above zero. Therefore, the RMST estimator would be asymptotically valid with $t < \tau$. However, such a practice is subjective and may be unsatisfactory, as it ignores information after the time point t. In this article, we show that under a rather mild condition on the censoring distribution, the RMST can be estimated well by choosing t, which is less than or equal to the largest follow-up time (either observed or censored). With such a data-dependent time window, we derive the large sample approximation to the distribution of the RMST estimator and study the corresponding procedure for inference, for example, the construction of confidence intervals for the underlying RMST, which is itself data-dependent. Data from the primary biliary cirrhosis study are used for illustration and a simulation study is conducted to investigate the adequacy of the proposed inferential procedure.

2 | METHOD

Suppose that our observations consist of n independent identically distributed copies of $(X, \delta) = (T \wedge C, I(T \leq C))$: $\{(X_i, \delta_i) = (T_i \wedge C_i, I(T_i \leq C_i)), i = 1, \ldots, n\}$, where $a \wedge b = \min(a, b)$, $I(A)$ is the indicator function for event A, and T and C are independent failure time and censoring time, respectively. We focus on the case that T is a continuous random variable and C has bounded support as in most randomized clinical trials. Let $\tau_C = \inf\{\tau \mid P(C \geq \tau) = 0\}$ be the upper end of the support of the censoring time C. We also assume that $P(T \geq \tau_C) > 0$, that is, the support of T is greater than that of C. Clearly, $\hat{\tau}_C = \max\{X_1, \ldots, X_n\}$ is a finite-sample approximation to τ_C. Let $\hat{S}(t)$ be the Kaplan-Meier estimator of the survival function of T, denoted by $S(t)$. Then, $R(t) = \int_0^t S(u)du$ and $\hat{R}(t) = \int_0^t \hat{S}(u)du$ are the RMST up to time t and its consistent estimator, respectively.

Using the standard martingale argument in Chapter 3 of Fleming and Harrington (2011), one can show that $\sqrt{n}\{\hat{S}(t) - S(t)\}$ converges weakly to a mean zero Gaussian process at $t \in [0, \tau]$, for any $\tau < \tau_C$ as $n \to \infty$. By the continuous mapping theorem, $D(t) = \sqrt{n}\{\hat{R}(t) - R(t)\}, t \in [0, \tau]$ also converges weakly to a mean zero Gaussian process (Zhao *et al*., 2016). To apply this approximation for making inferences about the RMST up to a specific time point, the time point has to be no greater than $\hat{\tau}_C$. Such a time point, denoted by $\hat{t}$, is data-dependent. If $\hat{t}$ converges to a fixed constant $t_0 < \tau_C$, the above large sample approximation is often used to justify the inferential procedure of the RMST up to t_0 by pretending $t_0 = \hat{t}$. However, the quantity of interest is $R(\hat{t})$, which is a random "parameter" and different from $R(t_0)$. This subtle but important issue is not unique for the RMST, it also arises from estimating the survival probability at a time point, which may also be selected according to observed data. It is not clear how to justify the validity of the standard inferential procedure treating the time point as fixed for these cases. Here, we first clarify how to construct a confidence interval for $R(\hat{t})$. Then, we extend our inferential procedure to the case in which $\hat{t} = \hat{\tau}_C$.

To this end, note that the tightness of a process $D(t)$ implies that $D(\hat{t}) - D(t_0) = o_p(1)$, i.e.,

$$\sqrt{n}\{\hat{R}(\hat{t}) - R(\hat{t})\} = \sqrt{n}\{\hat{R}(t_0) - R(t_0)\} + o_p(1)$$

converges weakly to mean zero Gaussian distribution with a variance of

$$\sigma^2(t_0) = \int_0^{t_0} \frac{\left\{\int_t^{t_0} S(u)du\right\}^2 d\Lambda(t)}{G(t)},$$

where $\Lambda(t)$ is the cumulative hazard function of T, and $G(t) = P(X \geq t)$. The variance $\sigma^2(t_0)$ can be consistently estimated by

$$\hat{\sigma}^2(\hat{t}) = \int_0^{\hat{t}} \frac{\left\{\int_t^{\hat{t}} \hat{S}(u)du\right\}^2 d\hat{\Lambda}(t)}{\hat{G}(t)},$$

where $\hat{\Lambda}(t) = -\log\{\hat{S}(t)\}$, and $\hat{G}(t) = n^{-1}\sum_{i=1}^n I(X_i \geq t)$. These imply that first

$$\hat{R}(\hat{t}) - R(\hat{t}) = o_p(1),$$

and second

$$P\left(\frac{\sqrt{n}|\hat{R}(\hat{t}) - R(\hat{t})|}{\hat{\sigma}(\hat{t})} \leq z_{0.975}\right) \to 0.95,$$

as $n \to \infty$, where z_α is the αth quantile of the standard normal, that is, $\hat{R}(\hat{t})$ and

$$\left[\hat{R}(\hat{t}) - z_{0.975}\frac{\hat{\sigma}(\hat{t})}{\sqrt{n}}, \hat{R}(\hat{t}) + z_{0.975}\frac{\hat{\sigma}(\hat{t})}{\sqrt{n}}\right] \tag{1}$$

can be viewed as a point estimator and a 95% confidence interval for $R(\hat{t})$, respectively. The inference results for $R(\hat{t})$ can be interpreted similarly as those from conventional inferences even though the "parameter" of interest, $R(\hat{t})$, is a random variable itself. For instance, for the 95% confidence interval (1), this means that if we hypothetically repeat the study, say, 100 times (including the current study), to generate 100 different observed $R(\hat{t})$ values and corresponding confidence intervals, there would be about 95 "good" intervals, which cover $R(\hat{t})$. As we only observe the data from a single study, $\hat{t}$ would be a fixed time point determined by the observed data and we are interested in the RMST up to this specific time point. The above argument suggests that the observed confidence interval is very likely to be a "good" one covering $R(\hat{t})$ in our actual study. Therefore, the conventional inferential procedure treating $\hat{t}$ as a fixed time point is valid for the empirically chosen $\hat{t}$. For example, we may choose $\hat{t}$ to be the 90th percentile of observed follow-up times as suggested in the literature.

Next, we show that our inferential procedure can even handle the case when $\hat{t} = \hat{\tau}_C$, the last observed follow-up time in the study under a mild condition on the censoring distribution. Note that $\hat{\tau}_C$ converges to τ_C and the aforementioned argument does not apply directly. However, via Theorem A1 in the Appendix, we can show that $D(t \wedge \hat{\tau}_C)$ converges weakly to a Gaussian process indexed by $t \in [0, \tau_C]$, which implies that $D(\hat{\tau}_C) = \sqrt{n}\{\hat{R}(\hat{\tau}_C) - R(\hat{\tau}_C)\}$ approximately follows a mean zero Gaussian distribution with a finite variance. The key condition in Theorem A1 is simply the finiteness of the asymptotic variance of $D(\hat{\tau}_C)$:

$$\sigma^2(\tau_C) = \int_0^{\tau_C} \frac{\left\{\int_t^{\tau_C} S(u)du\right\}^2 d\Lambda(t)}{G(t)} < \infty. \tag{2}$$

To make inference on $R(\hat{\tau}_C)$, we may estimate $\sigma^2(\tau_C)$ by $\hat{\sigma}^2(\hat{\tau}_C)$ and take

$$\left[\hat{R}(\hat{\tau}_C) - z_{0.975}\frac{\hat{\sigma}(\hat{\tau}_C)}{\sqrt{n}}, \hat{R}(\hat{\tau}_C) + z_{0.975}\frac{\hat{\sigma}(\hat{\tau}_C)}{\sqrt{n}}\right]$$

as a 95% confidence interval for $R(\hat{\tau}_C)$, whose interpretation is the same as that for $R(\hat{t})$ discussed above.

Now, we provide some physical, intuitive interpretations of Condition (2). To this end, we assume that $\lambda(t)$, the hazard rate function of T, is uniformly bounded within the interval $[0, \tau_C]$, and the censoring time C has a density function $f_C(t)$. Then, the integrand of the integral in (2) is bounded except at τ_C and

$$\frac{\left(\int_t^{\tau_C} S(u)du\right)^2 \lambda(t)}{G(t)} = O\left(\frac{\tau_C - t}{f_C(t)}\right),$$

as $t \to \tau_C$. One sufficient condition for (2) is thus that

$$\lim_{t \to \tau_C} \frac{f_C(t)}{(\tau_C - t)^{1+\delta}} > 0, \tag{3}$$

for a $\delta \in (0, 1)$. In other words, the density function of the censoring distribution cannot reach zero too fast when approaching τ_C, the upper end of its support. In clinical trials, censoring is often induced by a combination of attrition from the loss to follow-up and staggered entry (so-called administrative censoring). The latter is often the dominating factor and is caused by the fact that patients entered the study at different time points (Lachin and Foulkes, 1986). If we assume that patients entered the study uniformly over the accrual period, the density function of the induced administrative censoring distribution is bounded away from zero at τ_C and Condition (2) is trivially satisfied, as long as the probability of loss to follow-up is less than 1 during the study period. Condition (2) may be violated if the enrollment is very slow at the beginning of the study and greatly accelerates later. Empirically, this would be reflected by a flat tail in the Kaplan-Meier curve over a long time interval up to $\hat{\tau}_C$ containing very few censoring observations, which become more sparse toward $\hat{\tau}_C$. For example, if $\hat{\tau}_C$, the largest censoring time, is far away from other censoring times, then it is plausible that Condition (2) may not be satisfied and we should be cautious in making inference for the RMST up to $\hat{\tau}_C$.

There are parallel results for the asymptotic properties of the Kaplan-Meier estimator on the interval $[0, \tau_C]$ (Ying, 1989; Stute, 1995). However, the required condition

$$\int_0^{\tau_C} \frac{d\Lambda(t)}{G(t)} < \infty$$

is often violated in the context of clinical trials. For example, when t is close to τ_C, $G(t) = O(\tau_C - t)$ in the presence of uniform censoring caused by staggered entry and the integral above does not converge to a finite number. In other words, if we want to make inference on t-year survival probability based on the Kaplan-Meier estimator, then t needs to be chosen a priori such that $P(X \geq t) > 0$. In particular, t cannot be chosen to be $\hat{\tau}_C$.

3 | EXAMPLES

We used the data from the Mayo Clinic trial on primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984 as our initial illustrative example (Therneau and

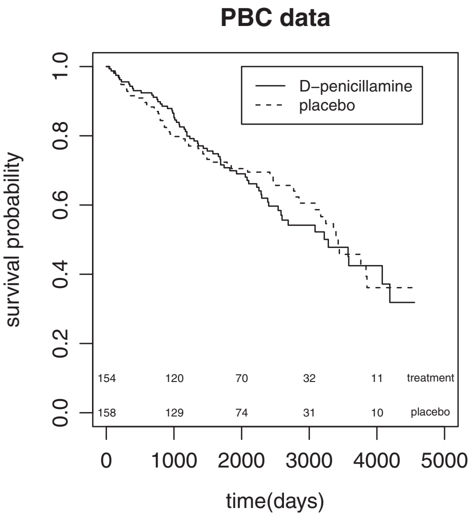

FIGURE 1 The Kaplan-Meier curves of the survival distribution by treatment arm in PBC and ECOG myeloma studies

Grambsch, 2013). Note that 158 and 154 PBC patients at Mayo Clinic were randomized to D-penicillamine and placebo arms, respectively, to investigate the potential benefit of D-penicillamine in prolonging the survival of PBC patients. The Kaplan-Meier curves by treatment arm were presented in Figure 1. There was no statistically significant difference between two groups based on the logrank test ($P = 0.75$). Suppose that we want to estimate the RMST up to time τ, which was chosen such that there would be at least 5% of the patients at risk at time τ in both arms. The largest of such $\tau = \hat{t}$ was 11.11 years and the estimated RMST up to 11.11 years was 7.62 (95% CI: 6.97-8.26) years in the treatment arm and 7.73 (95% CI: 7.07-8.39) years in the placebo arm. However, if we choose τ to be the minimum of the largest $\hat{\tau}_C$ from two arms, τ can be as large as 12.39 years, because the largest follow-up time is 12.48 years in the placebo arm and 12.39 years in the treatment arm. The estimated RMST up to 12.39 years was 8.05 (95% CI: 7.30-8.80) years in the treatment arm and 8.19 (95% CI: 7.42-8.97) years in the placebo arm. Note that $\tau = 12.39$ years was almost 11% longer than $\hat{t}$ in this example. The censoring distribution was also estimated by Kaplan-Meier curves presented in Figure 2. The near linear shape of the Kaplan-Meier curve from approximately year 2.74 to 12.48 was a strong indication that the censoring distribution induced by staggered entry followed a uniform distribution from approximately year 2.74 to τ_C, which was close to, but greater than year 12.48, the largest follow-up time in the study. This particular censoring pattern corresponds to uniform enrollment from the baseline to approximately year 9 followed by a 3-year follow-up. Thus Condition (2) was likely satisfied in this example, which ensured the validity of estimating the RMST up to year 12.39 in both arms. As a cautionary note, we do not suggest formally testing this condition, because its power may be affected by various factors and the final decision is still a subjective one.

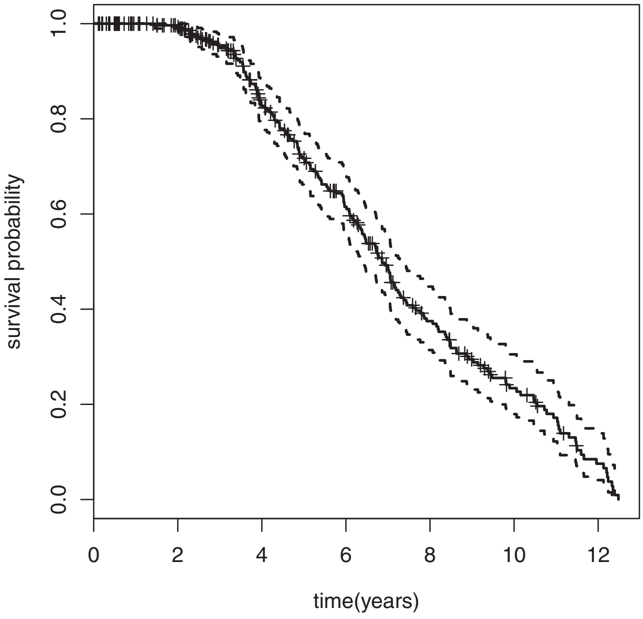

FIGURE 2 The Kaplan-Meier curve of the censoring distribution in PBC study

In the second example, we used a study recently conducted by the ECOG-ACRIN Cancer Research Group to compare low- and high-dose dexamethasone for treating newly diagnosed multiple myeloma patients (Rajkumar *et al*., 2010). In this study, 222 and 223 patients were randomized to the low-dose and high-dose dexamethasone arms, respectively. The resulting Kaplan-Meier curves of overall survival by treatment arms are presented in Figure 1. It appeared that patients in the low-dose group survived longer than those in the high-dose group but the difference in survival probability diminished toward the end of the study, which implied the presence of crossing hazards, a potential reason for the insignificant P-value from the logrank test ($P = 0.49$). If we

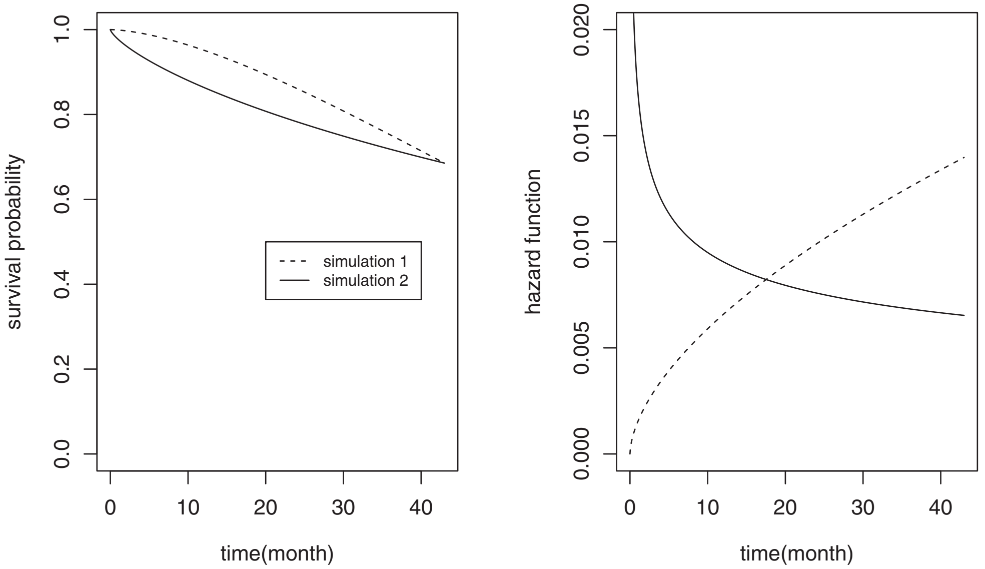

FIGURE 3 The survival and hazard functions of the survival distribution used in the simulation study

want to estimate the RMST up to 41.50 months, at which at least 5% of the patients were still at risk in both arms, then the estimated RMST was 36.48 (95% CI: 35.18-37.78) months in the low-dose arm and 34.32 (95% CI: 32.60-36.04) months in the high-dose arm. However, as described in the paper, we may choose the truncation time point to be the minimum of the largest $\hat{\tau}_C$ from two arms. Consequently, it can be as large as 42.67 months and the estimated RMST up to month 42.67 was 37.26 (95% CI: 35.89-38.62) months in the low-dose arm and 35.18 (95% CI: 33.40-36.96) months in the high-dose arm. In this example, $t = 42.67$ months were about 1 month longer than $\hat{t}$. Interestingly, the *P*-value for the two-group comparison was 0.070 based on the RMST up to month 42.67 and 0.049 based on the RMST up to month 41.50. Therefore, the RMST over a wider time window does not always generate more significant results in a two-group comparison, especially if these two survival curves start to converge to each other at the end of the follow-up period as in this example; note that there is also an increase in the variance of the RMST estimates at the later time point. However, this is not necessarily a reason against using larger truncation time points in the RMST analysis, because it is generally desirable to evaluate the treatment effect over a longer follow-up period. In an extreme case, where the survival curves crossed, a significant *P*-value obtained by comparing the survival curve at earlier time points could be premature and misleading.

4 | SIMULATION

In this section, we examined the finite sample performance of $\hat{R}(\hat{\tau}_C)$ in estimating $R(\hat{\tau}_C)$ and $\hat{R}(\hat{t})$ in estimating $R(\hat{t})$, where $\hat{t}$ converges to a limit less than τ_C. To this end, the survival time was generated from a Weibull distribution with a scale parameter of exp(4.37) and a shape parameter of 1.59. The true survival and hazard functions were presented in Figure 3. The censoring time was generated from $E \wedge U$, where $U \sim U(\tau_I, \tau_C) = U(24, 43)$, representing administrative censoring caused by staggered entry and E followed an exponential distribution such that $P(E \geq \tau_C) = 0.90$, representing the loss to follow-up. E and U were independently generated. All model parameters were chosen to mimic the high-dose arm of the ECOG myeloma study in the previous section. Under this simulation setting, $\tau_C = 43$ months. In each set of simulations, we generated n pairs of $(X_i, \delta_i), i = 1, \ldots, n$, to obtain $\hat{\tau}_C$, $R(\hat{\tau}_C)$, $\hat{R}(\hat{\tau}_C)$, $\hat{t}$, $R(\hat{t})$, and $\hat{R}(\hat{t})$, where $\hat{t}$ was the 95th percentile of $\{X_1, \ldots, X_n\}$. The 95% confidence intervals for $R(\hat{\tau}_C)$ and $R(\hat{t})$ were also constructed. Repeating this process 10 000 times, the empirical bias of point estimators and coverage level of 95% confidence intervals were summarized in Table 1. We also reported the average standard error estimates of $\hat{R}(\hat{\tau}_C) - R(\hat{\tau}_C)$ and $\hat{R}(\hat{t}) - R(\hat{t})$, which were compared with the corresponding empirical standard errors. The sample size n was set to be 30, 100, 300, and 1000. The simulation was also repeated with U being the sum of two independent uniform distributions $U(12, 21.5)$. In this setting, $f_C(t) = O(\tau_C - t)$ as $t \to \tau_C$ and still satisfied Condition (2) (Figure 4). For both censoring distributions, the proportion of censored observations was approximately 79%, representing heaving censoring.

The empirical bias of $\hat{R}(\hat{\tau}_C)$ in estimating $R(\hat{\tau}_C)$ was small relative to the truth and the 95% confidence interval covered $R(\hat{\tau}_C)$ in approximately 95% of the random data sets drawn for a moderate or big n. The only exception was that for small n, $\hat{\sigma}(\hat{\tau}_C)$ tended to slightly underestimate the standard error of $\hat{R}(\hat{\tau}_C) - R(\hat{\tau}_C)$, which led to mild under-coverage of the

TABLE 1 Finite sample performance of $\hat{R}(\hat{\tau}_C)$ and $\hat{R}(\hat{t})$, where $\hat{t}$ is the 95th percentile of observed follow-up times, and the survival time follows a Weibull distribution with increasing hazard

		$R(\hat{\tau}_C)$					$R(\hat{t})$				
Censoring	***n*(# events)**	$E(\hat{\tau}_C)$	**BIAS**	**ESE**	**ASE**	**COV(%)**	$E(\hat{t})$	**BIAS**	**ESE**	**ASE**	**COV(%)**
I	30 (6.4)	42.03	0.01	1.95	1.81	90.1	40.7	0.03	1.83	1.71	90.0
I	100 (21.5)	42.69	−0.02	1.10	1.07	93.6	41.2	−0.01	1.03	1.00	93.8
I	300 (64.2)	42.89	−0.00	0.64	0.63	94.3	41.4	0.00	0.60	0.59	94.4
I	1000 (213.8)	42.97	0.00	0.35	0.35	94.9	41.5	0.00	0.33	0.33	94.8
II	30 (6.4)	40.35	0.04	1.80	1.67	89.9	38.6	0.03	1.66	1.55	89.7
II	100 (21.4)	41.50	−0.01	1.04	1.01	93.6	39.0	−0.00	0.93	0.90	93.6
II	300 (64.1)	42.14	0.01	0.63	0.61	94.2	39.2	0.01	0.54	0.53	94.2
II	1000 (213.5)	42.53	0.00	0.35	0.35	94.7	39.3	0.00	0.29	0.29	94.7

Abbreviations: *n*, the sample size; # events, the average number of observed failures; $E(\hat{\tau}_C)$, the empirical average of $\hat{\tau}_C$; $E(\hat{t})$, the empirical average of $\hat{t}$; BIAS, the empirical bias in estimating $R(\hat{\tau}_C)$ and $R(\hat{t})$; ESE, the empirical standard error of $\hat{R}(\hat{\tau}_C) - R(\hat{\tau}_C)$ or $\hat{R}(\hat{t}) - R(\hat{t})$; ASE, the empirical average of standard error estimator of $\hat{R}(\hat{\tau}_C) - R(\hat{\tau}_C)$ or $\hat{R}(\hat{t}) - R(\hat{t})$; COV, the empirical coverage probability of the 95% confidence interval.

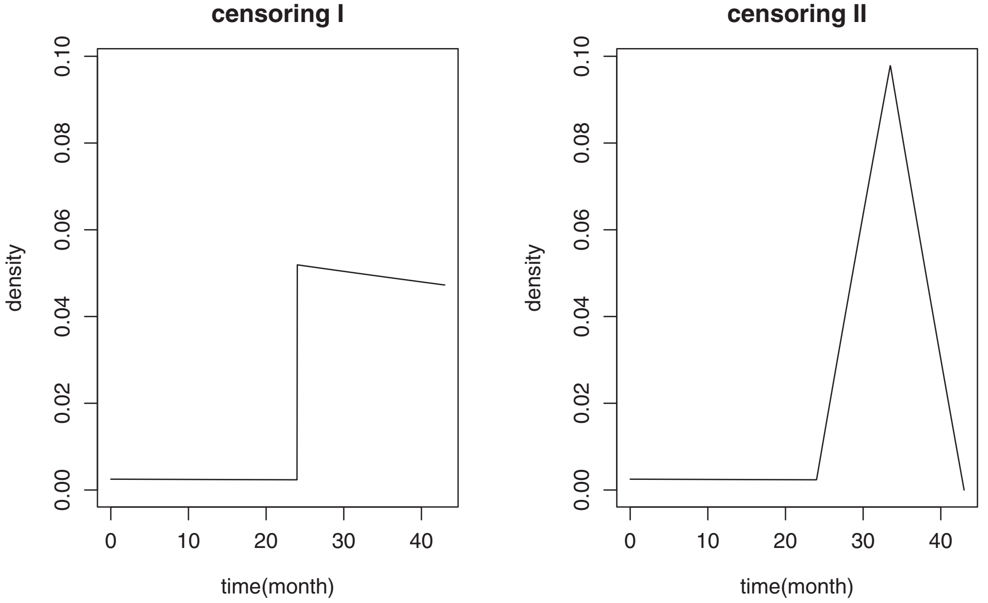

FIGURE 4 The density function of the censoring distribution used in the simulation study

corresponding 95% confidence intervals of $R(\hat{\tau}_C)$. However, this under-coverage quickly diminished when the sample size *n* increased. A similar pattern was observed for the inferential results for $R(\hat{t})$.

We also repeated this simulation by generating survival times from a different Weibull distribution with a scale parameter of exp(5.07) and a shape parameter of 0.74, which were chosen to mimic the low-dose arm of the ECOG myeloma study. The purpose of this new simulation was to investigate the finite sample performance of the proposed inferential procedure, when the hazard rate decreases toward the end of follow-up (Figure 3). This type of survival profile featured by a long nearly flat tail of the Kaplan-Meier estimator have been observed in many positive immunotherapy trials and attracted attention recently (Liu *et al*., 2018; Wei and Wu, 2020). The true survival function was given in Figure 3. The results were summarized in Table 2 and were similar to those in Table 1.

In the second set of simulations, we investigated the effect of extending the RMST up to the largest follow-up time on hypothesis testing. First, we studied the type one error by generating survival times in both arms from a common Weibull distribution with a scale parameter of exp(5.07) and a shape parameter of 0.74, mimicking the high-dose group in the ECOG study. The censoring distributions in two arms were chosen to be identical to those in the first set of simulations. For each set of simulated data, we conducted (a) the logrank test, (b) the test based on the difference in RMST up to the minimum of $\hat{\tau}_C$ from two arms, and (c) that based on the difference in RMST up to time $\hat{t}_\alpha$, at which at least $(100 - \alpha)\%$ of the patients were at risk in both arms and $\alpha = 80, 85, 90,$

TABLE 2 Finite sample performance of $\hat{R}(\hat{\tau}_C)$ and $\hat{R}(\hat{t})$, where $\hat{t}$ is the 95th percentile of observed follow-up times, and the survival time follows a Weibull distribution with decreasing hazard

		$R(\hat{\tau}_C)$					$R(\hat{t})$				
Censoring	***n* (# events)**	$E(\hat{\tau}_C)$	**BIAS**	**ESE**	**ASE**	**COV(%)**	$E(\hat{t})$	**BIAS**	**ESE**	**ASE**	**COV(%)**
I	30 (7.8)	42.02	0.00	2.56	2.42	91.4	40.63	0.02	2.45	2.32	91.4
I	100 (26.0)	42.69	−0.03	1.44	1.41	93.9	41.21	-0.02	1.37	1.34	93.9
I	300 (77.8)	42.90	−0.00	0.83	0.82	94.3	41.39	0.00	0.79	0.79	94.3
I	1000 (259.3)	42.97	−0.00	0.46	0.45	94.7	41.45	0.00	0.43	0.43	94.9
II	30 (7.8)	40.32	0.01	2.43	2.29	91.1	38.56	0.02	2.29	2.16	91.0
II	100 (26.1)	41.50	−0.02	1.38	1.35	93.9	38.99	-0.02	1.27	1.24	93.8
II	300 (78.1)	42.13	−0.00	0.82	0.80	94.3	39.15	0.00	0.74	0.73	94.2
II	1000 (259.9)	42.53	0.00	0.45	0.45	94.8	39.21	0.00	0.40	0.40	94.7

Abbreviations: *n*, the sample size; # events, the average number of observed failures; $E(\hat{\tau}_C)$, the empirical average of $\hat{\tau}_C$; $E(\hat{t})$, the empirical average of $\hat{t}$; BIAS, the empirical bias in estimating $R(\hat{\tau}_C)$ and $R(\hat{t})$; ESE, the empirical standard error of $\hat{R}(\hat{\tau}_C) - R(\hat{\tau}_C)$ or $\hat{R}(\hat{t}) - R(\hat{t})$; ASE, the empirical average of standard error estimator of $\hat{R}(\hat{\tau}_C) - R(\hat{\tau}_C)$ or $\hat{R}(\hat{t}) - R(\hat{t})$; COV, the empirical coverage probability of the 95% confidence interval.

and 95. The empirical type one error was calculated from 10 000 simulations.

We then studied the power of the relevant tests by generating survival times from a Weibull distribution with a scale parameter of exp(5.07) and a shape parameter of 0.74 in the control arm and from a Weibull distribution with a scale parameter of exp(4.37) and a shape parameter of 1.59 in the treatment arm, mimicking the observed ECOG myeloma data. Note that these two hazard functions crossed at month 17 and the logrank test was not powerful in this setting. The empirical power was also calculated from 10 000 simulations. Finally, we studied the power under the proportional hazards assumption. To this end, the survival times were generated from Weibull distributions with the scale parameter being exp(5.07) and exp(5.37) for placebo and treatment arms, respectively. Both Weibull distributions shared the same shape parameter of 0.74 and the proportional hazards assumption was satisfied with the corresponding hazard ratio of 0.80.

The simulation results were summarized in Table 3. The type one error was well preserved for moderate sample size and number of events. When the sample size per arm was 30 with less than 8 events on average, the type one error of the RMST-based test was slightly inflated but still below 0.06. Under the given nonproportional hazards alternative, the power of the logrank test is poor due to crossing hazards. On the other hand, powers of tests based on the RMST up to different truncation time points were higher. Interestingly, the power based on the RMST up to a smaller truncation time point actually was slightly higher in this setting. This was not a surprise, as the difference in RMST changed very little toward month 43, the maximum follow-up time, while the variance of the estimated RMST still increased. However, the power of the test was only one factor to consider. By extending the truncation time of the RMST to $\hat{\tau}_C$, we could compare two survival distributions and summarized their difference over a wider time window, better representing the entire survival distribution. Under the proportional hazards alternative, the logrank test had the highest power, when the type one error was controlled at the same level. The power of the RMST-based test up to $\hat{\tau}_C$ was almost identical to that of the logrank test. The powers of tests based on the RMST up to smaller truncation times were only slightly lower, which suggested that the power of these tests was fairly robust to the choice of truncation time point in this case.

5 | DISCUSSION

In this paper, under a rather mild condition likely satisfied in most clinical trial settings, we argue that one is able to estimate the RMST up to the largest follow-up time in the observed data set and make appropriate statistical inference. In practice, we may estimate the RMST up to any time rounded in weeks, months, or years before the largest follow-up time for easy interpretation. This alleviates the concern regarding the subjective choice of the truncation time in RMST-based inferences. The result also enables the potential utilization of maximum amount of information in the observed data in making relevant statistical inferences. Although the logrank test uses only information up to the minimum of $\hat{\tau}_C$ and the largest observed event time in the entire study, the RMST-based test can use information up to the minimum of $\hat{\tau}_C$ according to the result of this paper. In other words, the RMST-based test is always able to use observed follow-up information in a time window wider than or equal to that of the logrank test. However, we want to emphasize that this does not automatically translate into higher power or better statistical efficiency. On the contrary, as our simulation study demonstrated, the power of the logrank test can still be higher than that of the test based on the RMST in some scenarios, regardless of the truncation

TABLE 3 Finite sample power of the logrank test, and tests based on the difference in RMST with different truncation time points

		Null										
		Tests based on the RMST										
		$R(\hat{\tau}_C)$		$R(\hat{t}_{95})$		$R(\hat{t}_{90})$		$R(\hat{t}_{85})$		$R(\hat{t}_{80})$		logrank
Censoring	***n* (# events)**	$E(\hat{\tau}_C)$	$\alpha(\%)$	$E(\hat{t}_{95})$	$\alpha(\%)$	$E(\hat{t}_{90})$	$\alpha(\%)$	$E(\hat{t}_{85})$	$\alpha(\%)$	$E(\hat{t}_{80})$	$\alpha(\%)$	$\alpha(\%)$
I	60 (15.6)	41.5	5.6	39.9	5.5	38.3	5.5	36.9	5.5	35.5	5.4	4.7
I	200 (51.9)	42.5	5.6	40.8	5.4	39.3	5.4	37.8	5.4	36.4	5.5	5.1
I	600 (155.5)	42.8	5.4	41.2	5.3	39.6	5.2	38.2	5.2	36.7	5.3	5.0
I	2000 (518.8)	43.0	5.1	41.3	5.1	39.8	5.1	38.4	5.0	36.9	5.1	5.1
II	60 (15.7)	39.5	5.6	37.8	5.5	36.5	5.6	35.5	5.6	34.6	5.5	4.5
II	200 (50.0)	41.1	5.6	38.6	5.4	37.1	5.5	36.1	5.4	35.1	5.4	5.2
II	600 (155.8)	41.9	5.3	38.9	5.2	37.4	5.2	36.3	5.2	35.3	5.1	5.2
II	2000 (520.0)	42.4	5.1	39.1	4.9	37.6	4.9	36.4	5.0	35.5	4.9	5.1
		non-PH alternative										
		Tests based on the RMST										
		$R(\hat{\tau}_C)$		$R(\hat{t}_{95})$		$R(\hat{t}_{90})$		$R(\hat{t}_{85})$		$R(\hat{t}_{80})$		**logrank**
Censoring	***n*(# events)**	$E(\hat{\tau}_C)$	**Power(%)**	$E(\hat{t}_{95})$	**Power(%)**	$E(\hat{t}_{90})$	**power(%)**	$E(\hat{t}_{85})$	**power(%)**	$E(\hat{t}_{80})$	**power(%)**	**power(%)**
I	60(14.2)	41.6	14	40.0	14	38.4	15	37.0	16	35.7	17	8
I	200(47.3)	42.5	30	40.9	33	39.3	36	37.9	38	36.5	41	15
I	600(141.9)	42.8	71	41.2	76	39.7	79	38.2	83	36.8	85	37
I	2000(473.3)	43.0	100	41.3	100	39.8	100	38.4	100	37.0	100	84
II	60(14.3)	39.5	14	37.8	15	36.5	16	35.6	17	34.7	17	8
II	200(47.4)	41.1	33	38.6	37	37.2	40	36.1	42	35.2	44	15
II	600(142.0)	41.9	74	38.9	82	37.5	85	36.4	86	35.4	88	38
II	2000(473.7)	42.4	100	39.1	100	37.6	100	36.5	100	35.5	100	85
		PH alternative										
		Tests based on the RMST										
		$R(\hat{\tau}_C)$		$R(\hat{t}_{95})$		$R(\hat{t}_{90})$		$R(\hat{t}_{85})$		$R(\hat{t}_{80})$		**logrank**
Censoring	***n* (# events)**	$E(\hat{\tau}_C)$	**Power(%)**	$E(\hat{t}_{95})$	**Power(%)**	$E(\hat{t}_{90})$	**power(%)**	$E(\hat{t}_{85})$	**power(%)**	$E(\hat{t}_{80})$	**power(%)**	**power(%)**
I	60 (14.2)	41.6	8	40.0	8	38.5	8	37.1	8	35.7	8	7
I	200 (47.3)	42.6	12	40.9	12	39.4	12	37.9	12	36.6	11	12
I	600 (141.8)	42.9	26	41.2	25	39.7	25	38.3	24	36.9	24	27
I	2000 (473.1)	43.0	66	41.4	66	39.9	65	38.5	64	37.1	63	68
II	60 (14.3)	39.5	8	37.8	8	36.5	8	35.6	8	34.7	8	7
II	200 (47.4)	41.1	12	38.6	12	37.2	12	36.1	11	35.2	11	12
II	600 (142.2)	41.9	25	38.9	25	37.5	24	36.4	24	35.4	24	27
II	2000 (474.2)	42.4	66	39.1	64	37.6	63	36.5	62	35.5	62	68

Abbreviations: n, the total sample size; # events, the average number of observed failures; $\hat{\tau}_C$, the RMST up to the smallest $\hat{\tau}_C$ of two arms; $\hat{t}_\alpha$, the RMST up to the time point at which $(100-\alpha)\%$ of the patients are at risk in either arm; $E(\hat{\tau}_C)$, the empirical average of $\hat{\tau}_C$; $E(\hat{t})$, the empirical average of $\hat{t}$; α, the empirical type one error.

time point. Furthermore, the test based on the RMST up to a smaller truncation time point can also be more powerful than that based on the RMST up to a bigger truncation time point. But, if we view the RMST over a wider time window as a more "global" summary of the survival distribution of interest, then a larger truncation time point of RMST is more desirable. However, if there are clinical or scientific considerations preferring specific time points, then we may not want to choose a larger truncation time point, even if it is statistically feasible.

In practice, the censoring mechanism may not be completely known. We need to be cautious in making inference for the RMST up to $\hat{\tau}_C$, if there is a long time interval before $\hat{\tau}_C$, containing only very few sparsely distributed observed censored times, which become sparser toward the end of the follow-up. It can be an indication of small mass of the censoring distribution toward $\hat{\tau}_C$, and potential violation of Condition (2). In such a case, it is safer to choose a time point at which a proportion of patients are still at risk as the upper end of the RMST of interest. Finally, if one plans to estimate the RMST up to $\hat{\tau}_C$ in a future study, the operational planning in site initiation and accrual could take Condition (2) into consideration, so that the early accrual is not too slow to invalidate it.

The estimated RMST always needs to be accompanied explicitly by the corresponding truncation time point, which is often a random quantity with an unknown limit as discussed in the paper. The value of the proposed maximum truncation time point is determined by the censoring pattern and may vary from study to study, and from subgroup to subgroup, which presents challenges in meta analysis and subgroup analysis. However, the commonly used hazard ratio is not immune to this problem, because it also depends on the time window within which the hazard ratio is estimated.

In general, we may also use $\hat{R}(\hat{t})$ to estimate the RMST up to time t_0, where $\hat{t}$ is a consistent estimator of t_0. Furthermore, if the convergence of $\hat{t}$ is at a rate faster than root n, then the statistical inference can be made as if $\hat{t} = t_0$. However, if the convergence rate is root n, for example, when $\hat{t}$ is the maximum time point at which 95% of the patients are still at risk, then we need to consider the variance of $\hat{t} - t_0$ in making statistical inference on $R(t_0)$. However, in general, we do not recommend making inference on $R(t_0)$, as the value of t_0 and thus the estimand normally is unknown to us. For the rare case, where τ_C is known, one may estimate the RMST up to τ_C by the area under the extrapolated Kaplan-Meier curve up to τ_C. As $\hat{\tau}_C$ converges to τ_C at a rate faster than root n rate under mild conditions, the asymptotic variance of the resulting estimator can also be approximated by $\hat{\sigma}^2(\hat{\tau}_C)/n$. Unlike $R(\hat{\tau}_C)$, $R(\tau_C)$ is a deterministic parameter and the associated estimation and inference can then be carried out and interpreted in a more conventional manner.

ACKNOWLEDGMENTS

We thank the associate editor and two referees for very detailed and highly constructive comments that greatly improved the paper. Drs. Tian, Wei, and Uno's research was partially supported by R01 HL089778, R00 HS022193, and R21 AG049385 from National Institute of Health, USA. Dr. Jin's research was partially supported by grants from Guangdong Engineering Research Center for Data Science, Natural Science Foundation of Guangdong Province, China (2017A030313018, 2019A1515011717).

DATA AVAILABILITY STATEMENT

The data used to support the findings of this study are available from Therneau and Grambsch (2013).

ORCID

Lu Tian https://orcid.org/0000-0002-5893-0169
Hua Jin https://orcid.org/0000-0003-4511-6165
Hajime Uno https://orcid.org/0000-0003-0622-8471
Ying Lu https://orcid.org/0000-0002-7698-8962
Bo Huang https://orcid.org/0000-0002-3088-9328
Keaven M. Anderson https://orcid.org/0000-0003-3218-1363
LJ Wei https://orcid.org/0000-0002-8582-920X

REFERENCES

Fleming, T.R. and Harrington, D.P. (2011) *Counting Processes and Survival Analysis*. Honoken, NJ: John Wiley & Sons.

Gill, R. (1983) Large sample behaviour of the product-limit estimator on the whole line. *The Annals of Statistics*, 11, 49–58.

Huang, B. and Kuan, P.-F. (2018) Comparison of the restricted mean survival time with the hazard ratio in superiority trials with a time-to-event end point. *Pharmaceutical Statistics*, 17, 202–213.

Irwin, J. (1949) The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *Journal of Hygiene*, 47, 188–189.

Kaplan, E.L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457–481.

Karrison, T. (1987) Restricted mean life with adjustment for covariates. *Journal of American Statistical Association*, 82, 1169–1176.

Lachin, J.M. and Foulkes, M.A. (1986) Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics*, 507–519.

Liu, S., Chu, C. and Rong, A. (2018) Weighted log-rank test for time-to-event data in immunotherapy trials with random delayed treatment effect and cure rate. *Pharmaceutical statistics*, 17, 541–554.

Pak, K., Uno, H., Kim, D.H., Tian, L., Kane, R.C., Takeuchi, M., Fu, H., Claggett, B. and Wei, L.-J. (2017) Interpretability of cancer clinical trial results using restricted mean survival time as an alternative to the hazard ratio. *JAMA Oncology*, 3, 1692–1696.

Rajkumar, S.V., Jacobus, S., Callander, N.S., Fonseca, R., Vesole, D.H., Williams, M.E., Abonour, R., Siegel, D.S., Katz, M., Greipp, P.R. and Eastern Cooperative Oncology Group (2010) Lenalidomide plus high-dose dexamethasone versus lenalidomide plus low-dose dexamethasone as initial therapy for newly diagnosed multiple myeloma: an open-label randomised controlled trial. *The Lancet Oncology*, 11, 29–37.

Royston, P. and Parmar, M. (2011) The use of restricted mean survival time to estimate the treatment effect in randomized clincial trials when the proportional hazards assumption is in doubt. *Journal of American Statistical Association*, 30, 2409–2421.

Stute, W. (1995) The central limit theorem under random censorship. *The Annals of Statistics*, 422–439.

Therneau, T.M. and Grambsch, P.M. (2013) *Modeling Survival Data: Extending the Cox Model*. Berlin–Heidelberg: Springer Science & Business Media.

Tian, L., Fu, H., Ruberg, S.J., Uno, H. and Wei, L.-J. (2018) Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations. *Biometrics*, 74, 694–702.

Tian, L., Zhao, L. and Wei, LJ. (2014) Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics*, 15, 222–233.

Trinquart, L., Jacot, J., Conner, S.C. and Porcher, R. (2016) Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *Journal of Clinical Oncology*, 34, 1813–1819.

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., Skali, H., Solomon, S., Jacobus, S., Hughes, M., Packer, M. and Wei, L.J. (2014) Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*, 32, 2380–2385.

Uno, H., Wittes, J., Fu, H., Solomon, S.D., Claggett, B., Tian, L., Cai, T., Pfeffer, M.A., Evans, S.R. and Wei, L.-J. (2015) Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies. *Annals of Internal Medicine*, 163, 127–134.

Wei, J. and Wu, J. (2020) Cancer immunotherapy trial design with cure rate and delayed treatment effect. *Statistics in Medicine*, 39, 698–708.

Ying, Z. (1989) A note on the asymptotic properties of the product-limit estimator on the whole line. *Statistics & Pobability Letters*, 7, 311–314.

Zhao, L., Claggett, B., Tian, L., Uno, H., Pfeffer, M.A., Solomon, S.D., Trippa, L. and Wei, L. (2016) On the restricted mean survival time curve in survival analysis. *Biometrics*, 72, 215–221.

Zhao, L., Tian, L., Uno, H., Solomon, S., Pfeffer, M., Schindler, J. and Wei, L.J. (2012) Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clinical Trials*, 9, 570–577.

Zucker, D. (1998) Restricted mean life with covariates: modification and extension of a useful survival analysis method. *Journal of American Statistical Association*, 93, 702–709.

How to cite this article: Tian L, Jin H, Uno H, Lu Y, Huang B, Anderson KM, Wei LJ. On the empirical choice of the time window for restricted mean survival time. *Biometrics*. 2020;76:1157–1166. https://doi.org/10.1111/biom.13237

APPENDIX

Theorem A1. *Suppose that* $(X_i, \delta_i) = (T_i \wedge C_i, I(T_i \leq C_i)), i = 1, \dots, n,$ *are* n *independent identically distributed copies of* $(X, \delta) = (T \wedge C, I(T \leq C))$*, where* T *and* C *are independent failure time and censoring time, respectively. Let* $[0, \tau_C]$ *be the support of* X *and* $P(T \geq \tau_C) > 0$*. If*

$$\int_0^{\tau_C} \frac{\left(\int_t^{\tau_C} S(u)du\right)^2 d\Lambda(t)}{G(t)} < \infty,$$

then

$$P\left\{D(\hat{\tau}_C) \leq t\right\} - P\left\{\int_0^{\tau_C} Z(u)S(u)du \leq t\right\} = o(1)$$

for any t*, as* $n \to \infty$*, where* $\Lambda(t)$ *is the cumulative hazard function of* T*,* $S(t)$ *is the continuous survival function of* T*,* $G(t) = P(X \geq t)$*,* $D(t) = \sqrt{n}\{\hat{R}(t) - R(t)\}$*,* $R(t) = \int_0^t S(u)du$*,* $\hat{\tau}_C = \max\{X_1, \dots, X_n\}$*, and* $Z(u)$ *is a continuous independent increment Gaussian process on* $[0, \infty)$ *with mean zero and a covariance function of*

$$\text{cov}\left\{Z(u_1), Z(u_2)\right\} = \int_0^{u_1 \wedge u_2} \frac{d\Lambda(s)}{G(s)}.$$

Furthermore, the variance of $\int_0^{\tau_C} Z(u)S(u)du$ *is*

$$\int_0^{\tau_C} \frac{\left(\int_t^{\tau_C} S(u)du\right)^2 d\Lambda(t)}{G(t)}.$$

Proof. Let

$$h(t) = \int_t^{\tau_C} S(u)du,$$

which is clearly a nonnegative continuous nonincreasing function and $dh(t) = -S(t)dt$. Under Condition (2), Theorem 2.1 of Gill (1983) suggests that the stochastic process

$$\int_0^{t \wedge \hat{\tau}_C} \frac{\sqrt{n}\{\hat{F}(u) - F(u)\}}{S(u)} dh(u)$$
$$= \int_0^{t \wedge \hat{\tau}_C} \sqrt{n}\{\hat{S}(u) - S(u)\}du$$
$$= D(t \wedge \hat{\tau}_C) \to \int_0^t Z(u)S(u)du$$

weakly for $t \in [0, \tau_C]$, as $n \to \infty$. Let $t = \tau_C$, we have

$$D(\tau_C \wedge \hat{\tau}_C) = D(\hat{\tau}_C) \to \int_0^{\tau_C} Z(u)S(u)du$$

in distribution. Finally,

$$\text{Var} \int_0^{\tau_C} Z(u)S(u)du$$
$$= \int_0^{\tau_C} \int_0^{\tau_C} S(u)S(v)E\{Z(u)Z(v)\}dudv$$
$$= \int_0^{\tau_C} \left\{\int_0^u \frac{d\Lambda(s)}{G(s)}\right\} d\left\{-h(u)^2\right\} = \int_0^{\tau_C} \frac{h(u)^2 d\Lambda(u)}{G(u)}.$$

□