Received: 26 March 2017 | Revised: 13 August 2017 | Accepted: 10 November 2017

DOI: 10.1002/pst.1846

MAIN PAPER

WILEY

Comparison of the restricted mean survival time with the hazard ratio in superiority trials with a time-to-event end point

Bo Huang[1] | Pei-Fen Kuan[2]

[1] Pfizer Inc, Groton, CT, USA

[2] Department of Applied Math and Statistics, Stony Brook University, Stony Brook, NY, USA

Correspondence
Bo Huang, Pfizer Inc, 445 Eastern Point Rd, Groton, CT 06340, USA.
Email: bo.huang@pfizer.com

With the emergence of novel therapies exhibiting distinct mechanisms of action compared to traditional treatments, departure from the proportional hazard (PH) assumption in clinical trials with a time-to-event end point is increasingly common. In these situations, the hazard ratio may not be a valid statistical measurement of treatment effect, and the log-rank test may no longer be the most powerful statistical test. The restricted mean survival time (RMST) is an alternative robust and clinically interpretable summary measure that does not rely on the PH assumption. We conduct extensive simulations to evaluate the performance and operating characteristics of the RMST-based inference and against the hazard ratio–based inference, under various scenarios and design parameter setups. The log-rank test is generally a powerful test when there is evident separation favoring 1 treatment arm at most of the time points across the Kaplan-Meier survival curves, but the performance of the RMST test is similar. Under non-PH scenarios where late separation of survival curves is observed, the RMST-based test has better performance than the log-rank test when the truncation time is reasonably close to the tail of the observed curves. Furthermore, when flat survival tail (or low event rate) in the experimental arm is expected, selecting the minimum of the maximum observed event time as the truncation timepoint for the RMST is not recommended. In addition, we recommend the inclusion of analysis based on the RMST curve over the truncation time in clinical settings where there is suspicion of substantial departure from the PH assumption.

KEYWORDS

log-rank test, proportional hazard, restricted mean survival time, time to event

1 | INTRODUCTION

In a randomized clinical trial with a time-to-event end point, 1 primary objective is to quantify or measure the relative difference between the survival curves of the randomized arms, which is routinely characterized by a constant hazard ratio (HR) from the Cox proportional hazard (PH) model, under the assumption that the ratio of the 2 hazard functions is constant over time. The log-rank test is the most powerful nonparametric test for detecting a PH alternative and thus is the most commonly used testing procedure.

When the PH assumption holds, the HR captures the relative difference between 2 survival curves. However, the clinical meaning of such a ratio estimate is difficult, if not impossible, to interpret when the underlying PH assumption is violated (ie, the HR is not constant over time). When the HR varies over time, its value derived from the Cox-PH model depends on the accrual distribution, dropout pattern and the study follow-up time, leading to different trial results and parameter estimates in different studies even if patients come from the same population and survival curves are identical. With non-PH, the log-rank test may no longer provide the desired level of power.

Substantial departure from the PH assumption has been a common observation in the development of oncology drugs in recent years, with the emergence of targeted therapies and cancer immunotherapies. The identification of certain molecular mechanisms has led to the development of targeted agents against different families of growth factors and receptors. Targeted therapies can lead to fast tumor regression (or sometimes disease stabilization) by directly targeting the oncogenic driver mutations with fewer side effects than chemotherapies. However, cancer cells can develop resistance to such treatments through mutation and resume rapid growth. Early separation of Kaplan-Meier (KM) curves ($HR < 1$, treatment vs control) followed by subsequent crossing ($HR > 1$) is a clinical issue not uncommon for some of these drugs. On the other hand, immuno-oncology has emerged as a new prominence in cancer research with distinct mechanisms of action. Immunotherapies work by harnessing the immune system to induce antitumor response, with an indirect effect on cancer cells. Because of the delayed and durable antitumor effect on cancer cells, the survival curves may take a while to separate, and the immunotherapy agent curve can have a very long tail.

The restricted mean survival time (RMST) is a robust and clinically interpretable summary measure of the survival time distribution that does not rely on the PH assumption. Unlike the median survival time, it is estimable even under heavy censoring. There has been considerable methodological research (eg, Zucker,[1] Royston and Parmar,[2,3] and Uno et al[4,5]) on the use of RMST to estimate treatment effects as an alternative to the HR approach. The RMST methodology is applicable independent of the PH assumption, and a test of the difference or ratio between the RMST for the experimental arm and the control arm may be more appropriate to determine superiority with respect to the time-to-event end point. The RMST depends on the selection of cutoff (truncation) time t_{cut}, which needs to be prespecified to avoid selection bias (after seeing the data). Common selections include fixed landmark times of clinical relevance (eg, x-year), minimum of the largest observed event time in each of the 2 groups, or minimum of the largest observed time (event or censoring) in each of the 2 groups.

Trinquart et al[6] analyzed the efficacy results from 54 randomized controlled trials published in 5 leading oncology journals. They found that the HR- and RMST-based measures were in agreement in terms of the statistical significance of the effect, except in 1 case.

Although there has been considerable research in the literature on the statistical inference and practical considerations of the RMST using real clinical trial examples and with simulations under a specific non-PH circumstance motivated by case studies,[2-6] to our knowledge, there is no published work that systematically evaluates the properties and operating characteristics of the RMST-based inference under various PH and non-PH scenarios and trial parameter setups based on simulations. In this paper, we conduct an extensive simulation study comparing the performance and operating characteristics of the RMST with the HR, including treatment effect estimation and statistical power. A wide spectrum of potential scenarios under both PH assumptions and non-PH assumptions that could be encountered are investigated. In Section 2, we give a brief overview of the HR and the log-rank test and provide a statistical interpretation of the HR under non-PH scenarios. The definition and some notations is introduced for the RMST. A general inference procedure is provided for this method for estimation and hypothesis testing. In Section 3, a simulation study evaluating the RMST and the HR under PH and non-PH assumptions is described, and the results are summarized. We illustrate the use of the RMST method with a case study in acute lymphoblastic leukemia (ALL) in Section 4. Section 5 concludes with a discussion.

2 | METHOD

2.1 | HR and the log-rank test

The log-rank test and the Cox-PH model are the most commonly used method for the analysis of time-to-event data in randomized trials. Under the PH assumption, the log-rank test is the most powerful nonparametric test, and the HR can be interpreted as a constant relative measure of risk (hazard) over time, with hazard $\lambda(t) = \lim_{dt\to 0} Pr\{T \in (t, t+dt)|T > t\}/dt$.

When there is a substantial deviation from the PH assumption, the interpretation of the HR from the Cox-PH model is not straightforward. Assume the PH assumption holds in a piecewise fashion (K periods), based on the relationship between the HR estimate and the log-rank statistic Z:

$$\log(\hat{\mathrm{HR}}) \approx Z\sqrt{1/r(1-r)D_{\max}},$$

where $D_{\max}$ is the total number of events at the final analysis and r is the randomization ratio. By partitioning of Z, the HR on the log scale can be approximated as

$$\log(\hat{\mathrm{HR}}) \approx \sum_{i=1}^{K} p_i \log(\hat{\mathrm{HR}}_i),$$

where p_i and $\hat{\mathrm{HR}}_i$ are the proportions of events and estimated HR in each period in which the PH assumption holds. As a result, the HR can be interpreted as the weighted average of HR over time on the log scale. However, in the Cox-regression model, the weights depend on the censoring distribution that keeps changing in randomized clinical trials because it is a function of accrual, follow-up, and early dropout, leading to different trial results and parameter estimates in different studies even if survival curves are identical.

2.2 | Definition and inference procedure for the RMST

The RMST is a robust and clinically interpretable summary measure of the survival time distribution that does not rely on the PH assumption. The RMST μ of a random time-to-event variable T is the mean of the survival time $X = \min(T, t_{cut})$ truncated at a cutoff time $t_{cut} > 0$. It can be derived as the area under the survival curve $S(t) = P(T > t)$ from $t = 0$ to $t = t_{cut}$:

$$\mu(t_{cut}) = E(X) = \int_0^{t_{cut}} S(t)dt. \tag{1}$$

The variance term $\sigma^2(t_{cut})$ of X can also be derived accordingly using integration by part

$$\sigma^2(t_{cut}) = \mathrm{Var}(X) = 2\int_0^{t_{cut}} tS(t)dt - \left[\int_0^{t_{cut}} S(t)dt\right]^2. \tag{2}$$

A natural estimator for μ is

$$\hat{\mu}(t_{cut}) = \int_0^{t_{cut}} \hat{S}(t)dt, \tag{3}$$

where $\hat{S}(t)$ is the KM estimator for the survival function of T, a step function with mass at the time points $t_1, t_2, \ldots, t_D$. $\hat{\mu}(t_{cut})$ approximately follows a normal distribution with its variance term estimated below[7]:

$$V[\hat{\mu}(t_{cut})] = \sum_{i=1}^{D}\left[\int_{t_i}^{t_{cut}} \hat{S}(t)dt\right]^2 \frac{d_i}{Y_i(Y_i - d_i)}, \tag{4}$$

where d_i and Y_i are the number of events and number of patients at risk at t_i, respectively.

In a randomized 2-arm trial with survival function $S_T(t)$ and $S_C(t)$ for the treatment arm and control arm, respectively, the difference in RMST between arms can be estimated as

$$\int_0^{t_{cut}} [\hat{S}_T(t) - \hat{S}_C(t)]dt, \tag{5}$$

with estimated variance term $V[\hat{\mu}_T(t_{cut})] + V[\hat{\mu}_C(t_{cut})]$.

Alternatively, analogous to the HR as a measurement of the relative risk of event hazard, a similar measurement for RMST is the ratio of RMST between arms (control versus treatment), with ratio < 1 indicating survival improvement in the treatment arm. Unlike the HR, the RMST ratio does not rely on any model assumption, which can be estimated as

$$\frac{\int_0^{t_{cut}} \hat{S}_C(t)dt}{\int_0^{t_{cut}} \hat{S}_T(t)dt}, \tag{6}$$

with variance term estimated using the delta method.

For ease of exposition and without loss of generality, the term "survival time" refers to the time to event or event-free time for any event end point of interest (death, disease progression, relapse, etc).

3 | A SIMULATION STUDY TO COMPARE THE RMST WITH THE HR UNDER PH AND NON-PH ASSUMPTIONS

3.1 | Notations and simulation setup

In this simulation study, survival time is assumed to follow a piecewise exponential distribution with piecewise constant hazard

$$\lambda_k(t|\tau_0, \tau_1, \ldots, \tau_J) = \sum_{j=1}^{J} I_{[\tau_{j-1},\tau_j)}(t)\lambda_{kj}, \tag{7}$$

where $k = T, C$ as label for treatment arm and control arm, respectively. $I_A(t)$ is an indication function with value 1 if $t \in A$ and 0 if $t \notin A$. Time $\tau_1, < \ldots, < \tau_{J-1}$ is the change points for the piecewise exponential model with $\tau_0 = 0$ and $\tau_J = \infty$ as boundary values for formulation purpose.

The survival function as the probability of survival at time t can be derived as follows:

$$\begin{aligned} S_k(t) &= \exp\left(-\int_0^t \lambda_k(v|\tau_0, \tau_1, \ldots, \tau_J)dv\right) \\ &= \begin{cases} \exp\left(-\lambda_{k1}t\right), & 0 \le t < \tau_1 \\ \vdots & \vdots \\ \exp\left(-\left[\sum_{l=1}^{j-1}\lambda_{kl}(\tau_l - \tau_{l-1})\right] - \lambda_{kj}(t - \tau_{j-1})\right), & \tau_{j-1} \le t < \tau_j \\ \vdots & \vdots \\ \exp\left(-\left[\sum_{l=1}^{J-1}\lambda_{kl}(\tau_l - \tau_{l-1})\right] - \lambda_{kJ}(t - \tau_{J-1})\right), & \tau_{J-1} \le t < \infty \end{cases} . \end{aligned} \tag{8}$$

TABLE 1 Simulation scenarios (scenarios 1-12) in a randomized (1:1) clinical trial under PH and non-PH assumptions for the comparison of statistical measures using the RMST versus the HR

PH assumption	Sc 1:
	Sc 2:
		N = 300, *NE* = 200, *HR* = 0.67
		N = 600, *NE* = 500, *HR* = 0.8

Non-PH assumption: no/small early separation, late separation	Sc 3:	N = 450, *NE* = 300, *HR* = 1 up to 15 months, *HR* = 0.02 after 15 months
	Sc 4:	N = 250, *NE* = 150, *HR* = 1 up to 10 months, *HR* = 0.02 after 10 months
	Sc 5:	N = 700, *NE* = 500, *HR* = 1 up to 10 months, *HR* = 0.5 after 10 months
	Sc 6:	N = 300, *NE* = 200, *HR* = 1 up to 5 months, *HR* = 0.75 from 5 to 15 months, 0.02 after 15 months
	Sc 7:	N = 300, *NE* = 200, *HR* = 0.85 up to 15 months, *HR* = 0.05 after 15 months
Non-PH assumption: crossing hazards	Sc 8:	N = 450, *NE* = 300, *HR* = 0.67 up to 15 months, *HR* = 1.2 after 15 months
	Sc 9:	N = 700, *NE* = 250, *HR* = 0.67 up to 15 months, *HR* = 1.2 after 15 months
	Sc 10:	N = 350, *NE* = 250, *HR* = 0.1 up to 5 months, *HR* = 0.5 from 5 to 10 months, 2 after 10 months
	Sc 11:	N = 600, *NE* = 500, *HR* = 0.75 up to 24 months, *HR* = 1 from 24 to 30 months, 1.5 after 30 months
	Sc 12:	N = 600, *NE* = 500, *HR* = 0.75 up to 10 months, *HR* = 1.2 from 10 to 20 months, 1.5 after 20 months

Abbreviations: HR, hazard ratio; PH, proportional hazard; RMST, restricted mean survival time; Sc, scenario. Without loss of generality, the control-arm survival time is assumed to follow an exponential distribution with median of 10 months. N, sample size; *NE*, event size. Accrual is uniform, and accrual duration is 24 months for N $\le$ 500 and 36 months for N $>$ 500.

The simulation of survival time T for individual patient can be conducted by randomly drawing samples from $U(0, 1)$ and back transformed using the inverse function $S_k^{-1}(U)$.

We further assume that the dropout (or loss-to-follow-up) time variable Z follows an exponential distribution with rate parameters λ_{DT} in the test arm and λ_{DC} in the control arm. Let Y denote the patient entry time to the study. Because treatment assignment is randomized, the distribution of Y is the same in both groups. It is assumed that the distribution of (T, Z) is stationary (ie, does not depend on Y) and the accrual and event times from different patients are independent. It is also assumed that T is independent of Z.

We compare the performance and operating characteristics of statistical measurements and hypothesis tests based on the HR and the RMST in a clinical trial setting of a randomized (1:1) 2-arm study with a survival endpoint. Twelve simulation scenarios are formulated under both PH assumption and non-PH assumption to render a wide spectrum of potential scenarios that can be encountered in clinical trials (Table 1 and Figure 1). Scenarios 1 and 2 are under the PH assumption. Scenarios 3 to 7 are under the non-PH assumption with no or small early-on separation but large and late separation (long tail in the treatment arm). Scenarios 8 to 12 are under the non-PH assumption with crossing hazards or belly-shape curves. Without loss of generality, the control-arm survival time is assumed to follow an exponential distribution with median of 10 months.

For scenarios 3 to 7, it is expected that the first change point of curves' separation has an impact on the performance. We look at 3 different change points for the curves starting to separate: 5, 10, and 15 months.

Accrual rate is assumed to be constant with 24 months of accrual time for sample size ≤ 500 and 36 months of accrual time for sample size >500. Patient dropout follows an exponential distribution with hazard rates of 0.001, 0.01, or 0.003.

The RMST depends on the restricted (truncation) time t_{cut} to have a closed-form area under the survival curves. It is recommended that t_{cut} be prespecified to minimize selection bias and to protect the integrity of the trial. Furthermore, t_{cut} ought to be clinically meaningful and closer to the end of the study follow-up so that most survival outcomes will be covered by the time interval. The selection of t_{cut} was briefly discussed in several papers in the literature.[2,3,6] In our simulation study, to have an objective and fair comparison with the HR and the log-rank test, the restricted time point t_{cut} is linked to the data when the primary analysis based on the log-rank test is performed and is prespecified as either (a) minimum of the maximum observed event time of each arm (minimax event time) or (b) minimum of the maximum observed (event or censored) time (minimax observed time) of each arm.

We also evaluate the RMST as a function of t_{cut} over time under the non-PH scenarios (ie, RMST curve[8]). The RMST curve is constructed for the difference in RMSTs and ratio of RMSTs between the treatment and the control arms. It provides a temporal profile of RMSTs for evaluating the benefit of the experimental treatment over the control treatment over time and overcomes the restriction of selecting a single truncation time t_{cut}.

3.2 | Simulation results

For each scenario, 10 000 simulations are performed to evaluate the performance and operating characteristics of each method. The results are summarized in Tables 2 to 4 and Figure 2.

Under the PH scenarios (scenarios 1-2), the log-rank test remains the most powerful test regardless of the treatment effect and dropout patterns. However, the RMST-based test performs almost equally well, and the average difference in power is negligible. There is a little difference in power by selecting t_{cut} based on event time or observed (event or censored) time. Nevertheless, in scenario 1, when the treatment effect is large ($HR = 0.67$), using the minimax observed time as the cutoff leads to slightly higher power. On the other hand, when the treatment effect is small ($HR = 0.80$) as in scenario 2, allowing a longer follow-up by using the minimax observed time as the cutoff does not lead to higher power.

Under the non-PH scenarios of late separation of curves (scenarios 3-7), the choice of t_{cut} has a significant impact on the statistical power of the RMST-based test. This is because in these scenarios, the experimental arm has a long survival tail with few events, which leads to a substantial difference between the minimax event time and minimax observed time (also shown in Figure 2). When t_{cut} is equal to the minimax of the observed times, the RMST-based test has much higher power compared to the log-rank test regardless of the dropout pattern and treatment effect, with an average of 10% to 50% in absolute improvement over the log-rank test. On the other hand, the use of minimax event time, despite being recommended in the literature,[6,9] does not produce satisfactory results and instead results in substantial power loss due to omitting a large portion of the late-separating curves.

The relative improvement in the power of the RMST test over the log-rank test using the minimax observe time as t_{cut} partially depends on the change point for late separation, keeping other parameters constant. Based on additional

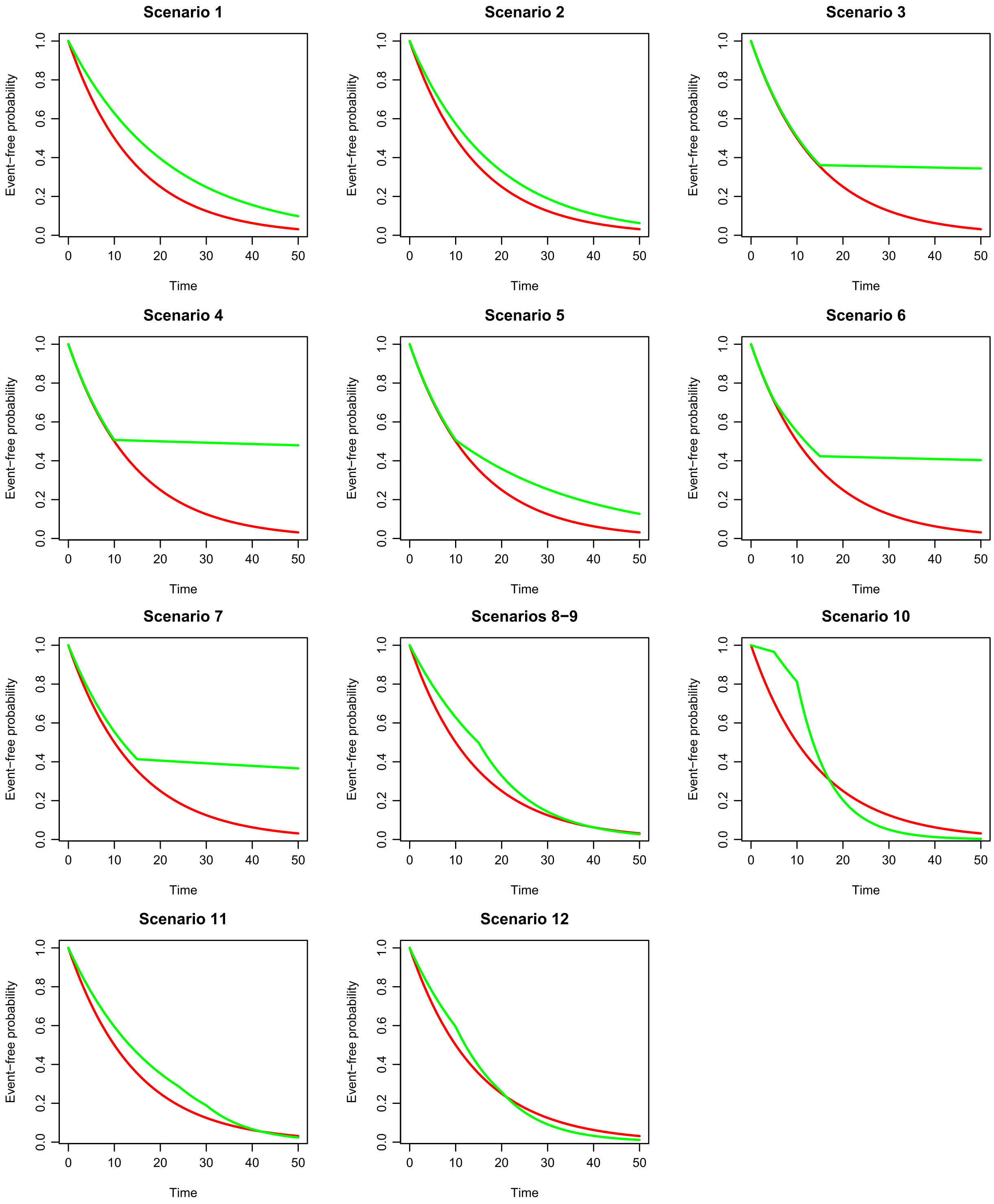

FIGURE 1 Simulation scenarios of event-free probability by time for a randomized (1:1) clinical trial comparing the treatment arm (green) with the control arm (red) under proportional hazard (PH) and non-PH assumptions for the comparison of statistical measures using the restricted mean survival time versus the hazard ratio

simulations (not shown) with varying change points for late separation, earlier separation is associated with less gain in power for the RMST test versus the log-rank test. For both tests, the power increases when the separation of curves occurs earlier.

Under the non-PH scenarios of crossing hazards or belly-shape curves (scenarios 8-12), the operating characteristics of these methods display different patterns. For belly-shape curves (scenarios 8-9 and 11) where HR reverses in the midst of

TABLE 2 Simulation results under scenarios 1 and2 (PH assumption) for the comparison of statistical measures using the RMST versus using the HR

		Log-Rank Test		**RMST Test (Event)**				**RMST Test (Observed)**				**Study**
		HR	**Power**	**Diff (m)**	**Ratio**	**Power**	**t_{cut} (m)**	**Diff (m)**	**Ratio**	**Power**	**t_{cut} (m)**	**Duration (m)**
Sc 1	A	0.677	80.3	3.11	0.799	79.2	26.7	3.67	0.780	80.2	31.2	32.7
	B	0.678	79.9	3.27	0.793	79.1	28.0	3.85	0.774	79.3	32.9	35.0
	C	0.678	79.8	3.29	0.793	79.0	28.0	3.88	0.773	79.6	33.0	34.9
Sc 2	A	0.802	69.8	2.61	0.841	69.6	41.5	2.82	0.833	69.4	46.9	50.3
	B	0.802	69.7	2.77	0.835	68.9	45.9	2.95	0.829	68.6	51.9	58.4
	C	0.803	69.7	2.82	0.833	69.5	46.6	3.01	0.826	69.3	52.7	58.8

Abbreviations: HR, hazard ratio; PH, proportional hazard; RMST, restricted mean survival time; Sc, scenario. The analyses are conducted at the same time for both methods. RMST test (event) refers to setting the truncated time t_{cut} as the minimum of the maximum observed event time of each treatment arm. RMST test (observed) refers to setting the truncated time t_{cut} as the minimum of the maximum observed (event or censored) time of each treatment arm, with power (%) as the average of the two. Patient dropout follows an exponential distribution with hazard rates: (A) $\lambda_{DT} = \lambda_{DC} = 0.0001$; (B) $\lambda_{DT} = 0.003$, $\lambda_{DC} = 0.01$; and (C) $\lambda_{DT} = 0.01$, $\lambda_{DC} = 0.003$, where λ_{DT} and λ_{DC} are the dropout hazard rates in the treatment and control arms, respectively.

TABLE 3 Simulation results under scenarios 3 to 7 (non-PH assumption: no/small early separation, late separation) for the comparison of statistical measures using the RMST versus using the HR

		Log-Rank Test		**RMST Test (Event)**				**RMST Test (Observed)**				**Study**
		HR	**Power**	**Diff (m)**	**Ratio**	**Power**	**t_{cut} (m)**	**Diff (m)**	**Ratio**	**Power**	**t_{cut} (m)**	**Duration (m)**
Sc 3	A	0.847	33.0	0.26	0.982	9.6	17.9	2.15	0.860	49.4	30.9	31.8
	B	0.824	41.3	0.46	0.969	13.7	19.3	2.78	0.830	61.4	33.4	34.8
	C	0.833	38.8	0.33	0.978	10.8	18.3	2.61	0.837	58.4	32.8	33.9
Sc 4	A	0.674	67.1	0.66	0.955	17.4	14.4	4.35	0.751	79.0	29.7	31.2
	B	0.642	73.9	1.12	0.930	24.7	16.3	5.40	0.717	85.2	32.4	34.6
	C	0.662	69.1	0.73	0.952	18.3	14.5	4.93	0.731	83.0	31.2	33.0
Sc 5	A	0.812	65.4	2.48	0.844	71.6	35.5	3.05	0.819	79.6	40.4	42.1
	B	0.807	67.7	2.69	0.835	75.2	37.3	3.29	0.809	82.5	42.7	45.5
	C	0.808	67.3	2.68	0.835	74.3	37.1	3.31	0.808	81.9	42.7	45.0
Sc 6	A	0.698	70.9	1.01	0.924	23.6	18.5	4.22	0.761	79.6	32.9	34.4
	B	0.656	80.2	1.86	0.887	35.0	22.0	5.78	0.711	87.9	37.5	40.3
	C	0.676	75.7	1.23	0.914	26.7	19.4	5.12	0.730	85.0	35.6	37.5
Sc 7	A	0.703	70.8	1.59	0.885	37.9	21.2	4.02	0.768	78.8	32.4	33.8
	B	0.670	78.3	2.43	0.847	49.9	24.9	5.14	0.730	86.0	36.0	38.5
	C	0.683	75.1	1.90	0.871	42.1	22.5	4.78	0.741	83.8	34.9	36.8

Abbreviations: HR, hazard ratio; PH, proportional hazard; RMST, restricted mean survival time; Sc, scenario. The analyses are conducted at the same time for both methods. RMST test (event) refers to setting the truncated time t_{cut} as the minimum of the maximum observed event time of each treatment arm. RMST test (observed) refers to setting the truncated time t_{cut} as the minimum of the maximum observed (event or censored) time of each treatment arm, with power (%) as the average of the two. Patient dropout follows an exponential distribution with hazard rates: (A) $\lambda_{DT} = \lambda_{DC} = 0.0001$; (B) $\lambda_{DT} = 0.003$, $\lambda_{DC} = 0.01$; and (C) $\lambda_{DT} = 0.01$, $\lambda_{DC} = 0.003$, where λ_{DT} and λ_{DC} are the dropout hazard rates in the treatment and control arms, respectively.

follow-up time ($HR < 1$ followed by $HR > 1$) but the treatment curve is consistently above the control curve throughout the course (Figure 1), the RMST-based and the log-rank tests have similar performance in terms of power. When t_{cut} is equal to the minimax event time, the RMST-based test performs slightly better; ehen t_{cut} is equal to the minimax observed time, the log-rank performs slightly better. For scenario 10 where crossing hazards are observed with the treatment curve above the control curve early on but the trend reverses later and the control curve moves above the treatment curve, the RMST-based test performs better with much higher power than that of the log-rank test regardless of the choice of t_{cut}. Under scenarios 8 to 11, for the RMST-based test, the selection of the minimax event time as t_{cut} leads to higher power than selecting the minimax observed time as t_{cut}.

Scenario 12 is another scenario of crossing hazards and crossing curves. However, the early positive separation mostly cancels out the late negative separation, rendering the RMST difference close to 0 and the RMST ratio and the estimated overall HR close to 1. Under this scenario, both methods have low power as expected (may also be interpreted as the type 1

TABLE 4 Simulation results under scenarios 8 to 12 (non-PH assumption: crossing hazards and belly shape) for the comparison of statistical measures using the RMST versus using the HR

		Log-Rank Test		RMST Test (Event)				RMST Test (Observed)				Study
		HR	Power	Diff (m)	Ratio	Power	t_{cut} (m)	Diff (m)	Ratio	Power	t_{cut} (m)	Duration (m)
Sc 8	A	0.751	71.2	2.36	0.838	74.6	26.6	2.46	0.839	68.0	30.2	31.4
	B	0.754	69.8	2.39	0.839	72.4	27.6	2.47	0.840	64.6	31.6	33.2
	C	0.755	69.5	2.40	0.838	72.7	27.8	2.49	0.839	65.3	31.7	33.3
Sc 9	A	0.723	74.5	2.15	0.841	77.5	22.1	2.32	0.839	70.6	25.3	26.1
	B	0.722	74.8	2.16	0.841	76.4	22.4	2.33	0.839	69.1	25.8	26.8
	C	0.723	74.6	2.17	0.841	76.5	22.5	2.34	0.839	69.0	25.8	26.8
Sc 10	A	0.789	48.3	2.51	0.828	80.5	26.3	2.27	0.848	64.9	29.7	32.2
	B	0.801	44.6	2.45	0.833	75.5	27.0	2.19	0.854	58.3	30.7	33.8
	C	0.808	41.6	2.44	0.835	75.6	27.3	2.20	0.853	59.5	30.8	34.4
Sc 11	A	0.793	74.1	2.78	0.832	76.6	41.1	2.78	0.834	72.2	45.7	49.7
	B	0.801	69.7	2.76	0.835	72.4	44.3	2.74	0.838	67.7	49.3	56.6
	C	0.803	69.2	2.78	0.834	72.8	45.1	2.77	0.836	68.7	49.7	57.5
Sc 12	A	0.960	8.6	0.57	0.961	10.3	38.0	0.45	0.970	8.2	42.3	47.0
	B	0.969	7.7	0.49	0.967	9.1	40.5	0.37	0.976	7.1	45.1	52.4
	C	0.969	6.9	0.50	0.966	8.4	40.7	0.40	0.974	6.7	45.0	53.0

Abbreviations: HR, hazard ratio; PH, proportional hazard; RMST, restricted mean survival time; Sc, scenario. The analyses are conducted at the same time for both methods. RMST test (event) refers to setting the truncated time t_{cut} as the minimum of the maximum observed event time of each treatment arm. RMST test (observed) refers to setting the truncated time t_{cut} as the minimum of the maximum observed (event or censored) time of each treatment arm, with power (%) as the average of the two. Patient dropout follows an exponential distribution with hazard rates: (A) $\lambda_{DT} = \lambda_{DC} = 0.0001$; (B) $\lambda_{DT} = 0.003$, $\lambda_{DC} = 0.01$; and (C) $\lambda_{DT} = 0.01$, $\lambda_{DC} = 0.003$, where λ_{DT} and λ_{DC} are the dropout hazard rates in the treatment and control arms, respectively.

error rate). The overall HR as a single summary statistics approximated by the weighted periodical HRs (on the log scale) cannot be easily interpreted (similar arguments can be made for other non-PH scenarios). Clearly, there is a treatment effect in the first 10 months ($HR = 0.75$), but the effect disappears and reverses after that ($HR = 1.2$ and $HR = 1.5$ afterwards). In practice, such results are very difficult to interpret from a benefit-risk perspective and usually indicate a heterogeneous mixture of subgroups with effect sizes in opposite directions. Subgroup analyses defined by biomarkers and other baseline characteristics should be conducted to identify patients who may benefit from the experimental treatment.

Three different dropout patterns are evaluated: (A) $\lambda_{DT} = \lambda_{DC} = 0.0001$; (B) $\lambda_{DT} = 0.003, \lambda_{DC} = 0.01$; and (C) $\lambda_{DT} = 0.01, \lambda_{DC} = 0.003$. We observe that patient dropout has an impact on the results of HR, RMST difference, RMST ratio, power, and study duration. However, it has minimal effect on the relative performance of the RMST-based test versus the log-rank test.

An interesting observation from the simulation results is that the RMST ratio (control vs treatment) is generally higher than the HR (treatment vs control), even if the statistical powers point to the opposite direction. Trinquart et al[6] analyzed the efficacy results from 54 randomized controlled trials published in 5 leading oncology journals. They found that on average, the HR provided larger treatment effect estimates than the ratio of RMST. Our simulation results are consistent with their findings in a broader range of clinical scenarios. This finding has some clinical implications. Unlike the HR, the RMST ratio derived from the KM curves does not rely on any model assumption and as such is more accurate in estimating the true treatment benefit. Trinquart et al[6] warned that in practice, clinicians may apply a similar standard when interpreting the magnitude of HRs and ratios of RMST, despite differences in the meanings of the 2 relative measures. It may not be a fair comparison as one term is the ratio of the hazards and the other is the ratio of the means.

The outcome and performance of the RMST difference or RMST ratio depend on the selection of t_{cut}. We further evaluate the statistical summaries in terms of the RMST curve for both the difference and the ratio as a function of t_{cut} over time for the non-PH scenarios, as illustrated in Figure 2. The Figure demonstrates that the minimax observed time is a better choice as t_{cut} for late-separating curves and long survival tail in the treatment arm. On the other hand, the selection between minimax event time and minimax observed time does not make significant difference for crossing hazards. The RMST curve is a useful and clinically meaningful visualization tool to evaluate the treatment benefit over time without the restriction of a single truncation time.

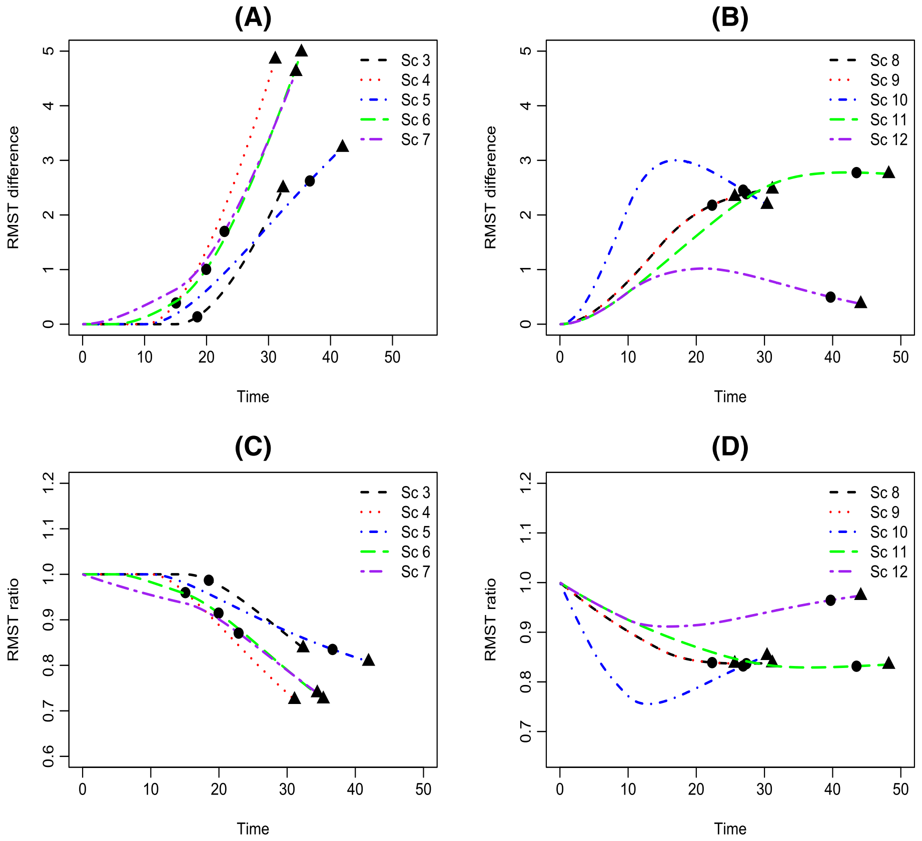

FIGURE 2 The RMST curves as a function of t_{cut} based on difference or ratio over time under scenarios 3 to 12. The average minimax event time from 10 000 simulations under each scenario is denoted by ●. The average minimax observed time from 10 000 simulations under each scenario is denoted by ▲. RMST, restricted mean survival time; Sc, scenario

4 | EXAMPLE: A PHASE 3 TRIAL OF INOTUZUMAB OZOGAMICIN IN ALL

Patients with ALL have poor prognosis. In this open-label, 2-group, randomized, phase 3 trial, patients 18 years of age or older were eligible for enrollment if they had relapsed or refractory, CD22-positive, Philadelphia chromosome–positive or –negative ALL. A total of 326 adults were randomly assigned (1:1 randomization ratio) to receive either inotuzumab ozogamicin (inotuzumab ozogamicin group) or standard intensive chemotherapy (standard-therapy group). The primary end points were complete remission (CR) and overall survival (OS), with 1-sided alpha level of 0.025 evenly split between the 2 primary end points.

The study was completed in 2016, and the results were published in the New England Journal of Medicine.[10] The primary analysis for CR was conducted in the first 218 patients. The rate of CR was significantly higher in the inotuzumab ozogamicin group than in the standard-therapy group (80.7% [95% CI, 72.1-88.7] vs 29.4% [95% CI, 21.0-38.8], $P < .001$).

In the intention-to-treat survival analysis, median OS was 7.7 months (95% CI, 6.0-9.2) in the inotuzumab ozogamicin group and 6.7 months (95% CI, 4.9-8.3) in the standard-therapy group, and the HR for death was 0.77 (97.5% CI, 0.58-1.03) (2-sided $P = .04$). Therefore, the second primary objective of showing significantly longer OS in the inotuzumab ozogamicin group than in the standard-therapy group, at a prespecified boundary of 2-sided $P = .0208$ (after adjusting for alpha spending at the interim analysis) was not met.[10]

However, data for OS appeared to depart from the PHs assumption (Figure 3). In fact, the shapes of the KM survival curves resemble those in simulation scenarios 3 to 7. An exploratory post hoc analysis of RMST was applied using the same data snapshot as the primary analysis of log-rank test. In this analysis, mean OS was longer in the inotuzumab ozogamicin group than in the standard-therapy group (mean [standard error]: 13.9 [1.10] months vs 9.9 [0.85] months; $P = .005$), with a clinically meaningful improvement in RMST of approximately 4 months. The RMST analysis was conducted with t_{cut} selected as the minimax of observed times in the treatment and control arms, which was approximately 38 months (Figure 3). Interestingly, should the minimax event time be selected, the truncation time would have moved

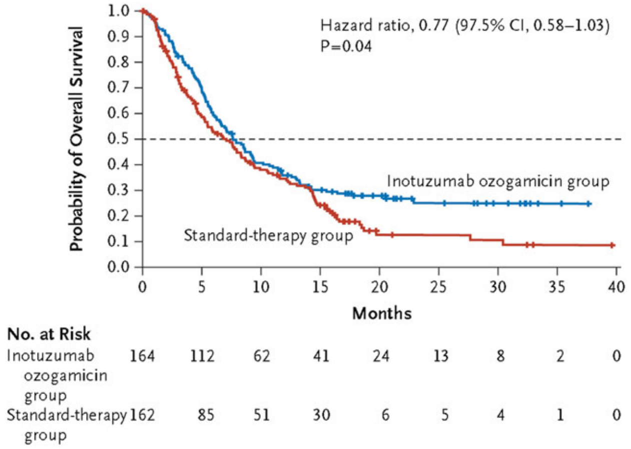

FIGURE 3 Overall survival Kaplan-Meier curves of the phase 3 randomized study in patients with relapsed or refractory, CD22-positive, Philadelphia chromosome (Ph)-positive or Ph-negative acute lymphoblastic leukemia. A total of 326 patients were 1:1 randomized to receive either inotuzumab ozogamicin (inotuzumab ozogamicin group) or standard intensive chemotherapy (standard-therapy group) (source: Kantarjian et al[10])

up to approximately 23 months, resulting in much information loss and higher *P* value due to missing out a large portion of the late-separated curves. This observation is consistent with our simulation results presented in section 3.2.

5 | DISCUSSION

With the emergence of novel cancer treatments such as targeted therapies and immunotherapies, observations of substantial departure from the PH assumption in clinical trials evaluating time-to-event end points are not uncommon. We conduct extensive simulations to evaluate the performance and operating characteristics of the RMST against the HR under various scenarios and design parameter setups. Under non-PH scenarios where late separation of survival curves is observed, the RMST-based test has better performance than the log-rank test in terms of power when the truncation time t_{cut} is reasonably close to the tail of the observed KM curves. Under non-PH scenarios of crossing hazards that lead to crossing curves (but the overall effect is still positive), the RMST test also performs better than the log-rank test. In other scenarios (PH and belly-shape scenarios), both the RMST-based and the log-rank tests have similar performance. We acknowledge that some of the scenarios under the non-PH assumption may look extreme. However, they are useful in delineating the nuances and variations of these methods under different situations and parameter setups, which otherwise may not be apparent.

In clinical trials, estimation of the treatment effect is as important as (if not more important than) the power for statistical inference. The RMST-based statistical measures derived from the KM estimates do not rely on any model assumptions. Thus, when there is departure from the PH assumption, the interpretation is still straightforward. In contrast, the HR varies with time, and the value derived from the Cox-PH model cannot be interpreted as the average HR across times. Furthermore, unlike median event-free times and time-specific probability end points, the RMST can capture the entire event-free distribution up to time t_{cut} as the area under the KM curve. Importantly, both the difference and the ratio in RMSTs provide a clinically meaningful summary of the group difference in a randomized study. Unlike the HR, the difference allows for quantifying the absolute survival difference and the magnitude of clinical benefit. The capability of dual presentation of both the relative and the absolute measures is an important benefit of using the RMST. Both measures (difference and ratio) have nearly identical performance in power (since the ratio is essentially the difference on the log scale). From a design perspective, the RMST difference may be preferred to the RMST ratio because an absolute increase in the mean of survival time is easier to interpret for the benefit-risk evaluation. The lack of an absolute measure from the HR is a major limitation in the evaluation of benefit-risk profile of the experimental drug, particularly when the expected mean or median event-free time from the control is small.

The RMST depends on the selection of truncation time t_{cut}, which needs to be prespecified to avoid selection bias. t_{cut} ought to be clinically meaningful and closer to the end of the study follow-up so that most survival outcomes will be covered by the time interval. To have an objective and fair comparison with the HR and the log-rank test, we look at 2 selections, minimum of the largest observed event time in each of the 2 groups (minimax event time) or minimum of the largest observed time (event or censoring) in each of the 2 groups (minimax observed time). We find that the performance of the RMST is sensitive to t_{cut} under the non-PH scenarios. For scenarios of late separation and long KM curve tail in the treatment arm, t_{cut} equal to the minimax event time could lead to poor outcomes, while for scenarios of crossing hazards and belly-shape curves, t_{cut} equal to the minimax event time performs better. The selection of t_{cut} equal to minimax observed time generally results in competitive and robust outcomes compared to the HR and the log-rank test.

If we design a study with the RMST as the primary analysis powered to detect a meaningful difference of 2 RMSTs, the selection of t_{cut} cannot be based on the minimax event time or minimax observed time when data are not available. Instead, t_{cut} should be a fixed timepoint. The time window $(0, t_{cut})$ should be large enough and expected to capture most of the survival curves for the RMST to be used as an adequate global summary statistic, which may be informed by considerations of both clinical significance and study feasibility. For example, if accrual is assumed to take 12 months, and patients will be followed for up to 24 months after the last randomized patient, a reasonable choice of t_{cut} may be 30 months. It is important to ensure that the potential follow-up time for a significant proportion of patients is adequate for estimating the RMST in the specified time window.

The need to prespecify a restricted or truncation time is a limitation of the RMST method especially since it is sensitive to the truncation time (as shown in our simulations). However, in the trial sample size calculation using the conventional log-rank test as the primary analysis, we also need to assume the patients' accrual profile and follow-up duration, which ultimately affect the estimation of the HR (section 2.1). The theoretical and practical considerations in designing a trial using the RMST is an interesting research topic and requires further work.

Because of censoring and early events, the number of patients in the later part of the curve may be small, resulting in increased variability of the curve shape by a small number of events. The RMST curve RMST(t) ($t \in [t_1, t_2]$) introduced earlier as an alternative summary to the survival function may be considered. The RMST curve can be constructed for each arm and for the difference or ratio in RMST between the treatment and the control arm. It provides a temporal profile of RMST or difference/ratio of RMSTs for evaluating the benefit of the experimental treatment over the control treatment by time and overcomes the restriction of selecting a single truncation time t_{cut}. The time interval $[t_1, t_2]$ can be selected to reflect the window of clinical relevance. For example, t_1 can be selected as the minimum of (median survival time for the experimental arm and median survival time for the control arm), and t_2 can be selected as the minimax observed time. We have evaluated the RMST curve under the non-PH scenarios in our simulations (Figure 2). Zhao et al[8] proposed inference based on simultaneous confidence bands for a single RMST curve and also the difference between 2 RMST curves. Compared to other conceptually similar plots such as HR by time, the RMST curve is easy to interpret and is clinically meaningful to characterize the treatment effect over time. Therefore, we recommend including this type of analysis for studies where there is suspicion of substantial departure from the PH assumption.

ORCID

Bo Huang http://orcid.org/0000-0002-3088-9328

REFERENCES

1. Zucker D. Restricted mean life with covariates: modification and extension of a useful survival analysis method. *J Am Stat Assoc*. 1998;93:702-709.
2. Royston P, Parmar M. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat Med*. 2011;30(19):2409-2421.
3. Royston P, Parmar M. 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 Method*. 2013;13(1):152-166.
4. 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(22):2380-2385.
5. Uno H, Wittes J, Fu H, et al. Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studiesalternatives to hazard ratios. *Ann Internal Med*. 2015;163(2):127-34.
6. Trinquart L, Jacot J, Conner S, Porcher R. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *J Clin Oncol*. 2016;34(15):1813-1819.
7. Klein J, Moeschberger M. Survival analysis: techniques for censored and truncated data. *Springer Science and Business Media*. 2005.
8. Zhao L, Claggett B, Tian L, et al. On the restricted mean survival time curve in survival analysis. *Biometrics*. 2016;72:215-221.
9. Uno H. Vignette for survRM2 package: comparing two survival curves using the restricted mean survival time. 2015.

10. Kantarjian H, DeAngelo D, Stelljes M, et al. Inotuzumab ozogamicin versus standard therapy for acute lymphoblastic leukemia. *New England J Med*. 2016;375(8):740-753.

How to cite this article: Huang B, Kuan P-F. Comparison of the restricted mean survival time with the hazard ratio in superiority trials with a time-to-event end point. *Pharmaceutical Statistics*. 2018;17:202–213. https://doi.org/10.1002/pst.1846