On assessing survival benefit of immunotherapy using long-term restricted mean survival time

Miki Horiguchi[1,2], Lu Tian[3], Hajime Uno[1,2,4]

[1]Department of Medical Oncology, Dana Farber Cancer Institute,

Boston, 02215, MA.

[2]Department of Internal Medicine, Harvard Medical School,

Boston, 02215, MA.

[3]Department of Biomedical Data Science, Stanford University School of Medicine,

Palo Alto, 94305, CA.

[4]Department of Data Science, Dana Farber Cancer Institute,

Boston, 02215, MA.

email: huno@ds.dfci.harvard.edu

SUMMARY: The pattern of the difference between two survival curves we often observe in randomized clinical trials for evaluating immunotherapy is not proportional hazards; the treatment effect typically appears several months after the initiation of the treatment (i.e., delayed difference pattern). The commonly used logrank test and hazard ratio estimation approach will be suboptimal concerning testing and estimation for those trials. The long-term restricted mean survival time (LT-RMST) approach is a promising alternative for detecting the treatment effect that potentially appears later in the study. A challenge in employing the LT-RMST approach is that it must specify a lower end of the time window in addition to a truncation time point that the RMST requires. There are several investigations and suggestions regarding the choice of the truncation time point for the RMST. However, little has been investigated to address the choice of the lower end of the time window. In this paper, we propose a flexible LT-RMST-based test/estimation approach that does not require users to specify a lower end of the time window. Numerical studies demonstrated that the potential power loss by adopting this flexibility was minimal, compared to the standard LT-RMST approach using a prespecified lower end of the time window. The proposed method is flexible and can offer higher power than the RMST-based approach when the delayed treatment effect is expected. Also, it provides a robust estimate of the magnitude of the treatment effect and its confidence interval that corresponds to the test result.

KEY WORDS: Delayed difference, hazard ratio, non-proportional hazards, versatile test, weighted logrank test

1. Introduction

The recent clinical success of the immune checkpoint inhibitors (CTLA-4, PD-1, and PD-L1 antagonists) has brought tremendous excitement and hope for patients with cancer (Emens et al., 2017). Cancer immunotherapy is now one of the primary treatment options for many cancer types; it has dramatically changed the treatment landscape. Nowadays, immunotherapy is one of the most exciting areas of cancer clinical research. In fact, as of November 26, 2021, 141 Phase 3 clinical trials for cancer immunotherapy are open and recruiting patients (*ClinicalTrials.gov*). It is expected that more immunotherapy trials will be conducted in the future for further indications or optimizations in combination with other therapies and biomarkers.

While other cancer treatments are usually developed to attack cancer directly, immunotherapy directs the immune system to do so. This unique mechanism has brought us several new challenges in assessing the efficacy of immunotherapy. Valuable summaries regarding statistical issues on immunotherapies were reported in recent publications (Chen, 2013; Huang, 2018). Among those new challenges, the one we specifically focus on in this paper is evaluating the treatment effect on time-to-event outcomes. In recent randomized clinical trials in oncology, the conventional logrank/hazard-ratio test/estimation method has been used in almost all trials (Uno et al., 2020). The logrank test is an asymptotically valid non-parametric test. It offers the highest power when the pattern of difference is the proportional hazards (PH). Also, the standard partial likelihood inference for hazard ratio (HR) provides a quantitative summary about the magnitude of treatment effect that corresponds to the logrank test. However, this routinely used approach is unlikely optimal for immunotherapy trials because we often observe that the PH assumption is unlikely valid. For example, Figure 1 shows the Kaplan-Meier curves for progression-free survival (PFS) with the data from a recently conducted randomized controlled trial to compare nivolumab plus ipilimumab

and sunitinib in patients with advanced renal cancer (Motzer et al., 2018). This represents a typical delayed difference pattern we often see in immunotherapy trials. In such non-proportional hazards (NPH) cases, the logrank test does not offer the highest power, and the interpretation of the HR estimate derived through the standard Cox's procedure is not obvious because it depends on an underlying study-specific censoring time distribution (Uno et al., 2014; Horiguchi et al., 2019).

[Figure 1 about here.]

Several alternative approaches to the conventional logrank/HR approach have been discussed to detect the delayed treatment effect. A simple approach would be using a weighted logrank (WLR) test that gives more weight to the tail part (Fleming and Harrington, 1991). For example, Xu et al. (2017) recently proposed the piecewise weighted logrank test that gives zero weight at early times when two survival curves seem identical. One of the drawbacks of the WLR test-based approach is that the treatment effect summary measure that corresponds to the test will be an HR-type measure and thus have the same issues as Cox's HR. Specifically, one may use the weighted Cox's model to estimate a weighted HR using the same weight used for the WLR test. The resulting weighted HR estimate will then be coherent to the test result. However, the adequacy of the model assumption would still be required for the resulting weighted HR to be independent of a study-specific censoring distribution. In addition, this approach does not convey the specific numbers from the two groups that yield the ratio. Reporting only a ratio without each group's absolute hazard "value" obscures the clinical significance of the magnitude of the treatment effect (Uno et al., 2014). Of note, the piecewise WLR test approach proposed by Xu et al. (2017) will have an additional significant limitation. This approach is essentially a landmark analysis and does not include all randomized patients in the analysis. Therefore, the analysis population used for inference may not represent the entire study population or the target population very

well. Also, because those patients who do not reach the landmark timepoint are excluded from the analysis, the random treatment allocation may not guarantee the compatibility of the treatment groups. Potential imbalance of patient characteristics between groups and generalizability will be of concern.

Another class of alternative approaches is based on restricted mean survival time (RMST). The RMST-based approach is gaining attention as an alternative to the conventional logrank/HR approach (Royston and Parmar, 2011, 2013; Uno et al., 2014, 2015; Chappell and Zhu, 2016; Saad et al., 2017). The RMST is defined as the area under the survival function and can be estimated non-parametrically via the Kaplan-Meier method. Therefore, it is robust and independent of study-specific censoring time distribution as long as the censoring is non-informative. However, the RMST-based approach is less powerful for detecting delayed treatment effects than the conventional logrank/HR approach (Tian et al., 2018). This limitation can be a critical barrier for the standard RMST-based approach to be employed for immunotherapy studies.

The long-term restricted mean survival time (LT-RMST) approach (Zhao et al., 2012; Horiguchi et al., 2018; Vivot et al., 2019) or equivalently, the window mean survival time approach (Paukner and Chappell, 2021) is a promising alternative for detecting the delayed treatment effect. A challenge in employing the LT-RMST approach is that it needs to prespecify the lower end of the time window as well as the upper end (i.e., truncation time point) that the standard RMST requires. However, unfortunately, at the design stage of a clinical trial, there would not be sufficient information to prespecify an optimal number for the lower end in practice. In this paper, to address this practical issue, we propose a flexible LT-RMST-based test/estimation approach that does not require users to specify a value for the lower end of the time window (Section 2). We also assess the performance of the proposed approach via extensive numerical studies (Section 3). We then present an

application of the proposed methods to the PFS data from the advanced renal cancer study (Section 4), followed by Remarks (Section 5).

2. Adaptive long-term restricted mean survival time procedure

Let T_j be a continuous non-negative random variable to denote the event time for group j ($j = 0, 1$). Let C_j denote the censoring time for group j. We assume that T_j is independent of C_j. Let $\{(T_{ji}, C_{ji});\ i = 1, \ldots, n_j\}$ denote independent copies from (T_j, C_j). Let $X_{ji} = T_{ji} \wedge C_{ji}$ and $\Delta_{ji} = I(T_{ji} \leqslant C_{ji})$, where $I(A)$ is the indicator function for event A. The observable data is then denoted by $\{(X_{ji}, \Delta_{ji});\ i = 1, \ldots, n_j,\ j = 0, 1\}$. We assume $p_j = \lim_{n\to\infty} n_j/n > 0$ for $j = 0, 1$, where $n = n_1 + n_0$.

The long-term restricted mean survival time (LT-RMST) was discussed in Zhao et al. (2012), Horiguchi et al. (2018), and Vivot et al. (2019). More recently, Paukner and Chappell (2021) also discussed it, calling it window mean survival time. Let $S_j(\cdot)$ be the survival function for T_j. The LT-RMST is defined as

$$R_j(\eta, \tau) = \int_\eta^\tau S_j(s)ds,$$

where $[\eta, \tau]$ is a fixed range on the study time ($0 \leqslant \eta < \tau$). This can be interpreted as mean survival time during the study time from η to τ. Also, standardized by the width of the time window,

$$\frac{R_j(\eta, \tau)}{\tau - \eta} \tag{1}$$

this can be interpreted as an average survival probability on $[\eta, \tau]$(Zhao et al., 2012; Horiguchi et al., 2018).

Let $\hat{S}_j(\cdot)$ be the Kaplan-Meier estimator for $S_j(\cdot)$. A non-parametric estimator for $R_j(\eta, \tau)$ is given by $\hat{R}_j(\eta, \tau) = \int_\eta^\tau \hat{S}_j(s)ds$, that is the area under $\hat{S}_j(t)$ on $t \in [\eta, \tau]$. One may consider a parametric model for estimating $S_j(\cdot)$, but we focus only on the non-parametric approach in this paper. For a between-group comparison, we consider $D(\eta, \tau) = R_1(\eta, \tau) - R_0(\eta, \tau)$

as a summary measure of the treatment effect magnitude. The asymptotic properties of $\hat{D}(\eta,\tau) = \hat{R}_1(\eta,\tau) - \hat{R}_0(\eta,\tau)$ has been shown in Zhao et al. (2012).

A practical issue on this LT-RMST approach is how to specify (η,τ). A method that is used in confirmatory trials should be a prespecified method. Thus, the requirement would be either specifying a specific (η,τ) or a rule to specify (η,τ) from the data. Regarding the truncation time point τ, there have been discussions and several proposals (Horiguchi et al., 2018; Eaton et al., 2020; Hasegawa et al., 2020; Tian et al., 2020). For example, Tian et al. (2020) provided the conditions for justifying empirically choosing the minimum of the maximum follow-up times from two groups as τ. Regarding η, however, few such investigations or discussions have been done. Although most immunotherapy studies share the same delayed separation pattern, the point of separation differs from study to study, and the pattern after the point of separation also differs from study to study. It would be challenging to specify a η at the design stage because we would not have sufficient information for it in most cases. Therefore, a procedure for choosing a reasonable η would be necessary for the LT-RMST approach to be eventually employed in practice. This paper extends the LT-RMST approach to select an optimal η data-dependently with the type I error rate controlled at a nominal level.

From results derived in Zhao et al. (2016), it is shown $W_n(\eta,\tau) = n^{1/2}\{\hat{D}(\eta,\tau) - D(\eta,\tau)\}$ converges weakly to a zero-mean Gaussian process indexed by $\eta \in [0,\tau)$. The covariance function of this process, $\mathrm{cov}\{W_n(\eta_1,\tau), W_n(\eta_2,\tau)\}$, is given in the Appendix A. For $\eta = \eta_1 = \eta_2$, we denote $\mathrm{cov}\{W_n(\eta,\tau), W_n(\eta,\tau)\} = V(W_n(\eta,\tau))$. Also, we denote estimates of $\mathrm{cov}\{W_n(\eta_1,\tau), W_n(\eta_2,\tau)\}$ and $V(W_n(\eta,\tau))$ by $\hat{\mathrm{cov}}\{W_n(\eta_1,\tau), W_n(\eta_2,\tau)\}$ and $\hat{V}(W_n(\eta,\tau))$, respectively, which can be calculated by replacing the unknown quantities involved in $\mathrm{cov}\{W_n(\eta_1,\tau), W_n(\eta_2,\tau)\}$ and $V(W_n(\eta,\tau))$ by their empirical counterparts.

Here, we assume that we cannot determine a value for η *a priori* but can determine a set of candidates for η or a range where η should be included. Let B denote such a set for η. For

example, expecting that the delayed treatment effect appears from around 6 months, one may set $B = \{b \mid b \in [0, 7 \text{ months}]\}$. For testing the null hypothesis, $H_0 : \ D(\eta, \tau) = 0$ for all $\eta \in B$, we propose the following test statistic,

$$Q = \sup_{b \in B} \left| \hat{D}(b, \tau) / \sqrt{\hat{V}(W_n(b, \tau))/n} \right|.$$

The reference distribution of Q under the null hypothesis can be derived by generating a zero-mean Gaussian process with the covariance function of the process $\{W_n(\eta, \tau) \ ; \eta \in B\}$. Let $\{ \ D_1^*(b, \tau), \ldots, D_M^*(b, \tau); b \in B\}$ be the set of the M sample paths generated randomly. For each path, $D_m^*(b, \tau)$, $m = 1, \ldots, M$, we calculate

$$Q_m^* = \sup_{b \in B} \left| D_m^*(b, \tau) / \sqrt{\hat{V}(W_n(b, \tau))/n} \right|.$$

A p-value is then given by $M^{-1} \sum_{m=1}^{M} I(Q_m^* > Q)$, and a corresponding $(1-\alpha)$ simultaneous confidence bands for $\{D(b, \tau); \ b \in B\}$ can be given by

$$\left\{ \hat{D}(b, \tau) \pm c_\alpha \sqrt{\hat{V}(W_n(b, \tau))/n} \ ; b \in B \right\},$$

where c_α is a smallest value that satisfies $M^{-1} \sum_{m=1}^{M} I(Q_m^* \geqslant c_\alpha) \leqslant \alpha$.

A treatment effect summary that is coherent to the test result is given as follows. In the above procedure, we choose a largest $\tilde{\eta}$ that satisfies

$$\left| \hat{D}(\tilde{\eta}, \tau) / \sqrt{\hat{V}(W_n(\tilde{\eta}, \tau))/n} \right| = \sup_{b \in B} \left| \hat{D}(b, \tau) / \sqrt{\hat{V}(W_n(b, \tau))/n} \right|.$$

A $(1 - \alpha)$ confidence interval (CI) for $D(\tilde{\eta}, \tau)$ is then given by

$$\hat{D}(\tilde{\eta}, \tau) \pm c_\alpha \sqrt{\hat{V}(W_n(\tilde{\eta}, \tau))/n}.$$

This interval will include 0 if the test result is not significant, and will exclude 0 otherwise.

Note that, when B is not a range but a set of several candidates $B = \{b_1, \ldots, b_L\}$, we will not have to generate the entire sample path $D_m^*(b, \tau)$ but generate $D_m^*(b, \tau)$, at $b = b_1, \ldots, b_L$, for $m = 1, \ldots, M$. We will generate $(D_m^*(b_1), \ldots, D_m^*(b_L))'$, for $m = 1, \ldots, M$,

from a multivariate normal distribution with a mean zero and variance-covariance matrix

$$\begin{pmatrix} \hat{V}(W_n(b_1,\tau)) & \cdots & \hat{\text{cov}}\{W_n(b_1,\tau), W_n(b_L,\tau)\} \\ \vdots & \ddots & \vdots \\ \hat{\text{cov}}\{W_n(b_L,\tau), W_n(b_1,\tau)\} & \cdots & \hat{V}(W_n(b_L,\tau)) \end{pmatrix}.$$

Also, note that when $B = \{b \mid b \in [0,\tau)\}$ is chosen, the proposed test involves tests based on difference in RMST, $R_1(0,\tau) - R_0(0,\tau)$, and difference in milestone survival probability $S_1(\tau) - S_0(\tau)$ on two extremes. The verification of these is trivial. For verifying the latter, let $\hat{\eta}$ be the last observed event time on $[0,\tau)$ in the cohort and consider the average survival probability on $[\hat{\eta},\tau]$ given by (1). Suppose $\hat{S}_1(\tau)$ and $\hat{S}_0(\tau)$ are estimable with the data. Then, $\frac{\hat{R}_1(\hat{\eta},\tau)}{\tau-\hat{\eta}} - \frac{\hat{R}_0(\hat{\eta},\tau)}{\tau-\hat{\eta}} = \hat{S}_1(\tau) - \hat{S}_0(\tau)$, because $\hat{S}_j(\hat{\eta}) = \hat{S}_j(\tau)$ for $j = 0, 1$.

3. Numerical Studies

We conducted extensive numerical studies to compare the performance of the adaptive LT-RMST approach with other approaches, which included four kinds of WLR tests from the $G^{\rho,\gamma}$ class (Fleming and Harrington, 1991), the combination of these WLR tests called the *Max-combo* test, the piecewise WLR test proposed by Xu et al. (2017), test based on event probability at a specific time point (or milestone survival probability), the RMST-based test, and the standard LT-RMST-based test using a fixed η.

Note that the $G^{\rho,\gamma}$ class (Fleming and Harrington, 1991) is a subclass of the class K statistics (Gill, 1980)

$$\int_0^\infty \sqrt{\frac{n_0+n_1}{n_0 n_1}} \frac{\bar{Y}_0(s)\bar{Y}_1(s)}{\bar{Y}_0(s)+\bar{Y}_1(s)} W(s) d(\hat{\Lambda}_0 - \hat{\Lambda}_1)(s), \tag{2}$$

where $\bar{Y}_j(s)$ is the number at risk at time s in group j, and the weight function $W(\cdot)$ has the following form, $W(s) = \hat{S}(s-)^\rho\{1-\hat{S}(s-)\}^\gamma$, where $\rho, \gamma \geqslant 0$ and $\hat{S}(\cdot)$ is the Kaplan-Meier estimator of the survival function of the pooled sample. In our simulations, we specifically included $G^{0,0}$ (i.e., logrank test), $G^{1,0}$ (i.e., Peto-Prentice generalized Wilcoxon test; PPW), $G^{0,1}$, and $G^{0,2}$. Because $G^{0,1}$ and $G^{0,2}$ give relatively more weights on the tail part of survival

time distribution, it is expected that these tests offer higher power than $G^{0,0}$ if a larger treatment effect appears over the course of the study time.

The delayed treatment effect has been seen in most immunotherapy trials, but it is not always the case. Also, the pattern of the difference after the separation point varies across studies. To address possible misidentification of the pattern of difference between two survival curves, one may employ a robust (or versatile) test (Fleming and Harrington, 1991), considering several WLR test statistics to capture various patterns of difference. In our simulations, we included the so-called *Max-combo* test recommended by a cross-pharma working group (Roychoudhury et al., 2021) for clinical trials with non-proportional hazards. The test statistic of the *Max-combo* test is defined as the maximum of the test statistics of $G^{0,0}$, $G^{1,0}$, $G^{0,1}$ and $G^{1,1}$.

The piecewise WLR test proposed by Xu et al. (2017) would be also an interesting alternative analytic method, particularly for immunotherapy trials. It sets $W(s) = 0$ for $s < k$ and 1 for $s \geqslant k$ in (2), where k is a pre-specified time point from which trial investigators expect that treatment effect is onset. In our simulations, we examined this test with several k's. Similarly, for the standard LT-RMST, we examined its performance with several η's. The truncation time point τ was set at 24 for the standard RMST, the standard LT-RMST, and adaptive LT-RMST.

The event time distributions we considered in our simulations are presented in Figure 2. We considered four patterns of between-group differences. Pattern 2(A) was no difference pattern. This was included to confirm the size of each test we investigated in this study. Patterns 2(B) and 2(C) were delayed difference patterns. The separation time point was $t = 4.8$ for both 2(B) and 2(C). The difference after $t = 4.8$ was PH for Pattern 2(B) and NPH for Pattern 2(C). Pattern 2(D) was a PH difference from time 0. This pattern was

included to check which tests can offer comparative power to the logrank test when the logrank test is optimal.

[Figure 2 about here.]

Regarding censoring, we considered three kinds of common censoring time distributions for both groups. Figure 3 shows the censoring distributions we considered in our simulations: 3(i) no censoring, 3(ii) light censoring, and 3(iii) moderate censoring, with an administrative censoring at 24. The proportions censored at 24^- were 0, 0.3, and 0.5 for 3(i), 3(ii), and 3(iii), respectively. Thus, a total of twelve combinations of event time distributions and censoring time distributions were considered. We considered sample sizes n=200, 300, and 500 for each group. Size and power for each test were calculated via 5000 iterations.

[Figure 3 about here.]

Table 1 shows the results regarding the empirical type I error of each test with Pattern 2(A). When $n = 200$ per arm, the size of adaptive LT-RMST with $B = [0, 24)$ was 0.06, which was slightly over the nominal level of 0.05. Except for it, we confirmed that the size of each test was sufficiently close to their nominal level of 0.05 for 5000 iterations. This result suggests that adaptive LT-RMST with $B = [0, 24)$ would require a relatively larger sample size.

[Table 1 about here.]

Tables 2 to 4 show the results regarding the power of each test with $n = 200$ per arm for the censoring distribution 3(i), 3(ii), and 3(iii), respectively. For the combination of pattern 2(B) and censoring 3(ii), the power of the logrank test was 0.837. The power of $G^{0,1}$ and $G^{0,2}$ were 0.919 and 0.868, respectively. Max-combo and the piecewise WLR tests offered 0.90 to 0.95 power. The PPW test was the worst for this pattern. The piecewise WLR test gave the highest power 0.951 when the true separation time point $t = 4.8$ was used for k. The power

of the test based on the event probability at $\tau = 24$ was 0.865. The power of the standard RMST was 0.716. This confirms that the RMST is less powerful than the logrank test for this type of pattern (Tian et al., 2018). The power of the standard LT-RMST ranged from 0.756 to 0.856. Interestingly, the power with $\eta = 8.8$ was higher than that with $\eta = 4.8$ for the standard LT-RMST, although the true separation time point in 2(B) was $t = 4.8$. The power of adaptive LT-RMST, excluding the one with $B = [0, 24)$, was from 0.818 to 0.841, which was close to the highest power of the standard LT-RMST (Table 3, Column 2(B)).

For pattern 2(C), the power of the logrank test was 0.594. While $G^{0,1}$ and $G^{0,2}$ offered even higher power than the logrank test for pattern 2(B), their power for the pattern 2(C) was quite lower than the logrank test (0.387 for $G^{0,1}$ and 0.164 for $G^{0,2}$). The power of the PPW test was 0.581 for this pattern. The power of the Max-combo test and the standard RMST were 0.679 and 0.662, respectively, which were higher than that of the logrank test. The power of the piecewise WLR tests ranged from 0.444 to 0.757 and was quite sensitive to the specified k. The power of the test based on the event probability at $\tau = 24$ was very low for this pattern (0.181). The power of the standard LT-RMST ranged from 0.710 to 0.776, which was also sensitive to the specified η but less sensitive than the piecewise WLR. The power of adaptive LT-RMST, excluding the one with $B = [0, 24)$, ranged from 0.762 to 0.765 and was close to the highest power of the standard LT-RMST (Table 3, Column 2(C)).

For pattern 2(D), the power of the logrank test was 0.490, which was higher than any other tests. The tests that gave similar power to the logrank test were the Max-combo test (0.452), the standard RMST (0.467), the standard LT-RMST (0.462 to 0.467), and the adaptive LT-RMST (0.471 to 0.481; excluding the one with $B = [0, 24)$). Of those tests we investigated, the power of the adaptive LT-RMST was closest to the power of the logrank test for 2(D). Unlike 2(B) and 2(C), the power of the standard LT-RMST was not sensitive

to the specified η for 2(D). Also, the power of adaptive LT-RMST was slightly higher than the standard LT-RMST tests (Table 3, Column 2(D)).

Overall, the standard LT-RMST and adaptive LT-RMST were robust for the patterns 2(B), 2(C), and 2(D). The power of the standard LT-RMST was sensitive to the specified η for the delayed difference patterns, 2(B) and 2(C). Adaptive LT-RMST gives us flexibility in specifying η, but the power loss for gaining the flexibility was not remarkable.

Similar findings were obtained for censoring patterns 3(i) and 3(iii) (see Tables 2 and 4), and for sample sizes 300 and 500 (data not shown).

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

4. Example

For illustration, we applied the adaptive LT-RMST method to the PFS data from the advanced renal cancer study we briefly introduced in the Introduction section (Figure 1). The study was a multinational, randomized phase III trial to compare nivolumab plus ipilimumab with sunitinib in patients with previously untreated clear-cell advanced renal-cell carcinoma. The primary analysis population was intermediate- and poor-risk patients. The PFS was one of the co-primary endpoints in this study. Data from 847 patients (425 for nivolumab plus ipilimumab and 422 for sunitinib) were served for the analyses. The primary results were reported by Motzer et al. (2018). Because the patient-level data were not publicly available, we reconstructed the PFS data of the 847 patients using the reported results and the method proposed by Guyot et al. (2012).

Figure 1 shows the estimated survival curves for both groups based on the reconstructed

data. A delayed difference pattern was observed, which suggested a violation of the PH assumption. While the p-value of the cumulative residual test (Lin et al., 1993) was not statistically significant (p=0.084), it is important to note that a non-significant p-value of the PH assumption test does not imply that the PH assumption is true (Stensrud and Hernán, 2020). The HR reported by Motzer et al. (2018) was 0.82 . However, if the PH assumption is not correct, the interpretation of HR=0.82 would be difficult because it depends on the study-specific censoring distribution (Horiguchi et al., 2019).

Table 5 shows the results of the standard RMST, the standard LT-RMST, and the adaptive LT-RMST. For this example, we set the truncation time point to τ=27 (months) for those RMST's. The RMST difference was 1.8 months (14.2 months for nivolumab plus ipilimumab group vs. 12.5 months for sunitinib group). The corresponding 0.95 CI and p-value were (0.2 to 3.3) and 0.025, respectively.

Regarding the standard LT-RMST, we chose η=5 (months). The difference in LT-RMST was 1.7 months (9.9 months for nivolumab plus ipilimumab vs. 8.2 months for sunitinib). The corresponding 0.95 CI and p-value are (0.3 to 3.2) and 0.018, respectively. It is interesting to see that the point estimate for the LT-RMST difference is almost identical to the standard RMST difference. Still, the CI was a little tighter, and the p-value was lower than the standard RMST-based analysis. This would be because the LT-RMST does not include the noise on the study time from 0 to 5 months when little signal of treatment effect appears.

Although we chose η for the LT-RMST ad hoc in this example, it would not be easy to prespecify a specific time point for η at the design stage. The proposed adaptive LT-RMST can work around this practical issue. In this example, we chose several candidates $B = \{0, 0.5, 1.0, 1.5, \ldots, 7.0\}$ (months) for η. The test result for the equality of LT-RMST on B was p=0.014. The procedure selected $\eta = 7$ (months). The difference in the LT-RMST with the range of (η, τ)=(7, 27) was 1.7 months (8.6 for nivolumab plus ipilimumab vs.

6.9 for sunitinib) and corresponding 0.95 CI was 0.3 to 3.1 months. We also calculated the RMST with the time range 0 to 7 months for reference. The estimated RMSTs were 5.6 months in nivolumab plus ipilimumab and 5.5 months in sunitinib. They were almost identical (Difference: 0.1, 0.95 CI: -0.2 to 0.3 months, p=0.725).

[Table 5 about here.]

5. Remarks

This paper extended the LT-RMST and proposed an adaptive LT-RMST approach that does not require users to specify η at the design stage. Numerical studies demonstrated that the potential power loss by adopting this flexibility was minimal. The test based on the adaptive LT-RMST is very powerful for delayed difference patterns and offers comparable power to the logrank test for PH scenarios. Also, the proposed approach can provide a robust estimate of the treatment effect summary that corresponds to the test result.

The proposed adaptive LT-RMST approach is the prespecified method. For a confirmatory study, one would specify B in the study protocol and the statistical analysis plan. Based on the empirical data, it would be reasonable to choose B to include the anticipated separation time point. An extreme one, $B = \{b \mid b \in [0, \tau)\}$, would be also an interesting choice because it includes the tests based on difference in RMST at τ and difference in milestone survival probability at τ as described in the end of the method section. However, based on our numerical studies, we would not recommend using $B = \{b \mid b \in [0, \tau)\}$ if the sample size is not so large, because the empirical type I error rate was slightly higher than the nominal level with the sample size $n = 200$ per arm. Our numerical studies suggested a larger sample size (e.g., 300 per arm or larger) would be required for employing $B = \{b \mid b \in [0, \tau)\}$. Further research is warranted to address the potential inflation of the type I error rate with $B = \{b \mid b \in [0, \tau)\}$ when the sample size is not large. For example, transformations to

improve the large sample approximation of the test statistic (Hashimoto and Kada, 2021), and also using permutations to get a robust reference distribution (Horiguchi and Uno, 2020) can be considered.

It is important to note that the η selected by the proposed procedure is not an estimate for the true separation time point of two survival curves but the one that gives the most significant difference in LT-RMST between the two groups. The procedure tends to select a η larger than the true separation time point. This was seen even in the PH difference scenario 2(D) in our numerical studies, where the adaptive LT-RMST offered higher power than the standard RMST. This phenomenon would be that the signal (i.e., area between two survival curves) around the separation time point is still tiny compared to the noise.

As Freidlin and Korn (2019) commented in their recent paper, a limitation of versatile tests, including the proposed adaptive LT-RMST, is that they can reject the null hypothesis both in favor of the treatment group and in favor of the control group. For example, consider the adaptive LT-RMST with $B = \{b \mid b \in [0, \tau)\}$, which includes the tests based on difference in RMST and difference in milestone survival probability at τ. It would not be impossible to see such a result where the RMST supports the superiority of the treatment group and the milestone survival probability supports the superiority of the control group statistically significantly. For such a cross survival situation, it would be hard to determine whether the treatment is superior to the control or not from a single test. However, for the adaptive LT-RMST, this is unlikely to happen as long as a relatively short range of B is used and the two survival curves in the early study time period are close to each other.

An R package (surv2sampleTest) to implement the proposed method is available on request from the first author (email: horiguchimiki@gmail.com).

Financial disclosure

This research was supported by McGraw/Patterson Research Fund.

Conflict of interest

The authors declare no potential conflict of interests.

Appendix : Covariance of $W_n(\eta_1, \tau)$ and $W_n(\eta_2, \tau)$

Let $\mathcal{T}_j(\tau)$ be $\sqrt{n}\int_0^\tau \{\hat{S}_j(u) - S_j(u)\}du$. From a result derived in Murray and Tsiatis (1999), it is shown

$$\mathrm{cov}\{\mathcal{T}_j(\tau_1), \mathcal{T}_j(\tau_2)\} = \int_0^{\tau_1} \left\{\int_u^{\tau_1} S_j(t)dt\right\}\left\{\int_u^{\tau_2} S_j(t)dt\right\} \frac{d\Lambda_j(u)}{p_j \pi_j(u)}, \tag{A.1}$$

for $0 < \tau_1 < \tau_2$, where $\pi_j(t) = P(X_j \geqslant t)$. Because $W_n(b, \tau) = \{\mathcal{T}_1(\tau) - \mathcal{T}_1(b)\} - \{\mathcal{T}_0(\tau) - \mathcal{T}_0(b)\}$,

$$\begin{aligned}
&\mathrm{cov}\{W_n(\eta_1, \tau), W_n(\eta_2, \tau)\} \\
&= \sum_{j=0,1} \mathrm{cov}\{\mathcal{T}_j(\tau), \mathcal{T}_j(\tau)\} - \mathrm{cov}\{\mathcal{T}_j(\tau), \mathcal{T}_j(\eta_1)\} - \mathrm{cov}\{\mathcal{T}_j(\tau), \mathcal{T}_j(\eta_2)\} + \mathrm{cov}\{\mathcal{T}_j(\eta_1), \mathcal{T}_j(\eta_2)\} \\
&= \sum_{j=0,1} \int_0^\tau \left\{A_j^\tau(u) - A_j^{\eta_1}(u) I(u < \eta_1)\right\}\left\{A_j^\tau(u) - A_j^{\eta_2}(u) I(u < \eta_2)\right\} \frac{d\Lambda_j(u)}{p_j \pi_j(u)},
\end{aligned}$$

where $A_j^x(u) = \int_u^x S_j(t)dt$.

References

Chappell, R. and Zhu, X. (2016). Describing differences in survival curves. *JAMA Oncol* **2,** 906–907.

Chen, T.-T. (2013). Statistical issues and challenges in immuno-oncology. *J Immunother Cancer* **1,** 18.

Eaton, A., Therneau, T., and Le-Rademacher, J. (2020). Designing clinical trials with (restricted) mean survival time endpoint: Practical considerations. *Clin. Trials* **17,** 285–294.

Emens, L. A., Ascierto, P. A., Darcy, P. K., Demaria, S., Eggermont, A. M. M., Redmond, W. L., Seliger, B., and Marincola, F. M. (2017). Cancer immunotherapy: Opportunities and challenges in the rapidly evolving clinical landscape. *Eur. J. Cancer* **81,** 116–129.

Fleming, T. and Harrington, D. (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons, New York.

Freidlin, B. and Korn, E. L. (2019). Methods for accommodating nonproportional hazards in clinical trials: ready for the primary analysis? *J. Clin. Oncol.* **37,** 3455–3459.

Gill, R. D. (1980). *Censoring and Stochastic Integrals*. Mathematisch Centrum, Amsterdam.

Guyot, P., Ades, A., Ouwens, M., and Welton, N. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol* **12,** 9.

Hasegawa, T., Misawa, S., Nakagawa, S., Tanaka, S., Tanase, T., Ugai, H., Wakana, A., Yodo, Y., Tsuchiya, S., Suganami, H., and JPMA Task Force Members (2020). Restricted mean survival time as a summary measure of time-to-event outcome. *Pharm. Stat.* **19,** 436–453.

Hashimoto, H. and Kada, A. (2021). A note on confidence intervals for the restricted mean survival time based on transformations in small sample size. *Pharm. Stat.* **.,** Online ahead of print.

Horiguchi, M., Cronin, A. M., Takeuchi, M., and Uno, H. (2018). 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* **37,** 2307–2320.

Horiguchi, M., Hassett, M. J., and Uno, H. (2019). How do the accrual pattern and follow-up duration affect the hazard ratio estimate when the proportional hazards assumption is violated? *Oncologist* **24,** 867–871.

Horiguchi, M., Tian, L., Uno, H., Cheng, S., Kim, D. H., Schrag, D., and Wei, L.-J. (2018). Quantification of long-term survival benefit in a comparative oncology clinical study. *JAMA Oncol* **4,** 881–882.

Horiguchi, M. and Uno, H. (2020). On permutation tests for comparing restricted mean survival time with small sample from randomized trials. *Stat. Med.* **39,** 2655–2670.

Huang, B. (2018). Some statistical considerations in the clinical development of cancer immunotherapies. *Pharm Stat* **17,** 49–60.

Lin, D., Wei, L., and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80,** 557–572.

Motzer, R. J., Tannir, N. M., McDermott, D. F., Arén Frontera, O., Melichar, B., Choueiri, T. K., and et al. (2018). Nivolumab plus ipilimumab versus sunitinib in advanced renal-cell carcinoma. *N Engl J Med* **378,** 1277–1290.

Murray, S. and Tsiatis, A. A. (1999). Sequential methods for comparing years of life saved in the two-sample censored data problem. *Biometrics* **55,** 1085–1092.

Paukner, M. and Chappell, R. (2021). Window mean survival time. *Stat Med* **40,** 5521–5533.

Roychoudhury, S., Anderson, K. M., Ye, J., and Mukhopadhyay, P. (2021). Robust design and analysis of clinical trials with nonproportional hazards: a straw man guidance from a cross-pharma working group. *Stat. Biopharm. Res.* **.,** Online ahead of print.

Royston, P. and Parmar, M. K. B. (2011). 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.* **30,** 2409–2421.

Royston, P. and Parmar, M. K. B. (2013). 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.* **13,** 152.

Saad, E. D., Zalcberg, J. R., Péron, J., Coart, E., Burzykowski, T., and Buyse, M. (2017). Understanding and communicating measures of treatment effect on survival: can we do better? *J. Natl. Cancer Inst.* **110,** 232–240.

Stensrud, M. J. and Hernán, M. A. (2020). Why test for proportional hazards? *JAMA* **323,** 1401–1402.

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., Jin, H., Uno, H., Lu, Y., Huang, B., Anderson, K. M., and Wei, L. J. (2020). On the empirical choice of the time window for restricted mean survival time. *Biometrics* **76,** 1157–1166.

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. *J Clin Oncol* **32,** 2380–2385.

Uno, H., Horiguchi, M., and Hassett, M. J. (2020). Statistical test/estimation methods used in contemporary phase III cancer randomized controlled trials with time-to-event outcomes. *Oncologist* **25,** 91–93.

Uno, H., Wittes, J., Fu, H., Solomon, S., 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. *Ann Intern Med* **163,** 127–134.

Vivot, A., Créquit, P., and Porcher, R. (2019). Use of late-life expectancy for assessing the long-term benefit of immune checkpoint inhibitors. *J Natl Cancer Inst* **111,** 519–521.

Xu, Z., Zhen, B., Park, Y., and Zhu, B. (2017). Designing therapeutic cancer vaccine trials with delayed treatment effect. *Stat Med* **36,** 592–605.

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

Zhao, L., Tian, L., Uno, H., Solomon, S. D., Pfeffer, M. A., Schindler, J. S., 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. *Clin Trials* **9,** 570–577.

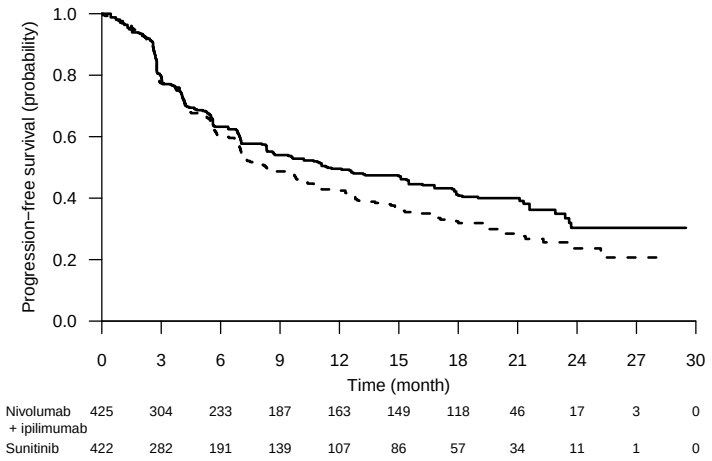

Figure 1. Kaplan-Meier curves of the progression-free survival based on reconstructed data from the advanced renal cancer study. Solid line, nivolumab plus ipilimumab group; dashed line, sunitinib group.

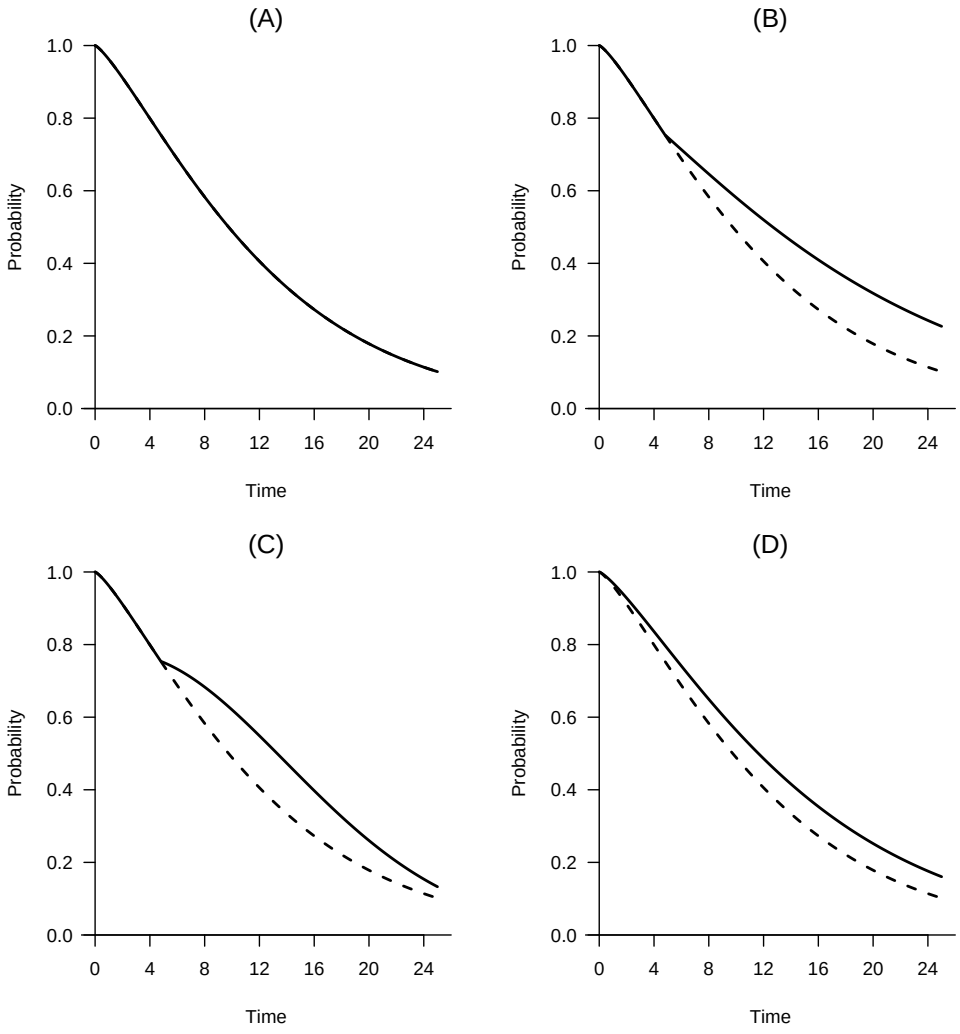

Figure 2. Survival functions considered in the numerical studies.

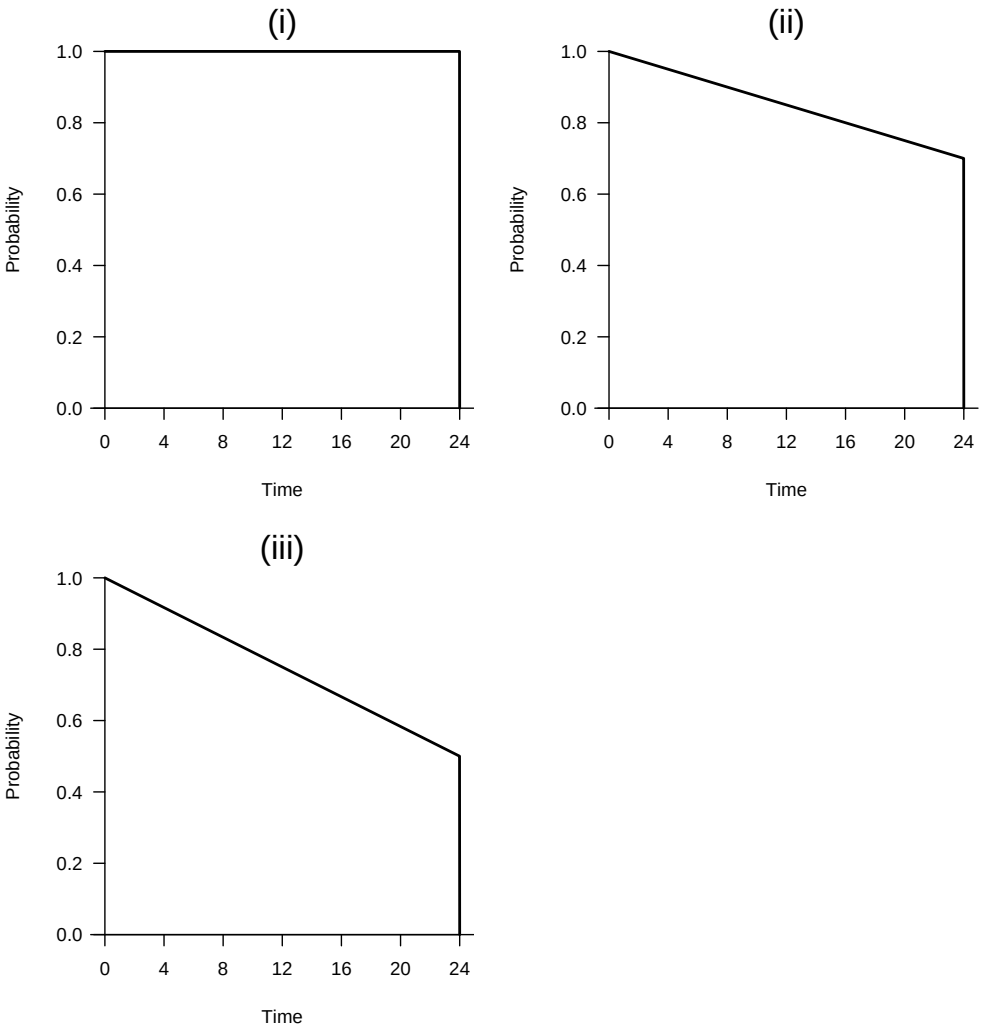

Figure 3. Survival functions of the underlying censoring distributions considered in the numerical studies.

Table 1

Size of tests evaluated in pattern 2(A) shown in Figure 2.

Sample size	**n=200/arm**			**n=300/arm**			**n=500/arm**		
Censoring distribution*	**3(i)**	**3(ii)**	**3(iii)**	**3(i)**	**3(ii)**	**3(iii)**	**3(i)**	**3(ii)**	**3(iii)**
Test									
$G^{\rho,\gamma}$ class of WLR									
$\{\rho,\gamma\}=\{0,0\}$ [logrank]	0.054	0.049	0.048	0.055	0.048	0.049	0.048	0.048	0.046
$\{\rho,\gamma\}=\{1,0\}$ [PPW]	0.053	0.049	0.050	0.055	0.048	0.050	0.048	0.048	0.047
$\{\rho,\gamma\}=\{0,1\}$	0.056	0.051	0.048	0.050	0.048	0.050	0.048	0.049	0.049
$\{\rho,\gamma\}=\{0,2\}$	0.053	0.051	0.054	0.051	0.048	0.052	0.049	0.051	0.049
Max-combo of $G^{\rho,\gamma}$ class tests**	0.056	0.051	0.048	0.054	0.048	0.051	0.051	0.047	0.050
Piecewise WLR (fixed k)									
$k=3.8$	0.055	0.053	0.050	0.049	0.049	0.049	0.049	0.048	0.047
$k=4.8$	0.054	0.051	0.049	0.051	0.050	0.049	0.050	0.050	0.045
$k=5.8$	0.050	0.051	0.048	0.052	0.047	0.047	0.048	0.045	0.047
$k=6.8$	0.053	0.054	0.051	0.050	0.049	0.047	0.047	0.044	0.048
Event probability at τ ***	0.051	0.053	0.048	0.050	0.052	0.051	0.049	0.054	0.051
Standard RMST $[0,\tau]$ ***	0.053	0.049	0.049	0.052	0.046	0.049	0.048	0.047	0.044
Standard LT-RMST (fixed range of $[\eta,\tau]$)***									
$\eta=3.8$	0.053	0.047	0.050	0.052	0.046	0.048	0.046	0.045	0.043
$\eta=4.8$	0.052	0.047	0.050	0.051	0.047	0.048	0.047	0.046	0.044
$\eta=5.8$	0.054	0.048	0.050	0.050	0.046	0.046	0.049	0.045	0.045
$\eta=6.8$	0.054	0.048	0.048	0.049	0.045	0.046	0.049	0.045	0.045
$\eta=7.8$	0.053	0.047	0.048	0.050	0.046	0.045	0.050	0.045	0.043
$\eta=8.8$	0.054	0.049	0.048	0.049	0.045	0.047	0.049	0.044	0.043
Adaptive LT-RMST (adaptive range of $[\eta,\tau]$)***									
$\eta\in B=\{3.0,3.1,\ldots,7.0\}$	0.056	0.048	0.052	0.053	0.048	0.049	0.049	0.046	0.046
$\eta\in B=\{0,0.1,\ldots,9.0\}$	0.058	0.051	0.053	0.055	0.049	0.052	0.052	0.049	0.047
$\eta\in B=[0,7]$	0.057	0.050	0.052	0.056	0.050	0.051	0.051	0.047	0.048
$\eta\in B=[0,9]$	0.058	0.051	0.053	0.055	0.049	0.051	0.051	0.048	0.046
$\eta\in B=[0,24)$	0.060	0.060	0.061	0.054	0.052	0.057	0.051	0.054	0.054

*: 3(i), no censoring; 3(ii), light censoring; 3(iii), moderate censoring (Figure 3).
**: Max-combo of $\{\rho,\gamma\}=\{0,0\},\{1,0\},\{0,1\}$, and $\{1,1\}$.
***: $\tau=24$.
Abbreviation: WLR, weighted logrank; PPW, Peto-Prentice generalized Wilcoxon; RMST, restricted mean survival time; LT-RMST, long-term restricted mean survival time.

Table 2

Power of tests evaluated with sample size 200 per arm in no censoring pattern shown in Figure 3(i).

Test	**Survival distribution*** **2(B)**	**2(C)**	**2(D)**
$G^{\rho,\gamma}$ class of WLR			
$\{\rho,\gamma\}=\{0,0\}$ [logrank]	0.896	0.609	0.539
$\{\rho,\gamma\}=\{1,0\}$ [PPW]	0.659	0.640	0.470
$\{\rho,\gamma\}=\{0,1\}$	0.951	0.376	0.426
$\{\rho,\gamma\}=\{0,2\}$	0.915	0.158	0.331
Max-combo of $G^{\rho,\gamma}$ class tests**	0.941	0.709	0.505
Piecewise WLR (fixed k)			
$k=3.8$	0.954	0.724	0.447
$k=4.8$	0.971	0.756	0.416
$k=5.8$	0.959	0.602	0.388
$k=6.8$	0.948	0.447	0.363
Event probability at τ ***	0.924	0.224	0.420
Standard RMST*** $[0,\tau]$	0.747	0.699	0.500
Standard LT-RMST*** (fixed range of $[\eta,\tau]$)			
$\eta=3.8$	0.787	0.742	0.509
$\eta=4.8$	0.815	0.769	0.508
$\eta=5.8$	0.839	0.798	0.511
$\eta=6.8$	0.856	0.811	0.512
$\eta=7.8$	0.872	0.814	0.515
$\eta=8.8$	0.886	0.810	0.513
Adaptive LT-RMST*** (adaptive range of $[\eta,\tau]$)			
$\eta\in B=\{3.0,3.1,3.2,\ldots,7.0\}$	0.850	0.801	0.516
$\eta\in B=\{0,0.1,0.2,\ldots,9.0\}$	0.875	0.801	0.527
$\eta\in B=[0,7]$	0.848	0.798	0.521
$\eta\in B=[0,9]$	0.874	0.801	0.526
$\eta\in B=[0,24)$	0.937	0.725	0.519

*: 2(B), PH difference after the separation time point $t=4.8$; 2(C), NPH difference after the separation time point $t=4.8$; 2(D), PH difference (Figure 2).

**: Max-combo of $\{\rho,\gamma\}=\{0,0\},\{1,0\},\{0,1\}$, and $\{1,1\}$.

***: $\tau=24$.

Abbreviation: WLR, weighted logrank; PPW, Peto-Prentice generalized Wilcoxon; RMST, restricted mean survival time; LT-RMST, long-term restricted mean survival time; PH, proportional hazards; NPH, non-proportional hazards.

Table 3

Power of tests evaluated with sample size 200 per arm in light censoring pattern shown in Figure 3(ii).

Test	**Survival distribution***		
	2(B)	**2(C)**	**2(D)**
$G^{\rho,\gamma}$ class of WLR			
$\{\rho,\gamma\}=\{0,0\}$ [logrank]	0.837	0.594	0.490
$\{\rho,\gamma\}=\{1,0\}$ [PPW]	0.559	0.581	0.427
$\{\rho,\gamma\}=\{0,1\}$	0.919	0.387	0.378
$\{\rho,\gamma\}=\{0,2\}$	0.868	0.164	0.293
Max-combo of $G^{\rho,\gamma}$ class tests**	0.900	0.679	0.452
Piecewise WLR (fixed k)			
$k=3.8$	0.918	0.715	0.395
$k=4.8$	0.951	0.757	0.364
$k=5.8$	0.929	0.602	0.331
$k=6.8$	0.905	0.444	0.312
Event probability at τ ***	0.865	0.181	0.351
Standard RMST*** $[0,\tau]$	0.716	0.662	0.467
Standard LT-RMST*** (fixed range of $[\eta,\tau]$)			
$\eta=3.8$	0.756	0.710	0.467
$\eta=4.8$	0.782	0.739	0.467
$\eta=5.8$	0.806	0.763	0.464
$\eta=6.8$	0.827	0.775	0.465
$\eta=7.8$	0.842	0.776	0.464
$\eta=8.8$	0.856	0.770	0.462
Adaptive LT-RMST*** (adaptive range of $[\eta,\tau]$)			
$\eta\in B=\{3.0,3.1,3.2,\ldots,7.0\}$	0.820	0.764	0.471
$\eta\in B=\{0,0.1,0.2,\ldots,9.0\}$	0.841	0.765	0.480
$\eta\in B=[0,7]$	0.818	0.762	0.477
$\eta\in B=[0,9]$	0.840	0.765	0.481
$\eta\in B=[0,24)$	0.897	0.674	0.473

*: 2(B), PH difference after the separation time point $t=4.8$; 2(C), NPH difference after the separation time point $t=4.8$; 2(D), PH difference (Figure 2).

**: Max-combo of $\{\rho,\gamma\}=\{0,0\},\{1,0\},\{0,1\}$, and $\{1,1\}$.

***: $\tau=24$.

Abbreviation: WLR, weighted logrank; PPW, Peto-Prentice generalized Wilcoxon; RMST, restricted mean survival time; LT-RMST, long-term restricted mean survival time; PH, proportional hazards; NPH, non-proportional hazards.

Table 4

Power of tests evaluated with sample size 200 per arm in moderate censoring pattern shown in Figure 3(iii).

Test	**Survival distribution***		
	2(B)	**2(C)**	**2(D)**
$G^{\rho,\gamma}$ class of WLR			
$\{\rho,\gamma\}=\{0,0\}$ [logrank]	0.762	0.562	0.441
$\{\rho,\gamma\}=\{1,0\}$ [PPW]	0.487	0.536	0.389
$\{\rho,\gamma\}=\{0,1\}$	0.882	0.401	0.338
$\{\rho,\gamma\}=\{0,2\}$	0.819	0.175	0.263
Max-combo of $G^{\rho,\gamma}$ class tests**	0.858	0.653	0.410
Piecewise WLR (fixed k)			
$k=3.8$	0.885	0.708	0.349
$k=4.8$	0.916	0.756	0.318
$k=5.8$	0.893	0.593	0.293
$k=6.8$	0.857	0.437	0.270
Event probability at τ ***	0.791	0.155	0.291
Standard RMST*** $[0,\tau]$	0.684	0.634	0.427
Standard LT-RMST*** (fixed range of $[\eta,\tau]$)			
$\eta=3.8$	0.725	0.680	0.431
$\eta=4.8$	0.750	0.708	0.428
$\eta=5.8$	0.775	0.731	0.424
$\eta=6.8$	0.793	0.743	0.422
$\eta=7.8$	0.808	0.744	0.423
$\eta=8.8$	0.821	0.734	0.421
Adaptive LT-RMST*** (adaptive range of $[\eta,\tau]$)			
$\eta\in B=\{3.0,3.1,3.2,\ldots,7.0\}$	0.784	0.731	0.433
$\eta\in B=\{0,0.1,0.2,\ldots,9.0\}$	0.808	0.731	0.442
$\eta\in B=[0,7]$	0.782	0.731	0.439
$\eta\in B=[0,9]$	0.807	0.732	0.442
$\eta\in B=[0,24)$	0.857	0.630	0.426

*: 2(B), PH difference after the separation time point $t=4.8$; 2(C), NPH difference after the separation time point $t=4.8$; 2(D), PH difference (Figure 2).

**: Max-combo of $\{\rho,\gamma\}=\{0,0\},\{1,0\},\{0,1\}$, and $\{1,1\}$.

***: $\tau=24$.

Abbreviation: WLR, weighted logrank; PPW, Peto-Prentice generalized Wilcoxon; RMST, restricted mean survival time; LT-RMST, long-term restricted mean survival time; PH, proportional hazards; NPH, non-proportional hazards.

Table 5
Analysis results with the advanced renal cancer data

		Treatment	**Control**		**Difference**	
Method	Range	Est.	Est.	Est.	0.95 CI	P
RMST (month)	0 to 27	14.2	12.5	1.8	0.2 to 3.3	0.025
Standard LT-RMST (month)	5 to 27	9.9	8.2	1.7	0.3 to 3.2	0.018
Adaptive LT-RMST (month)	7^* to 27	8.6	6.9	1.7	0.3 to 3.1	0.014

*: 7 (months) was selected among $B = \{0, 0.5, 1.0, \ldots, 7.0\}$.
Abbreviation: RMST, restricted mean survival time; LT-RMST, long-term restricted mean survival time; Est., estimate; CI, confidence interval; P, p-value.