Ratio and Difference of Average Hazard with Survival Weight: New Measures to Quantify Survival Benefit of New Therapy

Hajime Uno[1,2*], Miki Horiguchi[1,2]

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

Boston, Massachusetts 02215, U.S.A.

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

Boston, Massachusetts 02215, U.S.A.

**email:* huno@ds.dfci.harvard.edu

SUMMARY: The hazard ratio (HR) has been the most popular measure to quantify the magnitude of treatment effect on time-to-event outcomes in clinical research. However, the HR estimated by Cox's method has several drawbacks. One major issue is that there is no clear interpretation when the proportional hazards (PH) assumption does not hold, because it is affected by study-specific censoring time distribution in non-PH cases. Another major issue is that the lack of a group-specific absolute hazard value in each group obscures the clinical significance of the magnitude of the treatment effect. Given these, we propose average hazard with survival weight (AH-SW) as a summary metric of event time distribution and will use difference in AH-SW (DAH-SW) or ratio of AH-SW (RAH-SW) to quantify the treatment effect magnitude. The AH-SW we propose is a new digestible metric interpreted as a person-years event rate when random censoring would not exist. It is defined as the ratio of τ-year event rate and restricted mean survival time, which can be estimated non-parametrically. Numerical studies demonstrate that DAH-SW and RAH-SW offer almost identical power to Cox's method under PH scenarios and can be more powerful for delayed-difference patterns that are often seen in immunotherapy trials. The proposed metrics (i.e., AH-SW, DAH-SW and RAH-SW) and the inferential methods for them offer a digestible interpretation that the conventional Cox's method could not provide about the survival benefit of a new therapy. These metrics will increase the likelihood that results from clinical studies are correctly interpreted.

KEY WORDS: Hazard ratio; immunotherapy studies; non-proportional hazards; person-years event rate; t-year survival rate; restricted mean survival time.

1. Introduction

The magnitude of treatment effect on time-to-event outcomes, such as overall survival (OS) and progression-free survival (PFS), is almost routinely explained using the hazard ratio (HR). A recent study showed that more than 95% studies used the HR in contemporary phase III randomized trials in oncology (Uno et al., 2020). The standard method to estimate HR is based on the partial likelihood score equation proposed by Cox (1972, 1975). However, the HR based on Cox's method has several drawbacks (Kalbfleisch and Prentice, 1981; Struthers and Kalbfleisch, 1986; Lin and Wei, 1989; Hernán, 2010; Uno et al., 2014). One notable issue is that when the proportional hazards (PH) assumption does not hold, the result depends on the underlying study-specific censoring time distribution (Kalbfleisch and Prentice, 1981; Schemper, 1992; Xu and O'Quigley, 2000; Horiguchi et al., 2019). In these non-PH cases, interpretation of the HR is not clear.

Another major issue is the lack of a reference number from the control group. This reference number is a summary metric of event time distribution in the control group that can serve as a reference for the corresponding between-group contrast measure to quantify the treatment effect magnitude. For example, when the difference or ratio of response rates from two groups is used for reporting the treatment effect on tumor response in a cancer trial, the reference number will be the response rate in the control group. The lack of a reference number for Cox's HR is not a statistical problem but rather a practical one that arises during clinician/patient treatment decision making. This can be seen in an example clinical study that uses the risk ratio (RR) to report the treatment effect magnitude. The risk reduction from 50% (control) to 40% (treatment) and that from 1% (control) to 0.8% (treatment) give the exact same RR=0.8 (i.e., 20% of risk reduction). However, the clinical implication would be quite different, depending on the absolute risk in the control group. As such, a reference number from the control group plays an important role in determining whether the observed

between-group summary measure indicates a clinically meaningful treatment benefit/risk. Unfortunately, we do not have such a reference number for Cox's HR. Cox's model has a baseline hazard function; however, this is not a number but a function of time.

These points regarding the issues of HR have been reported in the General Statistical Guidance provided by the Annals of Internal Medicine. It states, *"... Hazard ratios are notoriously difficult to interpret clinically, may be sensitive to the length of follow-up, and rely on model assumptions, such as proportional hazards. In addition, presenting estimates of effect in both absolute and relative terms increases the likelihood that results will be correctly interpreted..."* (Annals of Internal Medicine). However, with a reference number it would be possible to present the treatment effect in both absolute and relative terms.

To address the first issue of Cox's HR, several alternative approaches have been already proposed (Kalbfleisch and Prentice, 1981; Schemper, 1992; Xu and O'Quigley, 2000). Kalbfleisch and Prentice (1981) proposed the "average" hazard ratio (AHR). Borrowing the expression by Schemper et al. (2009), a general form of the AHR is given by

$$\text{AHR} = \frac{\int \left[h_1(t)/\{h_0(t)+h_1(t)\}\right] w(t)\{f_0(t)+f_1(t)\}dt}{\int \left[h_0(t)/\{h_0(t)+h_1(t)\}\right] w(t)\{f_0(t)+f_1(t)\}dt}, \tag{1}$$

where $h_k(\cdot)$ and $f_k(\cdot)$ are the hazard function and the density function of the event time T_k for group k, respectively, for $k = 0, 1$, and $w(\cdot)$ is a weight function. When $w(t) = 1$, AHR is a HR of the standard Cox (Schemper et al., 2009). Let $S_k(t)$ denote the survival function for group k. When $w(t) = \{S_0(t)f_1(t) + S_1(t)f_0(t)\}/\{f_0(t) + f_1(t)\}$, the AHR is simplified to

$$\frac{\int h_1(t)S_0(t)S_1(t)dt}{\int h_0(t)S_0(t)S_1(t)dt} = \frac{P(T_1 < T_0)}{1 - P(T_1 < T_0)}, \tag{2}$$

which is called the odds-of-concordance (Schemper et al., 2009).

Another approach to address the first issue of Cox's HR is using weighted Cox regression (Schemper, 1992; Xu and O'Quigley, 2000). The problem with Cox's method is that the limiting quantity of the partial likelihood score equation involves censoring time distribution when the PH assumption does not hold. Xu and O'Quigley (2000) proposed using inverse

probability censoring weights, so that the censoring time distribution can be removed from the limiting quantity of the estimating equation without imposing the PH assumption. The resulting HR estimate from this estimating equation can be interpretable as an AHR.

Unlike Cox's method, the estimates for the AHR derived from these alternative methods will not depend on a study-specific random censoring time distribution regardless of whether the PH assumption holds or not. Therefore, the interpretation of the resulting AHR estimates would still be possible even under non-PH scenarios. However, although these AHR approaches address the first issue of Cox's HR, it is not clear if they address the second issue of Cox's HR. One may try to calculate the average hazard in the control group, as a reference number for AHR, by standardizing the term in the denominator of the equation (1) or (2). However, it may be difficult to interpret this number as an absolute hazard of the control group because it involves the event time distribution of the treatment group.

Given the limitations on Cox's HR and these AHR approaches, non-hazard-based alternative metrics to summarize treatment effect magnitude are gaining attention (Royston and Parmar, 2011; Uno et al., 2014, 2015; A'Hern, 2016; Chappell and Zhu, 2016; Péron et al., 2016; Saad et al., 2018). For example, the difference or ratio of restricted mean survival time (RMST) is a good alternative measure that does not have the first or second issue we discussed above. However, availability of such non-hazard-based alternative measures will not resolve all the problems in practice. Since the summary measure should be selected to address clinical research questions that are highly variable, there will still exist many situations where investigators prefer to use a summary metric based on "hazard."

These provide motivation for developing novel summary measures for quantifying the treatment effect based on average hazards (AH) from two groups. We propose a ratio of average hazard (RAH) and difference in average hazard (DAH) that can be consistently estimated and have interpretations regardless of whether the PH assumption holds or not.

In the proposed method, a reference number from the control group is also available. We provide the details of the proposed measures and inference procedures for them (Section 2). We conduct numerical studies to assess the performance of the proposed method in finite sample size situations (Section 3). We also compare performance of the proposed method with Cox's method in the real-world setting using empirical data from recently conducted phase III cancer trials (Section 4). We illustrate how the proposed method can help clinical investigators with the interpretation of treatment effect by using the data from a recently conducted immunotherapy trial (Section 5), followed by some remarks (Section 6).

2. Method

2.1 *Average Hazard with Survival Weight*

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

Let $h_k(\cdot)$ be the hazard function for T_k. For a given weight function, $w_k(\cdot)$, the average hazard over a given time range $[0, \tau]$ is defined by

$$\eta_k(\tau) = \frac{\int_0^\tau h_k(u)w_k(u)du}{\int_0^\tau w_k(u)du}.$$

When $w_k(t) = 1$ for $t \in [0, \tau]$, $\eta_k(\tau) = H_k(\tau)/\tau$, where $H_k(\cdot)$ is the cumulative hazard function of T_k. It can be also expressed as $\eta_k(\tau) = -\log\{S_k(\tau)\}/\tau$, where $S_k(\cdot)$ is the survival function for T_k. Let us call this the average hazard with equal weight (AH-EW). Difference or ratio of the AH-EW can be a between-group summary measure to quantify the treatment effect measure. However, we pursue neither in this paper, because it is essentially the same as the

difference or ratio of the cumulative hazard function at τ. The inference procedures of these quantities were already investigated extensively (Fleming and Harrington, 1991).

Here, we propose the average hazard with survival weight (AH-SW) using $S_k(t)$ for the weight function. When $w_k(t) = S_k(t)$, the numerator of $\eta_k(\tau)$ is denoted by the event rate at τ, $F_k(\tau) = 1-S_k(\tau)$, and the denominator is denoted by the RMST at τ, $R_k(\tau) = \int_0^\tau S_k(u)du$. That is, the AH-SW is written by

$$\eta_k(\tau) = \frac{F_k(\tau)}{R_k(\tau)}.$$

It is interesting to rewrite this by $\eta_k(\tau) = \int_0^\tau F_k'(t)dt / \int_0^\tau S_k(t)dt$, and compare it with the definition of the hazard function $h_k(t) = F_k'(t)/S_k(t)$, where $F_k'(t) = \frac{d}{dt}F(t)$. We notice that the AH-SW has a somehow similar form to the hazard function.

The rationale for proposing $w_k(t) = S_k(t)$ as the weight are as follows. First, $S_k(t)$ is a decreasing function of time and independent of study-specific censoring time distribution. Because the number of subjects at risk is decreasing as t increases, the precision in estimating $h_k(t)$ will also be decreasing along with t. From this point of view, one may consider $\Pr(X_k \geqslant t)$ as a weight function in order to take the size of risk set into account for calculating the AH. However, the interpretation of the AH with $w(t) = \Pr(X_k \geqslant t) = \Pr(\min(T_k, C_k) \geqslant t)$ will be rather challenging, because this weight involves the study-specific censoring distribution, C_k.

Second, for a pair of event time random variables, T_0 and T_1, the AH-SW does not contradict with their stochastic order. For a given τ, suppose T_0 is less than T_1 in the usual stochastic order, that is, $\Pr(T_0 > t) \leqslant \Pr(T_1 > t)$ for all $t \in (0, \tau)$. This implies that $F_0(\tau) \geqslant F_1(\tau)$ and $R_0^{-1}(\tau) \geqslant R_1^{-1}(\tau)$. Therefore, $\eta_0(\tau) \geqslant \eta_1(\tau)$. This would an ideal characteristic of the AH-SW because it will not produce a counterintuitive result when the survival function from one group is uniformly higher than that from another group. The

AH-EW also has this characteristic, but AH with some other weight functions may not hold this.

Third, with $w_k(t) = S_k(t)$, the AH becomes a metric that would significantly help interpretation of the analytical results of time-to-event outcomes. Specifically, the AH-SW can be interpreted as *the person-years event rate when general random censoring would not exist.* To show this, consider the conventional method of calculating the person-years event rate, that is,

$$\hat{\lambda}_k(\tau) = \frac{\sum_{i=1}^{n_k} \Delta_{ki} I(X_{ki} \leqslant \tau)}{\sum_{i=1}^{n_k} (X_{ki} \wedge \tau)},$$

where τ it typically the maximum event time or censoring time in the observed data. In that case, $\hat{\lambda}_k(\tau)$ will be simply denoted by $\sum_{i=1}^{n_k} \Delta_{ki} / \sum_{i=1}^{n_k} X_{ki}$. It is well known that, when the distribution of T_k follows an exponential distribution with a parameter λ_k, $\hat{\lambda}_k(\tau)$ is the maximum likelihood estimator for λ_k and it is consistent. However, when this distribution assumption is not correct, $\hat{\lambda}_k(\tau)$ will converge to a quantity that involves a study-specific censoring time distribution. In fact, when the distribution assumption is not correct, $\hat{\lambda}_k(\tau)$ coverges in probability to

$$\lambda_k^*(\tau) = \frac{E\left\{I(T_k < C_k \wedge \tau)\right\}}{E\left\{(T_k \wedge C_k) \wedge \tau\right\}} = \frac{E\left\{F_k(C_k \wedge \tau) \mid C_k\right\}}{\int_0^\tau S_k(u) S_{C_k}(u) du}, \tag{3}$$

where $S_{C_k}(\cdot)$ is the survival function for C_k. Now, suppose there is no random censoring in the sense that $\Pr(C_k \geqslant \tau) = 1$. In this case, $F_k(C_k \wedge \tau) = F_k(\tau)$ and $S_{C_k}(u) = 1$ for $u \in (0, \tau)$ in the equation (3). Therefore, $\lambda_k^*(\tau) = F_k(\tau)/R_k(\tau)$, which is identical to the AH-SW. This derivation demonstrates that the AH-SW is a summary metric of T_k that has a clear interpretation — the average person-years event rate of T_k on $t \in (0, \tau)$ when all T_k before τ would have been observed without being censored by study-specific censoring time C_k.

A natural non-parametric estimator for the AH-SW is $\hat{\eta}_k(\tau) = \hat{F}_k(\tau)/\hat{R}_k(\tau)$, where $\hat{R}_k(\tau) = \int_0^\tau \left\{1 - \hat{F}_k(u)\right\} du$, and $\hat{F}_k(\cdot)$ is the Kaplan-Meier estimator for $F_k(\cdot)$. For the inference of

the AH-SW for group k, we consider $Q_k = n_k^{1/2}\left\{\log\hat{\eta}_k(\tau) - \log\eta_k(\tau)\right\}$. In Appendix A, we show that Q_k converges weakly to a normal distribution with mean zero and variance

$$V(Q_k) = \int_0^\tau \left\{\frac{1}{F_k(\tau)} - \frac{R_k(u)}{R_k(\tau)}\right\}^2 \frac{dH_k(u)}{G_k(u)},$$

where $G_k(t) = \Pr(X_k \geqslant t)$. This variance can be estimated by replacing the unknown quantities by their empirical counterparts. That is,

$$\hat{V}(Q_k) = \int_0^\tau \left\{\frac{1}{\hat{F}_k(\tau)} - \frac{\hat{R}_k(u)}{\hat{R}_k(\tau)}\right\}^2 \frac{d\hat{H}_k(u)}{\hat{G}_k(u)},$$

where $\hat{G}_k(t) = n_k^{-1}\sum_{i=1}^{n_k} I(X_{ki} \geqslant t)$, and $\hat{H}_k(\cdot)$ is the Nelson-Aalen estimator for the cumulative hazard function for group k.

Using these results, an $(1-\alpha)$ asymptotic confidence interval (CI) for the AH-SW in group k is

$$\exp\left\{\log\hat{\eta}_k(\tau) \pm z_{1-\alpha/2}\sqrt{\hat{V}(Q_k)/n_k}\right\},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)\times 100$-percentile of the standard normal distribution.

2.2 *Ratio of Average Hazard with Survival Weight*

The ratio of average hazard with survival weight (RAH-SW) is defined by

$$\theta(\tau) = \frac{\eta_1(\tau)}{\eta_0(\tau)} = \frac{F_1(\tau)}{F_0(\tau)}\frac{R_0(\tau)}{R_1(\tau)}. \tag{4}$$

Interestingly, the RAH-SW is expressed as a product of the ratio of τ-year event rate (treatment ($k=1$) versus control ($k=0$)) and the ratio of RMST (control ($k=0$) versus treatment ($k=1$)) at τ. The equation (4) gives another insight about the RAH-SW. Suppose $\lim_{t\to\infty} F_k(t) = 1$, for $k = 0, 1$. This implies that $\lim_{t\to\infty} R_k(t) = E(T_k)$, for $k = 0, 1$. Therefore, for a large τ such that $F_k(\tau)$ is close to 1 for $k = 0, 1$, $\theta(\tau)$ can be viewed as an approximation of the ratio of mean survival times from two groups (control versus treatment), similar to the ratio of RMST.

The estimator for (4) is $\hat{\theta}(\tau) = \hat{\eta}_1(\tau)/\hat{\eta}_0(\tau)$. For hypothesis testing and interval estimation,

we consider the asymptotic distribution of

$$n^{1/2}\left\{\log\hat{\theta}(\tau)-\log\theta(\tau)\right\}=n^{1/2}\left\{\log\hat{\eta}_1(\tau)-\log\eta_1(\tau)\right\}-n^{1/2}\left\{\log\hat{\eta}_0(\tau)-\log\eta_0(\tau)\right\}.$$

From the results regarding Q_k described in Section 2.1, it is obvious that $n^{1/2}\left\{\log\hat{\theta}(\tau)-\log\theta(\tau)\right\}$ converges weakly to a normal distribution with mean zero and variance $p_1^{-1}V(Q_1)+p_0^{-1}V(Q_0)$. The variance can be estimated by $n\left\{n_1^{-1}\hat{V}(Q_1)+n_0^{-1}\hat{V}(Q_0)\right\}$. Therefore, an $(1-\alpha)$ asymptotic CI for $\theta(\tau)$ is

$$\exp\left\{\log\hat{\theta}(\tau)\pm z_{1-\alpha/2}\sqrt{n_1^{-1}\hat{V}(Q_1)+n_0^{-1}\hat{V}(Q_0)}\right\}. \tag{5}$$

For testing the null hypothesis $\log\theta(\tau)=0$,

$$\log\hat{\theta}(\tau)/\sqrt{n_1^{-1}\hat{V}(Q_1)+n_0^{-1}\hat{V}(Q_0)} \tag{6}$$

is used as the test statistic, which asymptotically follows the standard normal distribution under the null hypothesis.

2.3 *Difference in Average Hazard with Survival Weight*

We consider the difference in average hazard with survival weight (DAH-SW). Using the same notations as we used for the RAH-SW, the DAH-SW is given by

$$\xi(\tau)=\eta_1(\tau)-\eta_0(\tau)=\frac{F_1(\tau)}{R_1(\tau)}-\frac{F_0(\tau)}{R_0(\tau)}.$$

Similar to the RAH-SW, it is trivial that this can be consistently estimated by

$$\hat{\xi}(\tau)=\frac{\hat{F}_1(\tau)}{\hat{R}_1(\tau)}-\frac{\hat{F}_0(\tau)}{\hat{R}_0(\tau)}.$$

Now, we consider the asymptotic distribution of

$$n^{1/2}\left\{\hat{\xi}(\tau)-\xi(\tau)\right\}=n^{1/2}\left\{\hat{\eta}_1(\tau)-\eta_1(\tau)\right\}-n^{1/2}\left\{\hat{\eta}_0(\tau)-\eta_0(\tau)\right\}.$$

In Appendix B, we show that $U_k=n_k^{1/2}\left\{\hat{\eta}_k(\tau)-\eta_k(\tau)\right\}$ converges weakly to a normal distribution with mean zero and variance

$$V(U_k)=\int_0^\tau\left\{\frac{1}{R_k(\tau)}-\frac{F_k(\tau)R_k(u)}{R_k^2(\tau)}\right\}^2\frac{dH_k(u)}{G_k(u)}.$$

From this, it is shown that $n^{1/2}\left\{\hat{\xi}(\tau)-\xi(\tau)\right\}$ converges weakly to a normal distribution with mean zero and variance $p_1^{-1}V(U_1)+p_0^{-1}V(U_0)$, which can be estimated by replacing the unknown quantities by their empirical counterparts. Thus, an $(1-\alpha)$ asymptotic CI for $\xi(\tau)$ is given by

$$\hat{\xi}(\tau)\pm z_{1-\alpha/2}\sqrt{n_1^{-1}\hat{V}(U_1)+n_0^{-1}\hat{V}(U_0)}, \tag{7}$$

where $\hat{V}(U_k)$ is the variance estimator for $V(U_k)$, $(k=0,1)$.

For testing no treatment effect (i.e., $\xi(\tau)=0$), we will use

$$\hat{\xi}(\tau)/\sqrt{n_1^{-1}\hat{V}(U_1)+n_0^{-1}\hat{V}(U_0)} \tag{8}$$

as the test statistic, which asymptotically follows the standard normal distribution under the null.

As described above, RAH-SW and DAH-SW can be consistently estimated non-parametrically regardless of whether the PH assumption holds or not. Also, a reference number from the control group, $\eta_0(\tau)$, will help clinical investigators assess if the resulting RAH-SW (or DAH-SW) indicates a clinically meaningful magnitude or not. These will provide new interpretation of the hazard-based treatment effect that the existing methods could not offer. In Section 5, we will illustrate how to use the reference number $\eta_0(\tau)$ to aid interpretation of the result in practice.

3. Numerical Studies

3.1 *Configurations*

We evaluated finite sample properties of the proposed asymptotic CIs for RAH-SW and DAH-SW and asymptotic tests for no treatment effect (i.e., RAH-SW=1 and DAH-SW=0) via numerical studies. Four patterns of difference between two event time distributions were

considered — (A) No difference, (B) PH difference, (C) Early difference, and (D) Delayed difference (Figure 1).

[Figure 1 about here.]

The pattern (A) was included for evaluating the empirical type I error rate of the proposed tests. We used Weibull distributions to generate the event times for both groups. Specifically, for the treatment group, we used the Weibull distributions with shape and scale parameters of (1, 10), (1, 12.5), (1.5, 10), and (0.8, 15) for the patterns (A) to (D), respectively. For the control group, we used the Weibull distribution with shape and scale parameters of (1, 10) for all the patterns. Regarding censoring, we considered three patterns — (I) No censoring, (II) Light censoring, and (III) Heavy censoring (Figure 2), all of which had an administrative censoring at time 10. The fractions of censored observations at time 10^- were 0, 0.3, and 0.7 for (I), (II), and (III), respectively. Regarding the sample size, we considered two scenarios — n=100 and n=300 per arm. As such, we simulated a total of 24 configurations.

[Figure 2 about here.]

For each of the 24 simulation configurations, first, we generated n pairs of event time and censoring time for each group $\{(T_{ki}, C_{ki});\ k = 0, 1,\ i = 1, \ldots, n_k\}$ independently. Note that censoring time was independent of the event time and the same censoring distribution was used for both groups. We then derived the observable data $\{(X_{ki}, \Delta_{ki});\ k = 0, 1,\ i = 1, \ldots, n_k\}$, where $X_{ki} = \min(T_{ki}, C_{ki})$ and Δ_{ki} is equal to 1 if $T_{ki} \leqslant C_{ki}$ and 0 otherwise. With this data, we estimated RAH-SW and DAH-SW, constructed the 0.95CI for them using (5) and (7), respectively. Between-group comparisons were performed using the test statistics presented in (6) and (8) at two-sided 0.05 α level. We used τ=10 for the truncation time point for all scenarios. Repeating this process 5,000 times, we assessed the empirical bias of the proposed estimators for RAH-SW and DAH-SW, the empirical coverage probability of the proposed CIs, the average length of the CIs, and the empirical size and power of the tests.

As a reference, we included the HR based on Cox's method and assessed bias and coverage probability. It is somewhat difficult to determine what the true value for Cox's HR is under non-PH scenarios (Early difference (Figure 1C) and Delayed difference (Figure 1D)), because the population parameter to be estimated by Cox's method depends on the censoring time distribution when the PH assumption does not hold. In our simulations, we calculated the value for each non-PH scenario with the no censoring pattern (Figure 2-I), and called it "true" for Cox's HR.

For assessment of the empirical size and power of the tests, we also included tests based on the ratio of τ-year event rate and ratio of RMST as well as Cox's method. Because the RAH-SW is denoted by a product of the ratio of τ-year event rate and ratio of RMST (4), inclusion of these components would be interesting. Regarding the ratio of τ-year event and ratio of RMST, details of the test statistics used in our numerical studies are given in Appendix C. For the group comparison by Cox's method, the Wald test was used.

3.2 *Results*

We confirmed that the empirical bias of the proposed estimators for RAH-SW and DAH-SW was negligibly small and the coverage probability is sufficiently close to the nominal level for all scenarios with n=100 per arm (Table 1) and n=300 per arm (Table 2). On the other hand, as expected, we observed that Cox's HR gave us a biased estimate and the CI did not achieve the nominal coverage level under non-PH scenarios (i.e., Early and Delayed differences; Figures 1C and 1D) with the presence of light and heavy censoring (Figures 2-II and 2-III). The most pronounced case in our study was the combination of the early difference pattern (Figure 1C) and heavy censoring pattern (Figure 2-III) with n=300 per arm (Table 2). In this case, the bias of Cox's HR was -0.091, roughly 10% of the true HR (0.901). The coverage probability of 0.95CI based on Cox's method was 0.854, which was much lower than the nominal level 0.95. The average length of the CI of the RAH-SW was

almost identical to that of Cox's method for the no and light censoring patterns (Figures 2-I and 2-II), but it was slightly wider with heavy censoring scenarios (Figure 2-III) except for the delayed difference scenario (Figure 1D). For the delayed difference scenario (Figure 1D), no remarkable difference was seen in the average length of the CIs between RAH-SW and Cox's HR regardless of the censoring pattern.

[Table 1 about here.]

[Table 2 about here.]

The results of the five tests we considered are presented in Table 3 (for n=100) and Table 4 (for n=300.) First, the empirical size was assessed with the no difference pattern (Figure 1A). Since the number of iterations was 5,000, we considered that the empirical size should be within 0.044 to 0.056 with 95% chance if the true type I error rate is 5.0%. The empirical sizes of these tests were within this range except for the test based on the ratio of τ-year event rate with the combination of n=100 and heavy censoring (Figure 2-III).

Under the PH difference pattern (Figure 1B), power of the tests based on RAH-SW and DAH-SW were comparable to that of Cox's HR with no, light and heavy censoring patterns for both n=100 (Table 3) and n=300 (Table 4).

Under the early difference pattern (Figure 1C), the power of RAH-SW, DAH-SW and Cox's HR were similar with the no censoring case, but Cox's HR offered higher power with light and heavy censoring cases. The test based on RMST was superior for all censoring patterns for this scenario. The test based on τ-year event rate was the worst.

Under the delayed difference pattern (Figure 1D), again, for the no censoring case, no remarkable difference was seen among RAH-SW, DAH-SW and Cox's HR. However, with the presence of censoring, RAH-SW and DAH-SW gave higher power than Cox's HR. For this scenario, the test based on RMST was the worst for all censoring patterns. The test based on τ-year event rate gave the highest power.

RAH-SW and DAH-SW are comprised of ratio of τ-year event rate and ratio of RMST. As expected, the power of RAH-SW and DAH-SW were between tests based on ratio of τ-year event rate and ratio of RMST under the non-PH scenarios (Figures 1C and 1D). Interestingly, under the PH scenario (Figure 1B), RAH-SW and DAH-SW were more comparable to Cox's method than tests based on ratio of τ-year event rate and ratio of RMST.

In sum, in terms of power, RAH-SW and DAH-SW can be recommended when the expected pattern is PH difference or delayed difference. Throughout the scenarios, the performance of RAH-SW and DAH-SW were almost identical.

[Table 3 about here.]

[Table 4 about here.]

4. Empirical power comparisons with data from recent cancer clinical trials

The performance of a new method is usually assessed by Monte Carlo simulation studies as we presented in the previous section. However, these numerical studies rely on artificial data and simulation configurations are limited. In this section, we assessed the proposed method using empirical data to provide more convincing real-world evidence. Specifically, we used the same set of data used by Horiguchi et al. (2020). It consists of reconstructed patient-level data for OS and PFS from 69 and 54 phase III cancer trials, respectively. These studies were selected from the papers published in one of seven journals (Journal of the American Medical Association, JAMA Oncology, Journal of Clinical Oncology, Journal of the National Cancer Institute, Lancet, Lancet Oncology, and New England Journal of Medicine) between July 1st, 2016 and June 30th, 2017. Details of the eligibility criteria for the papers and studies are found in Horiguchi et al. (2020). The algorithm proposed by (Guyot et al., 2012) was used to reconstruct patient-level data from the figures in these papers.

We applied the tests based on RAH-SW and DAH-SW, and the Wald test via Cox's model

to each study data and compared their empirical power for OS and PFS, separately. The truncation time point for RAH-SW and DAH-SW was set to the study time point where the number at risk was at least 30 in both groups. Here, the empirical power was defined as the proportion of studies where the test gave a significant p-value (i.e., $\leqslant 0.05$ (two-sided)).

Figure 3 shows scatter plots of p-values from the tests based on RAH-SW and Cox's method for OS (3A) and PFS (3B). Most of the studies were distributed around the 45 degree diagonal line for both OS and PFS. We did not see a significant difference between the two tests in terms of empirical power. For OS, the empirical power of RAH-SW was 43.5%, which was numerically higher than that of Cox's method (37.7%). The difference (RAH-SW minus Cox's) was 5.8% and a corresponding 0.95CI (Liu et al., 2002) was -2.1% to 13.7%. We also did not observe a notable difference between the two tests for PFS. The empirical power of RAH-SW was identical to that of Cox's method (59.3%). The 0.95CI for the difference in the empirical power was -5.1% to 5.1%. The same analyses were performed for the test based on DAH-SW. Similar to the results of the numerical studies in the previous section, RAH-SW and DAH-SW provided almost identical performance. The empirical power of DAH-SW was almost identical to that of RAH-SW (results not shown). These results suggest that there is no clear power advantage of exclusively using Cox's method for all studies.

[Figure 3 about here.]

5. Example

We illustrate the proposed method using the data from a randomized phase III trial to compare nivolumab plus ipilimumab (treatment) with sunitinib (control) in patients with previously untreated clear-cell advanced renal-cell carcinoma (CheckMate 214 study). A total of 847 patients (425 for the treatment group and 422 for the control group) served for the analysis of PFS. We reconstructed patient-level data from the results reported by Motzer

et al. (2018), using the method proposed by Guyot et al. (2012). Figure 4 presents the Kaplan-Meier curves for PFS with the reconstructed patient-level data.

[Figure 4 about here.]

The two curves were similar to each other up to around 6 months and then diverged. This pattern of difference is the so-called "delayed difference" that is often seen in immunotherapy trials. Cancer immunotherapy does not attack cancer cells directly but instead directs the immune system to do so. It is believed that it takes time for immunotherapy to activate the immune system and thus, this would be a reason why the two Kaplan-Meier curves show such a delayed difference pattern. The HR (treatment over control) based on Cox's method was 0.82 (0.95CI: 0.68 to 0.99, p-value=0.037). The p-value of the cumulative residual test (Lin et al., 1993) for this study was not significant (0.084). However, it is important to note that a non-significant p-value from the PH assumption test does not imply that the PH assumption is indeed true.

We estimated the AH-SW, τ-year event rate, and RMST for each group. For each metric, we estimated both absolute difference (treatment minus control) and ratio (treatment over control) and corresponding 0.95CIs and p-values. The results were summarized in Table 5. Note that, in this example, we chose 21 (months) for τ because the number at risk at 21 months was greater than 30 for both groups (Figure 4). The estimated AH-SWs for the treatment and control groups were 0.049 (0.95CI: 0.042 to 0.057) and 0.066 (0.95CI: 0.057 to 0.076), respectively. The RAH-SW was 0.747 and the corresponding 0.95CI was (0.608 to 0.917). The p-value for testing RAH-SW=1 was 0.005. The estimated DAH-SW was -0.017 (0.95CI: -0.029 to -0.005). The resulting p-value for testing DAH-SW=0 was 0.006.

Similar to the τ-year event rate and RMST, using the AH-SW as a summary of event time distribution allows us to express the treatment effect in both absolute and relative terms (i.e., DAH-SW and RAH-SW). Moreover, since the AH-SW is the person-time event rate

when random censoring would not exist, DAH and RAH would be much more digestible than Cox's HR for clinical researchers. From the results, the event rates in the treatment group and control group, on average, are 4.9 events and 6.6 events per 100 at risk subjects per month, respectively. The RAH-SW 0.747 means that the combination of immunotherapy (nivolumab plus ipilimumab) reduces the event rate in the control (sunitinib) group by 25.3% on average. Note that Cox's HR approach cannot provide such a digestible interpretation due to the lack of a group-specific absolute hazard value in each group. We consider this is a notable advantage of using RAH-SW or DAH-SW over the conventional Cox's HR approach.

[Table 5 about here.]

6. Remarks

In this paper, we proposed AH-SW, a new summary metric of event time distribution, and RAH-SW and DAH-SW for quantifying the treatment effects and inference procedure for these metrics. The proposed method offers clinical investigators with a tool to summarize the hazard-based treatment effect magnitude in both absolute and relative terms. We believe this can be a solution that addresses the comments in the General Statistical Guidance by the Annals of Internal Medicine we introduced in Section 1.

RAH-SW and DAH-SW have several beneficial properties. First, they can be estimated consistently without imposing a strong model assumption (such as the PH assumption) between two event time distributions. Second, the AH-SW values from two groups used for RAH-SW and DAH-SW are available. These help us assess if the resulting RAH-SW (or DAH-SW) indicates a clinically meaningful treatment effect magnitude or not. As we showed in Section 5, using the proposed method enables us to provide a new digestible interpretation of the hazard-based treatment effect that was not possible with the conventional Cox's method. Third, neither RAH-SW or DAH-SW contradicts the stochastic ordering of two

event time distributions on $[0, \tau]$. That is, when a survival curve from Group A is always located higher than that from Group B on the time range $[0, \tau]$, the AH-SW of Group A is always lower than that of Group B, which indicates that Group A is better than Group B. Forth, they will always be reported along with the truncation time point τ, which will enhance understanding regarding the limitations about generalization of the study findings. Lastly, RAH-SW and DAH-SW offer similar power to Cox's method when the PH assumption holds. Also, our numerical studies demonstrated that they can be more powerful than Cox's method for the delayed difference patterns that are often seen in immunotherapy trials.

Regarding adjustment for prognostic factors, a stratified analysis with the proposed method would be straightforward. The proposed method can be extended to apply to observational data, such as when Conner et al. (2019) used it for estimating adjusted difference in RMST. Specifically, one can use the inverse probability weighting approach to get the adjusted Kaplan-Meier curve for the event time distribution in each group. Then, τ-year event rate, RMST, and AH-SW can be easily derived for each group.

Similar to RMST-based analyses, the choice of the truncation time τ is an important point to be addressed. For confirmatory studies, τ should be pre-specified in the study protocol. Given that the study findings are limited to the range of the study time, it would be possible to elicit a pre-specified τ at the design stage by considering the clinical questions investigators would like to have answered by that study. For example, if investigators believe that evaluating long-term treatment effect (e.g., 2 years) is necessary to determine whether the investigative drug is useful, τ will be 2 years. The patient accrual schedule and the additional follow-up time will be determined accordingly, so that the size of risk set at τ can be sufficiently large to rely on the large sample theories in both groups. When designing a study using conventional Cox's HR approach, investigators are projecting how long the study follow-up needs to be in order to observe a required number of events to achieve a

desired power. Therefore, we believe that such a number must exist in investigators' mind, even if vaguely.

Because the study findings are limited to the duration of the study time, clarifying the truncation time τ is important to enhance understanding of the treatment effect and its limitation regarding generalization. Related to this, it is important to note that Cox's HR also has an implicit truncation time point since information beyond that time point (i.e., the time when the last event was observed or the time when the size of risk set in either group became 0, whichever occurred first) does not contribute inference of the HR. Conventionally, such an implicit truncation time point has not been reported along with Cox's HR. However, we believe that it would be a good practice to explicitly report the truncation time τ regardless of the metric used for summarizing the treatment effect.

Since the variance formulae for $\log \hat{\theta}(\tau)$ and $\hat{\xi}(\tau)$ are given in this paper, the sample size calculation for designing a study using RAH-SW or DAH-SW would be straightforward. To calculate, one will specify the distributions of event time and censoring time for each group, the truncation time τ, and the ratio to allocate subjects to group 1, and determine $\log \theta(\tau)$ (or $\xi(\tau)$) and the variances $V(Q_0)$ and $V(Q_1)$ (or $V(U_0)$ and $V(U_1)$). The required total sample size to achieve a $(1-\beta)$ power, at a two-sided α level, will then be

$$n = \left\{ \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right) \sqrt{p_1^{-1} V(Q_1) + p_0^{-1} V(Q_0)}}{\log \theta(\tau)} \right\}^2$$

or

$$n = \left\{ \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right) \sqrt{p_1^{-1} V(U_1) + p_0^{-1} V(U_0)}}{\xi(\tau)} \right\}^2 .$$

Financial disclosure

This research was supported by McGraw/Patterson Research Fund.

Conflict of interest

The authors declare no potential conflict of interests.

Appendix A. Large sample properties of Q_k

We use the same notation and the assumption in Section 2. First, we note well-known results about $\hat{F}_k(\cdot)$ and $\hat{R}_k(\cdot)$. As it is shown by Fleming and Harrington (1991),

$$\sqrt{n}_k \left\{ \frac{\hat{F}_k(\tau) - F_k(\tau)}{1 - F_k(\tau)} \right\} = n_k^{-1/2} \sum_{i=1}^{n_k} \int_0^\tau \frac{dM_{ki}(u)}{G_k(u)} + o_p(1), \tag{A.1}$$

and this converges weakly to a zero-mean normal distribution, where $G_k(t) = \Pr(X_k \geqslant t)$, $M_{ki}(t) = N_{ki}(t) - \int_0^t Y_{ki}(s) dH_k(s)$, $N_{ki}(t) = I(X_{ki} \leqslant t, \Delta_{ki} = 1)$, and $Y_{ki}(t) = I(X_{ki} \geqslant t)$. Also, from the results shown by Zhao et al. (2012),

$$\sqrt{n}_k \left\{ \hat{R}_k(\tau) - R_k(\tau) \right\} = -n_k^{-1/2} \sum_{i=1}^{n_k} \int_0^\tau \left\{ \int_u^\tau S_k(t) dt \right\} \frac{dM_{ki}(u)}{G_k(u)} + o_p(1), \tag{A.2}$$

which converges weakly to a zero-mean normal distribution.

Next, applying the Taylor series expansion, coupled with the results of (A.1) and (A.2), $Q_k = n_k^{1/2} \{ \log \hat{\eta}_k(\tau) - \log \eta_k(\tau) \}$ is denoted by

$$Q_k = n_k^{1/2} F_k^{-1}(\tau) \left\{ \hat{F}_k(\tau) - F_k(\tau) \right\} - n_k^{1/2} R_k^{-1}(\tau) \left\{ \hat{R}_k(\tau) - R_k(\tau) \right\} + o_p(1).$$

Using (A.1) and (A.2),

$$Q_k = n_k^{-1/2} \sum_{i=1}^{n_k} \int_0^\tau \left\{ \frac{1}{F_k(\tau)} - \frac{R_k(u)}{R_k(\tau)} \right\} \frac{dM_{ki}(u)}{G_k(u)} + o_p(1).$$

By the martingale central limit theorem, it is shown that Q_k converges weakly to a normal distribution with mean zero and variance

$$V(Q_k) = \int_0^\tau \left\{ \frac{1}{F_k(\tau)} - \frac{R_k(u)}{R_k(\tau)} \right\}^2 \frac{dH_k(u)}{G_k(u)}.$$

Appendix B. Large sample properties of U_k

We rewrite $U_k = n_k^{1/2} \{ \hat{F}_k(\tau)/\hat{R}_k(\tau) - F_k(\tau)/R_k(\tau) \}$ by

$$U_k = n_k^{1/2} \{ \hat{F}_k(\tau)/\hat{R}_k(\tau) - F_k(\tau)/\hat{R}_k(\tau) \} + n_k^{1/2} \{ F_k(\tau)/\hat{R}_k(\tau) - F_k(\tau)/R_k(\tau) \}. \tag{A.3}$$

By the application of Taylor series expansion, coupled with the results of (A.1) and (A.2), it is shown that the first term of (A.3) is $n_k^{1/2} \{ \hat{F}_k(\tau)/R_k(\tau) - F_k(\tau)/R_k(\tau) \} + o_p(1)$. Also, by the application of Taylor series expansion to the second term of (A.3), coupled with the

result of (A.2), U_k is denoted by

$$U_k = n_k^{1/2} R_k^{-1}(\tau) \left\{ \hat{F}_k(\tau) - F_k(\tau) \right\} - n_k^{1/2} F_k(\tau) R_k^{-2}(\tau) \left\{ \hat{R}_k(\tau) - R_k(\tau) \right\} + o_p(1). \quad \text{(A.4)}$$

From (A.1) and (A.2),

$$U_k = n_k^{-1/2} \sum_{i=1}^{n_k} \int_0^\tau \left\{ \frac{1}{R_k(\tau)} - \frac{F_k(\tau) R_k(u)}{R_k^2(\tau)} \right\} \frac{dM_{ki}(u)}{G_k(u)} + o_p(1).$$

Therefore, by the martingale central limit theorem, it is shown that U_k converges weakly to a normal distribution with mean zero and variance

$$V(U_k) = \int_0^\tau \left\{ \frac{1}{R_k(\tau)} - \frac{F_k(\tau) R_k(u)}{R_k^2(\tau)} \right\}^2 \frac{dH_k(u)}{G_k(u)}.$$

Appendix C. Test statistics used in the simulation study

For group comparisons based on ratio of τ-year event rate and ratio of RMST, we used the following results. From (A.1) and (A.2), $\mathcal{W}_k^F = n_k^{1/2}\{\log \hat{F}_k(\tau) - \log F_k(\tau)\}$ and $\mathcal{W}_k^R = n_k^{1/2}\{\log \hat{R}_k(\tau) - \log R_k(\tau)\}$ converge to zero-mean normal distribution with the variance $V(\mathcal{W}_k^F) = \int_0^\tau \left\{ \frac{1-F_k(\tau)}{F_k(\tau)} \right\}^2 \frac{dH_k(u)}{G_k(u)}$ and $V(\mathcal{W}_k^R) = \int_0^\tau \left\{ \frac{R_k(u)}{R_k(\tau)} \right\}^2 \frac{dH_k(u)}{G_k(u)}$, respectively. Thus, for testing the null hypotheses, $\log F_1(\tau) - \log F_0(\tau) = 0$ and $\log R_1(\tau) - \log R_0(\tau) = 0$, we used $\log \left\{ \frac{\hat{F}_1(\tau)}{\hat{F}_0(\tau)} \right\} / \sqrt{n_1^{-1}\hat{V}(\mathcal{W}_1^F) + n_0^{-1}\hat{V}(\mathcal{W}_0^F)}$ and $\log \left\{ \frac{\hat{R}_1(\tau)}{\hat{R}_0(\tau)} \right\} / \sqrt{n_1^{-1}\hat{V}(\mathcal{W}_1^R) + n_0^{-1}\hat{V}(\mathcal{W}_0^R)}$ as the test statistic, respectively.

References

A'Hern, R. P. (2016). Restricted mean survival time: an obligatory end point for time-to-event analysis in cancer trials? *J Clin Oncol* **34,** 3474–3476.

Annals of Internal Medicine, ., Information for authors: General statistical guidance (section 4: Measures of effect and risk)., https://www.acpjournals.org/journal/aim/authors/statistical-guidance (Accessed May. 21, 2021).

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

Conner, S. C., Sullivan, L. M., Benjamin, E. J., LaValley, M. P., Galea, S., and Trinquart, L. (2019). Adjusted restricted mean survival times in observational studies. *Stat Med* **38,** 3832–3860.

Cox, D. R. (1972). Regression models and life-tables. *J R Stat Soc Series B Stat Methodol* **34,** 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62,** 269–276.

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

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

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology* **21,** 13–15.

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., Hassett, M. J., and Uno, H. (2020). Empirical power comparison of statistical tests in contemporary phase III randomized controlled trials with time-to-event outcomes in oncology. *Clin Trials* **17,** 597–606.

Kalbfleisch, J. D. and Prentice, R. L. (1981). Estimation of the average hazard ratio. *Biometrika* **68,** 105–112.

Lin, D. and Wei, L. J. (1989). The robust inference for the cox proportional hazards model. *J. Am. Stat. Assoc.* **84,** 1074–1078.

Lin, D. Y., Wei, L. J., and Ying, Z. (1993). Checking the Cox model with cumulative sums

of martingale-based residuals. *Biometrika* **80,** 557–572.

Liu, J. P., Hsueh, H. M., Hsieh, E., and Chen, J. J. (2002). Tests for equivalence or non-inferiority for paired binary data. *Stat Med* **21,** 231–245.

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.

Péron, J., Roy, P., Ozenne, B., Roche, L., and Buyse, M. (2016). The net chance of a longer survival as a patient-oriented measure of treatment benefit in randomized clinical trials. *JAMA Oncol* **2,** 901–905.

Royston, P. and Parmar, M. K. (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.

Saad, E. D., Zalcberg, J. R., Péron, J., Coart, E., Burzykowski, T., and Buyse, M. (2018). Understanding and communicating measures of treatment effect on survival: can we do better? *J Natl Cancer Inst* **110,** 232–240.

Schemper, M. (1992). Cox analysis of survival data with non-proportional hazard functions. *Statistician* **41,** 455–465.

Schemper, M., Wakounig, S., and Heinze, G. (2009). The estimation of average hazard ratios by weighted Cox regression. *Stat Med* **28,** 2473–2489.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika* **73,** 363–369.

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., Skali, H., Solomon, S. D., 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. D., Claggett, B., Tian, L., Cai, T., Pfeffer, M. A., Evans, S. R., and Wei, L. J. (2015). Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies. *Ann Intern Med* **163,** 127–134.

Xu, R. and O'Quigley, J. (2000). Estimating average regression effect under non-proportional hazards. *Biostatistics* **1,** 423–439.

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.

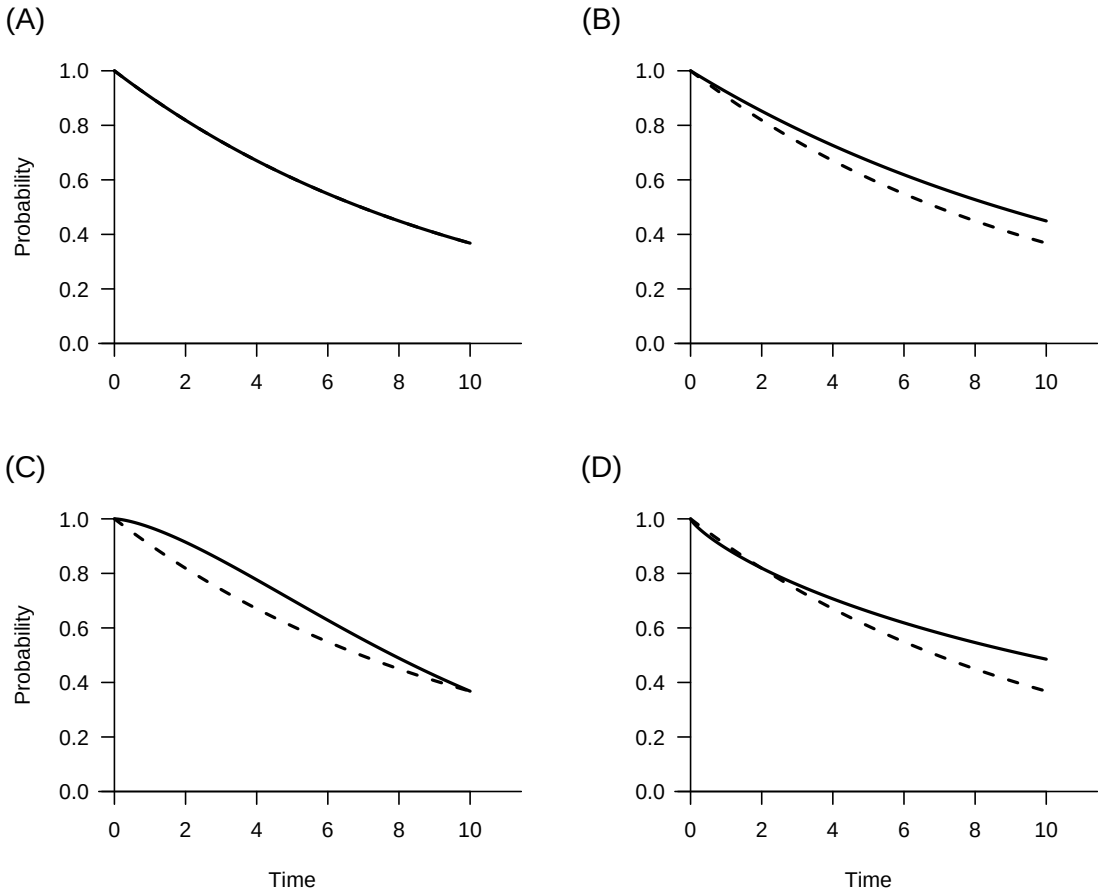

Figure 1. Survival functions of event time distributions of the treatment group (solid line) and control group (dashed line) used in the numerical studies. A, no difference; B, proportional hazards difference; C, early difference; D, delayed difference.

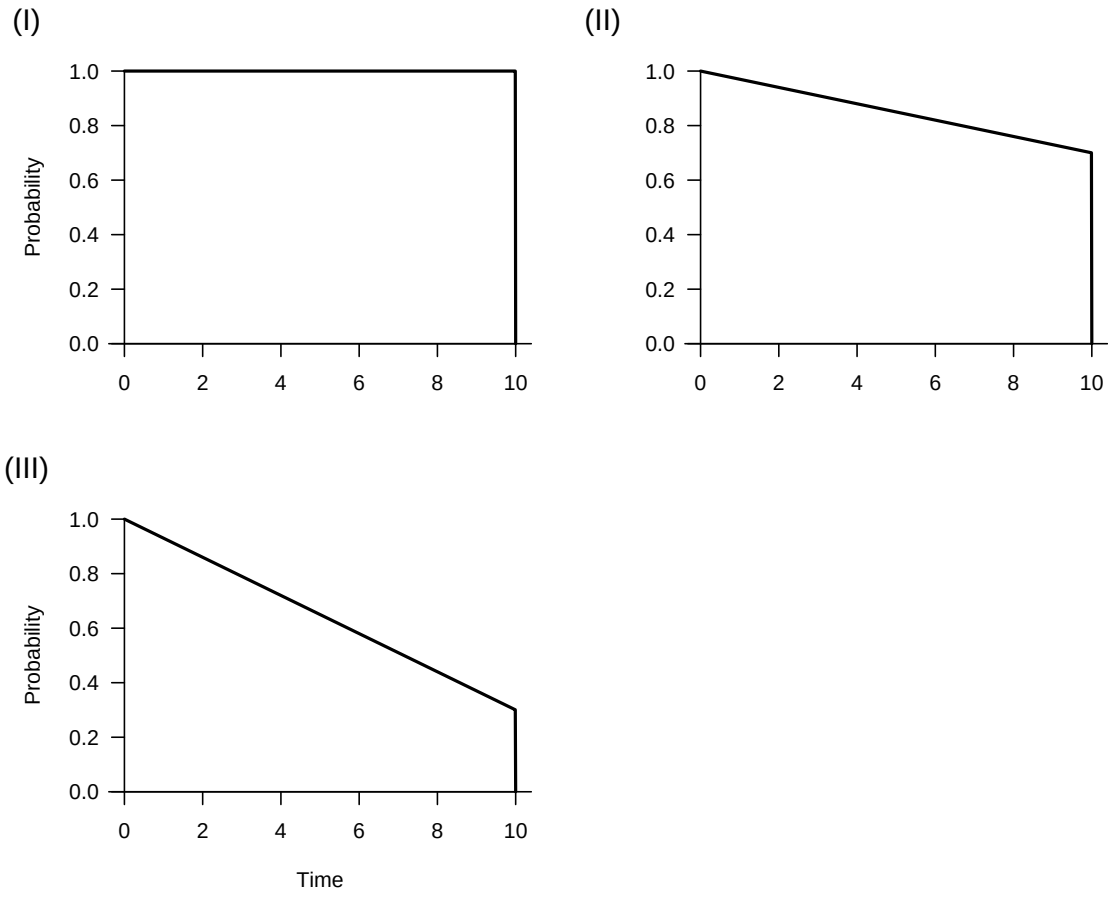

Figure 2. Survival functions of censoring time distributions used in the numerical studies. I, no censoring; II, light censoring; III, heavy censoring.

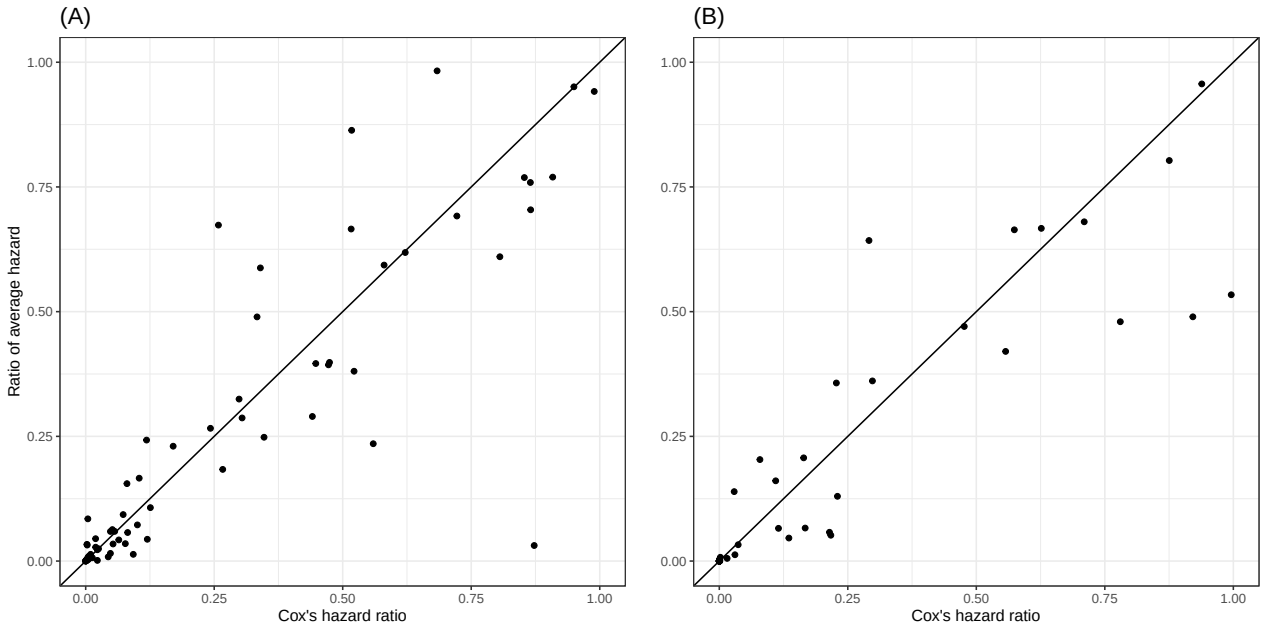

Figure 3. Scatter plots of p-values from the tests based on ratio of average hazard with the survival weight (RAH-SW) and Cox's method for A) overall survival (69 studies) and B) progression-free survival (54 studies).

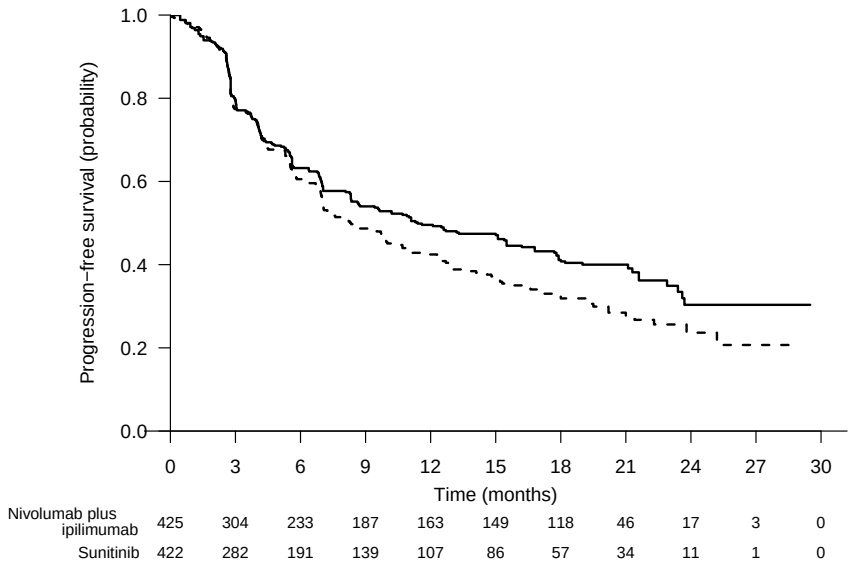

Figure 4. The Kaplan-Meier curves of progression-free survival time for nivolumab plus ipilimumab group (solid line) and sunitinib group (dashed line) with the data reconstructed from the publication of the CheckMate 214 study.

Table 1

Performance of the difference and ratio of average hazard with survival weight, and Cox's hazard ratio with sample size 100 per arm.

Event time distribution pattern: No difference												
	DAH-SW				RAH-SW				Cox's HR			
Censoring	True	Bias	CP	AL	True	Bias	CP	AL	True	Bias	CP	AL
None	0.000	0.000	0.950	0.070	1.000	0.018	0.948	0.720	1.000	0.018	0.952	0.727
Light	0.000	-0.000	0.947	0.075	1.000	0.016	0.946	0.774	1.000	0.015	0.950	0.779
Heavy	0.000	-0.000	0.949	0.086	1.000	0.025	0.949	0.911	1.000	0.024	0.953	0.880
Event time distribution pattern: PH difference												
	DAH-SW				RAH-SW				Cox's HR			
Censoring	True	Bias	CP	AL	True	Bias	CP	AL	True	Bias	CP	AL
None	-0.020	-0.000	0.950	0.065	0.800	0.013	0.949	0.597	0.800	0.014	0.951	0.603
Light	-0.020	-0.000	0.950	0.070	0.800	0.014	0.949	0.645	0.800	0.015	0.952	0.649
Heavy	-0.020	-0.000	0.944	0.081	0.800	0.019	0.946	0.760	0.800	0.018	0.952	0.733
Event time distribution pattern: Early difference												
	DAH-SW				RAH-SW				Cox's HR			
Censoring	True	Bias	CP	AL	True	Bias	CP	AL	True	Bias	CP	AL
None	-0.010	-0.000	0.952	0.063	0.903	0.014	0.949	0.616	0.901	0.017	0.951	0.655
Light	-0.010	-0.000	0.950	0.068	0.903	0.014	0.949	0.669	0.901	-0.018	0.947	0.683
Heavy	-0.010	-0.000	0.944	0.080	0.903	0.020	0.946	0.800	0.901	-0.080	0.920	0.728
Event time distribution pattern: Delayed difference												
	DAH-SW				RAH-SW				Cox's HR			
Censoring	True	Bias	CP	AL	True	Bias	CP	AL	True	Bias	CP	AL
None	-0.025	-0.000	0.950	0.065	0.754	0.013	0.948	0.589	0.759	0.013	0.950	0.583
Light	-0.025	-0.000	0.949	0.070	0.754	0.014	0.947	0.633	0.759	0.030	0.947	0.638
Heavy	-0.025	-0.000	0.943	0.080	0.754	0.020	0.943	0.739	0.759	0.064	0.944	0.741

Event time distribution pattern: 1A, no difference; 1B, PH difference; 1C, early difference; 1D, delayed difference (see Figure 1). Censoring time distribution pattern: 2-I, no censoring; 2-II, light censoring; 2-III, heavy censoring (see Figure 2).

Abbreviations: DAH-SW, difference in average hazard with the survival weight; RAH-SW, ratio of average hazard with the survival weight; HR, hazard ratio; PH, proportional hazards; Censoring, censoring time distribution pattern; True, the true value; Bias, the empirical bias (estimate minus true value); CP, the empirical coverage probability of the 0.95 confidence interval; AL, the average length of the 0.95 confidence interval.

Table 2

Performance of the difference and ratio of average hazard with survival weight, and Cox's hazard ratio with sample size 300 per arm.

Event time distribution pattern: No difference												
	DAH-SW				RAH-SW				Cox's HR			
Censoring	True	Bias	CP	AL	True	Bias	CP	AL	True	Bias	CP	AL
None	0.000	0.000	0.950	0.040	1.000	0.006	0.949	0.407	1.000	0.006	0.950	0.408
Light	0.000	-0.000	0.950	0.043	1.000	0.003	0.950	0.436	1.000	0.003	0.951	0.436
Heavy	0.000	-0.000	0.947	0.050	1.000	0.007	0.948	0.510	1.000	0.007	0.952	0.488
Event time distribution pattern: PH difference												
	DAH-SW				RAH-SW				Cox's HR			
Censoring	True	Bias	CP	AL	True	Bias	CP	AL	True	Bias	CP	AL
None	-0.020	-0.000	0.950	0.037	0.800	0.005	0.950	0.338	0.800	0.005	0.952	0.339
Light	-0.020	-0.000	0.952	0.040	0.800	0.004	0.952	0.364	0.800	0.005	0.954	0.363
Heavy	-0.020	-0.000	0.950	0.047	0.800	0.006	0.952	0.426	0.800	0.006	0.951	0.406
Event time distribution pattern: Early difference												
	DAH-SW				RAH-SW				Cox's HR			
Censoring	True	Bias	CP	AL	True	Bias	CP	AL	True	Bias	CP	AL
None	-0.010	0.000	0.950	0.037	0.903	0.005	0.949	0.349	0.901	0.006	0.949	0.368
Light	-0.010	-0.000	0.951	0.040	0.903	0.005	0.952	0.379	0.901	-0.029	0.934	0.383
Heavy	-0.010	-0.000	0.949	0.047	0.903	0.006	0.950	0.450	0.901	-0.091	0.854	0.404
Event time distribution pattern: Delayed difference												
	DAH-SW				RAH-SW				Cox's HR			
Censoring	True	Bias	CP	AL	True	Bias	CP	AL	True	Bias	CP	AL
None	-0.025	0.000	0.952	0.038	0.754	0.005	0.950	0.332	0.759	0.005	0.949	0.328
Light	-0.025	-0.000	0.949	0.040	0.754	0.004	0.950	0.356	0.759	0.020	0.946	0.356
Heavy	-0.025	-0.000	0.948	0.047	0.754	0.006	0.946	0.412	0.759	0.051	0.927	0.411

Event time distribution pattern: 1A, no difference; 1B, PH difference; 1C, early difference; 1D, delayed difference (see Figure 1). Censoring time distribution pattern: 2-I, no censoring; 2-II, light censoring; 2-III, heavy censoring (see Figure 2).

Abbreviations: DAH-SW, difference in average hazard with the survival weight; RAH-SW, ratio of average hazard with the survival weight; HR, hazard ratio; PH, proportional hazards; Censoring, censoring time distribution pattern; True, the true value; Bias, the empirical bias (estimate minus true value); CP, the empirical coverage probability of the 0.95 confidence interval; AL, the average length of the 0.95 confidence interval.

Table 3

Size and power of tests based on difference in average hazard, ratio of average hazard, Cox's hazard ratio, ratio of τ-year event rate, and ratio of restricted mean survival time for sample size of 100 per arm.

Size of tests
Event time distribution pattern: No difference

	Test				
Censoring	DAH-SW	RAH-SW	Cox's HR	Ratio (τ-year)	Ratio (RMST)
None	0.050	0.052	0.048	0.050	0.051
Light	0.053	0.054	0.050	0.055	0.056
Heavy	0.051	0.051	0.047	0.059	0.056

Power of tests
Event time distribution pattern: PH difference

	Test				
Censoring	DAH-SW	RAH-SW	Cox's HR	Ratio (τ-year)	Ratio (RMST)
None	0.226	0.228	0.221	0.216	0.201
Light	0.206	0.208	0.205	0.190	0.193
Heavy	0.167	0.169	0.166	0.155	0.179

Event time distribution pattern: Early difference

	Test				
Censoring	DAH-SW	RAH-SW	Cox's HR	Ratio (τ-year)	Ratio (RMST)
None	0.086	0.093	0.086	0.050	0.291
Light	0.081	0.087	0.112	0.048	0.278
Heavy	0.079	0.081	0.167	0.064	0.252

Event time distribution pattern: Delayed difference

	Test				
Censoring	DAH-SW	RAH-SW	Cox's HR	Ratio (τ-year)	Ratio (RMST)
None	0.315	0.313	0.305	0.396	0.171
Light	0.281	0.279	0.242	0.341	0.163
Heavy	0.229	0.229	0.156	0.255	0.153

Event time distribution pattern: 1A, no difference; 1B, PH difference; 1C, early difference; 1D, delayed difference (see Figure 1). Censoring time distribution pattern: 2-I, no censoring; 2-II, light censoring; 2-III, heavy censoring (see Figure 2).
Abbreviations: DAH-SW, difference in average hazard with the survival weight; RAH-SW, ratio of average hazard with the survival weight; HR, hazard ratio; Ratio (τ-year), ratio of τ-year event rate; Ratio (RMST), ratio of restricted mean survival time; PH, proportional hazards; Censoring, censoring time distribution patterns.

Table 4

Size and power of tests based on difference in average hazard, ratio of average hazard, Cox's hazard ratio, ratio of τ-year event rate, and ratio of restricted mean survival time for sample size of 300 per arm.

Size of tests
Event time distribution pattern: No difference

	Test				
Censoring	DAH-SW	RAH-SW	Cox's HR	Ratio (τ-year)	Ratio (RMST)
None	0.050	0.051	0.050	0.050	0.049
Light	0.050	0.050	0.049	0.050	0.052
Heavy	0.053	0.052	0.048	0.055	0.051

Power of tests
Event time distribution pattern: PH difference

	Test				
Censoring	DAH-SW	RAH-SW	Cox's HR	Ratio (τ-year)	Ratio (RMST)
None	0.553	0.555	0.551	0.537	0.486
Light	0.503	0.505	0.502	0.460	0.460
Heavy	0.391	0.394	0.417	0.324	0.414

Event time distribution pattern: Early difference

	Test				
Censoring	DAH-SW	RAH-SW	Cox's HR	Ratio (τ-year)	Ratio (RMST)
None	0.175	0.182	0.172	0.050	0.676
Light	0.157	0.163	0.248	0.049	0.647
Heavy	0.123	0.128	0.402	0.057	0.582

Event time distribution pattern: Delayed difference

	Test				
Censoring	DAH-SW	RAH-SW	Cox's HR	Ratio (τ-year)	Ratio (RMST)
None	0.717	0.715	0.710	0.831	0.399
Light	0.674	0.672	0.596	0.760	0.372
Heavy	0.549	0.549	0.400	0.595	0.341

Event time distribution pattern: 1A, no difference; 1B, PH difference; 1C, early difference; 1D, delayed difference (see Figure 1). Censoring time distribution pattern: 2-I, no censoring; 2-II, light censoring; 2-III, heavy censoring (see Figure 2).
Abbreviations: DAH-SW, difference in average hazard with the survival weight; RAH-SW, ratio of average hazard with the survival weight; HR, hazard ratio; Ratio (τ-year), ratio of τ-year event rate; Ratio (RMST), ratio of restricted mean survival time; PH, proportional hazards; Censoring, censoring time distribution patterns.

Table 5

Estimated τ-year event rates, restricted mean survival times, and average hazards with survival weight for treatment group (nivolumab plus ipilimumab) and control group (sunitinib) with the data reconstructed from the publication of the CheckMate 214 study.

	Treatment (0.95CI)	Control (0.95CI)	Difference* (0.95CI; p-value)	Ratio** (0.95CI; p-value)
τ-year event rate [%]	60.0 (54.6 to 65.3)	71.5 (66.4 to 76.9)	-11.5 (-19.3 to -3.8; 0.004)	0.8 (0.7 to 0.9; 0.003)
RMST [month]	12.2 (11.4 to 13.0)	11.0 (10.2 to 11.8)	1.2 (0.0 to 2.4; 0.043)	1.1 (1.0 to 1.2; 0.041)
AH-SW	0.049 (0.042 to 0.057)	0.066 (0.057 to 0.076)	-0.017 (-0.029 to -0.005; 0.006)	0.747 (0.608 to 0.917; 0.005)

Abbreviations: RMST, restricted mean survival time; AH-SW, average hazard with survival weight; 0.95CI, 0.95 confidence interval. We set τ to be 21 months for estimating the τ-year event rate, RMST, and AH-SW.

* Difference: Treatment − Control. A value below 0 is in favor of treatment group for τ-year event rate and AH-SW, and that above 0 is in favor of treatment group for RMST.

** Ratio: Treatment/Control. A value below 1 is in favor of treatment group for τ-year event rate and AH-SW, and that above 1 is in favor of treatment group for RMST.