The Oncologist®

Commentary

How Do the Accrual Pattern and Follow-Up Duration Affect the Hazard Ratio Estimate When the Proportional Hazards Assumption Is Violated?

Miki Horiguchi,[a] Michael J. Hassett,[b] Hajime Uno[b,c]

[a]Department of Clinical Medicine (Biostatistics), Kitasato University Graduate School of Pharmaceutical Sciences, Tokyo, Japan; Departments of [b]Medical Oncology and [c]Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA

Disclosures of potential conflicts of interest may be found at the end of this article.

Key Words. Clinical trials • Cox proportional hazards model • Survival data analysis • Time-to-event outcomes

Time-to-event outcomes, such as overall survival and progression-free survival, are often selected as key endpoints in comparative clinical studies. When analyzing results from these studies, estimating the magnitude of the between-group difference is a fundamental step. This is especially important in clinical trials that compare treatments, because quantitative information about the treatment effect is necessary to balance benefits/risks and inform clinical decision making. Conventionally, many comparative oncology studies use the hazard ratio (HR) and a corresponding 95% confidence interval to report the magnitude of the between-group difference, despite a number of important limitations [1–6].

One well-known limitation of HR is the need for a proportional hazards (PH) assumption [1–6]. Specifically, the ratio of the hazard functions for the two groups must be constant over time. Although this limitation is well known, many papers report the HR even when the PH assumption is not met [7–9]. Arguing in favor of using the standard Cox's method [10] and reporting the HR in all circumstances, investigators assert that it represents an approximate average of the ratio of two hazard functions [11, 12]. However, this is a misunderstanding, because study-specific parameters that should have no bearing on the between-group difference can affect the hazard ratio estimation. For example, two clinical trials evaluating the same intervention and conducted on the same population could produce different HR estimates in a non-PH scenario, no matter how many subjects enroll, just because the duration of follow-up time or the accrual profile differ. Consequently, when the HR is used to summarize the between-group difference using the standard Cox's method [10] in non-PH scenarios, it is not obvious what "average" quantity the HR is estimating; the resulting HR estimate is clinically uninterpretable and may mislead clinical decision [1, 2]. Our goals were to demonstrate this critical limitation using simulated patient data for PH and non-PH scenarios and to illustrate this limitation by using results from a recent clinical study.

Numerical Studies

First, we described two patterns of between-group differences common to comparative oncology clinical trials. The first pattern was PH, with a constant HR of 0.78. Figure 1A shows the survival functions of this pattern (top) and corresponding hazard functions (middle). Figure 1A (bottom) shows the ratio of the hazard functions for the two groups (i.e., time-specific HR). The second pattern was non-PH. Survival functions and hazard functions by group, and their corresponding time-specific HRs, are depicted in Figure 1B (top, middle, and bottom, respectively). This pattern (i.e., survival curves are separated for early time points and then the curves overlap later) is one of the common non-PH scenarios in cancer clinical trials (e.g., Paridaens et al. [7] and Wilke et al. [8]).

Second, we identified two study parameters that should not affect the treatment effect estimate: the patient accrual profile and the follow-up time. In clinical trials, patient enrollment does not necessarily follow a consistent pattern. Sometimes, accrual takes only a few months, but other times, accrual takes much longer. Also, the duration of follow-up after accrual may vary from study to study. Figure 2 illustrates several accrual and follow-up patterns. The *y*-axis indicates the total number of patients enrolled, and the *x*-axis indicates calendar time. *N* denotes the total sample size of the study. T_1 and T_2 indicate the lengths of the accrual period and the postaccrual follow-up period, respectively. Therefore, $T_1 + T_2$ denotes the length of the total study time (i.e., time from the first patient enrolled to the end of trial). Note that in typical event-driven trials, the end of trial is determined by the time when a required number of events are observed and is therefore not a fixed time point.

For our numerical studies, four combinations of T_1 and T_2 were considered: (a) immediate enrollment, (b) short accrual, (c) intermediate accrual, and (d) long accrual compared with follow-up time (Fig. 2). Specifically, (T_1, T_2) was

Correspondence: Hajime Uno, Ph.D., Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, Massachusetts 02215, USA. Telephone: 1-617-632-6893; e-mail: huno@jimmy.harvard.edu Received March 6, 2018; accepted for publication June 25, 2018; published Online First on September 10, 2018. http://dx.doi.org/10.1634/theoncologist.2018-0141

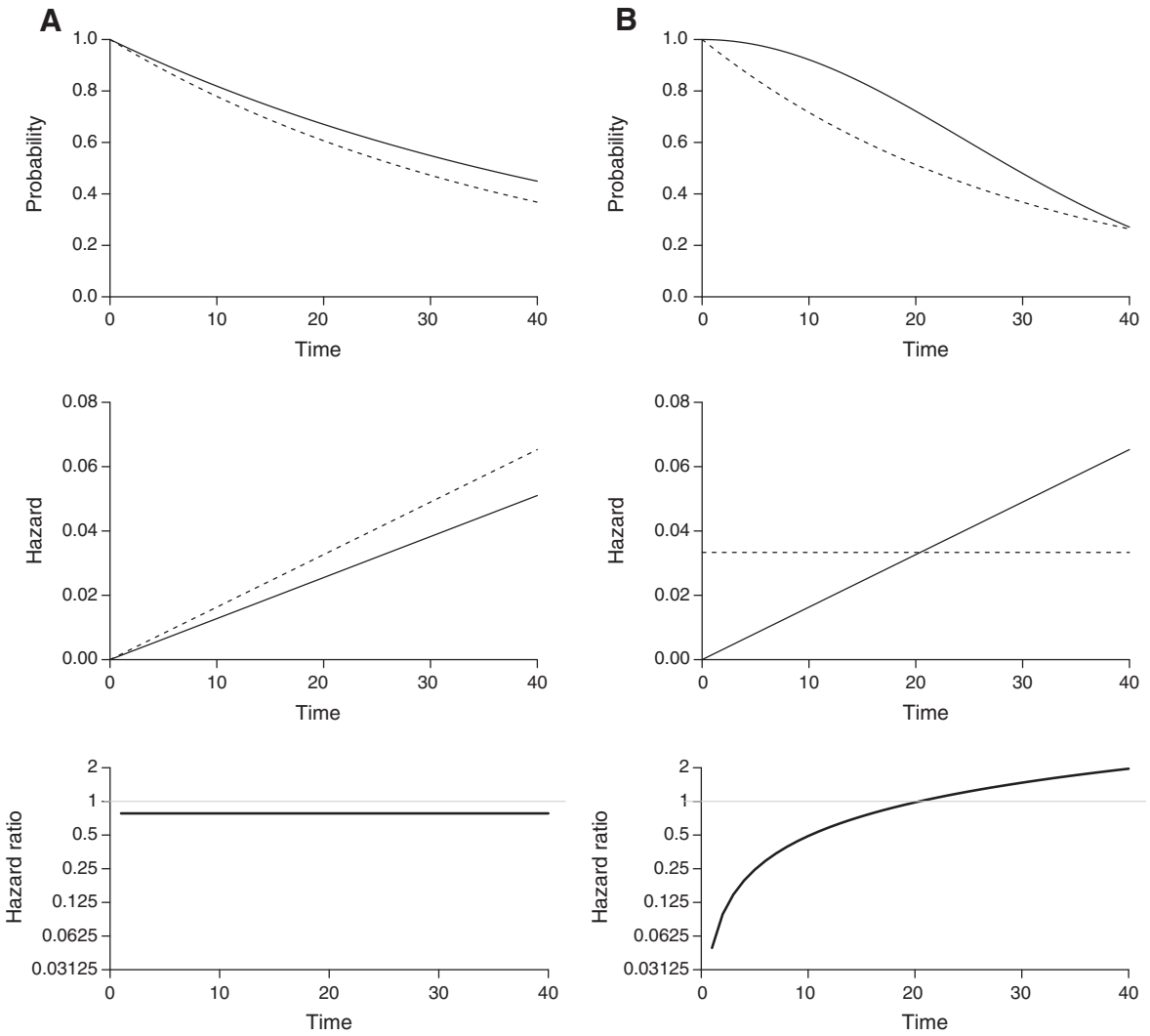

Figure 1. Survival functions, hazard functions, and the ratio of two hazard functions. Survival functions (top) and corresponding hazard functions (middle) by group, and the ratio of two hazard functions (time-specific hazard ratio) (bottom) in a pattern of proportional hazards **(A)** and one of non-proportional hazards **(B)**.

(0 and 40 months) for (a), (8 and 32 months) for (b), (16 and 24 months) for (c), and (32 and 8 months) for (d). Note that we had the immediate enrollment pattern (a) as a reference, although it is not realistic to enroll all subjects simultaneously. Combining these four patterns of accrual and follow-up with the two patterns of between-group differences (i.e., PH and non-PH), eight clinical trial scenarios were defined. All scenarios assumed a uniform enrollment rate throughout the given accrual time, except the immediate enrollment case.

The objective of our numerical studies was to determine what population quantity (i.e., HR) the standard Cox's inference procedure [10] estimates, and to demonstrate how the HR estimate depends on study-specific parameters (i.e., the accrual and/or follow-up time). Using a substantial number of data points for each scenario allowed us to be confident that observed differences in the computed HR across different scenarios were not due to sampling variability. Specifically, for each scenario, we generated 2,000,000 observations (i.e., 1,000,000 simulated observations for each treatment group). The number of events for each of the eight scenarios was at least 773,000, making the length of the 0.95 confidence interval for the HR no greater than 0.006 for all scenarios. This allowed us to compare the computed HRs across different scenarios without regard to the sampling variability.

For each subject, we generated time from the study activation to the enrollment and time from the enrollment to event occurrence. The former was generated from the uniform distribution as described in Figure 2, and the latter was generated from the event time distribution shown in Figure 1 (top). Event time was censored at the end of the follow-up period (T_1 + T_2) or the timing of the final data analysis (see supplemental online Data A for details of the data generation procedure). With these simulated data, the HR between two groups was calculated using the Cox procedure. After calculating HR estimates for all scenarios, we looked for variation in these estimates across the four PH scenarios and across the four non-PH scenarios. We reasoned that differences in T_1 and T_2 should not translate into differences in HR estimates.

Results from all eight scenarios appear in Table 1. In PH scenarios, all estimated HRs were the same as the true HR

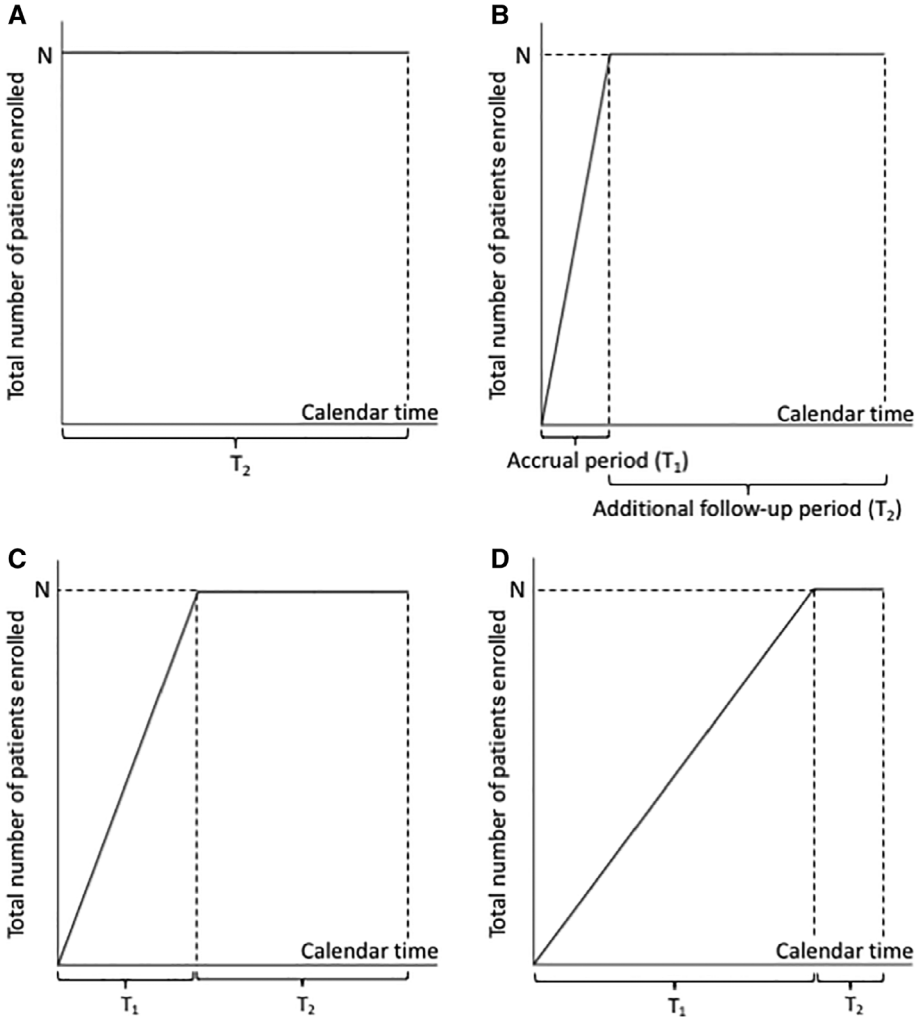

Figure 2. Four patterns of patient accrual and length of additional follow-up. Panels **(A)** through **(D)** correspond to patterns (a)–(d) in Table 1.
Abbreviations: N, total number of patients enrolled; T_1, accrual period; T_2, additional follow-up period after the end of the accrual.

of 0.78. In other words, different accrual and follow-up patterns did not yield different HR estimates (Table 1). However, in non-PH scenarios, HR estimates for the four accrual and follow-up patterns were 0.78, 0.73, 0.67, and 0.56, respectively (Table 1). So each HR estimate was not a simple average of the time-specific HR (Fig. 1B, bottom) but rather depended on study-specific parameters that should have had no impact on the treatment effect estimate. The amount of variation in the HR estimates for the non-PH scenarios was not negligible, ranging from 0.56 to 0.78. Arguably, these differences are of sufficient magnitude to affect clinical decision making. If treatment plans were created from these HR estimates, then study parameters unrelated to the actual treatment benefit could have affected clinical decision making.

Next, we examined the magnitude of variation in the HR estimate using data from a recent clinical trial for immunotherapy—a study that evaluated the efficacy and safety of nivolumab versus docetaxel in patients with advanced nonsquamous non-small-cell lung cancer [9]. The top chart in Figure 3 shows the Kaplan-Meier curves of overall survival by group; the reported HR was 0.73 [9]. As is often seen in immunotherapy trials, the survival curves present a delayed difference pattern. The bottom chart in Figure 3 plots the corresponding time-specific HR [13], which was generated by individual-level data that we reconstructed by scanning the reported Kaplan-Meier curves [9, 14]. The HR appears to fluctuate over time, suggesting a departure from the PH assumption. We conducted a similar numerical study generating 2,000,000 data points based on

Table 1. Hazard ratio estimated by Cox's procedure with eight scenarios

Accrual and follow-up pattern (T_1, T_2)	(a) (0, 40)	(b) (8, 32)	(c) (16, 24)	(d) (32, 8)
PH difference	0.78	0.78	0.78	0.78
Non-PH difference	0.78	0.73	0.67	0.56

$T_1 + T_2$ = total study duration.
Abbreviations: PH, proportional hazards; T_1, accrual period; T_2, additional follow-up period after the end of the accrual.

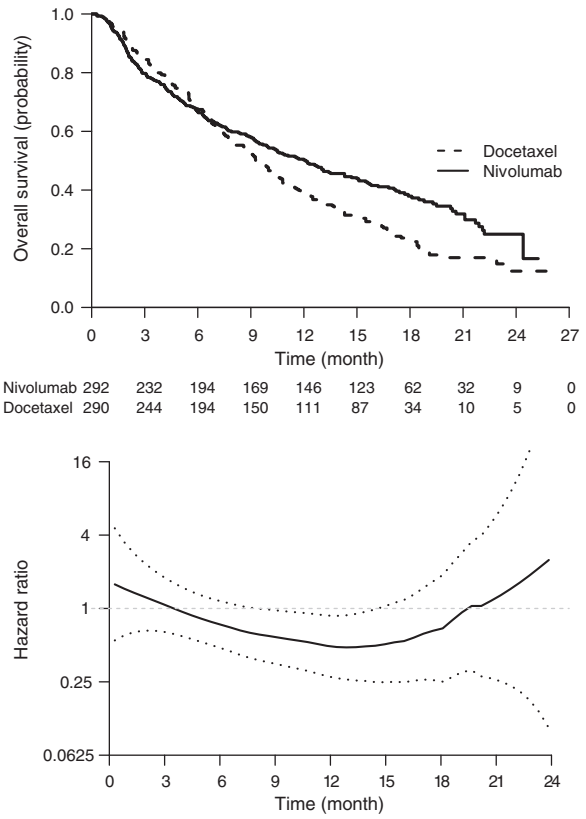

Figure 3. Overall survival by group with reconstructed data of the CheckMate 057 study. Kaplan-Meier curves (top) and estimated time-specific hazard ratio (bottom).

the estimated survival functions presented in Figure 3 (top) (see supplemental online Data B for details) to demonstrate the extent to which the HR estimate could have varied if the accrual period (T_1) or additional follow-up period (T_2) had differed. Figure 4 shows the estimated HR with various combinations of (T_1, T_2). In our numerical study, each scenario has the same total study duration ($T_1 + T_2$) for simplicity. The total study duration ($T_1 + T_2$) could vary in practice, because the study termination is usually not fixed but rather occurs when a prespecified number of events are observed. We found that changes to T_1 or T_2 could have produced a HR anywhere in the range from 0.73 to 0.81 (Figure 4). Note that the HR variation we observed in this numerical study was a function of the (T_1, T_2) combinations we chose, where we fluctuated (T_1, T_2), forcing $T_1 + T_2$ to be a constant. If (T_1, T_2) varied more than we fluctuated, then the HR range could have been even larger.

Discussion

We demonstrated that the HR estimated by Cox's procedure [10] is affected by two study-specific parameters (i.e., accrual pattern and follow-up time) in non-PH scenarios. Other factors that are possibly independent of the treatment difference (e.g., random dropout rate, withdrawal from the study) will also affect the HR estimate. We have examined two non-PH cases commonly seen in cancer clinical trials. One is the pattern in which survival curves are separated during early time points and then glued together later on [7, 8]; another

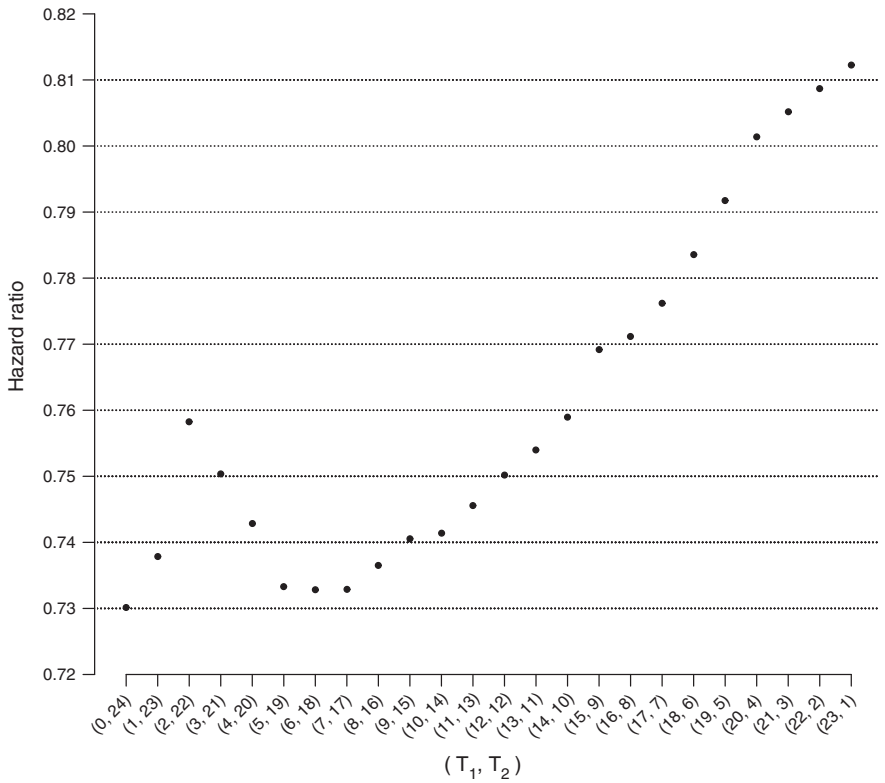

Figure 4. Hazard ratio estimated by Cox's procedure as a summary of the difference between two survival time distributions presented in Figure 3 with various scenarios of accrual period (T_1) and additional follow-up period after the end of accrual (T_2), where $T_1 + T_2$ is the total study duration.

is the pattern of delayed separation, which is typical for immunotherapy trials [9]. The variation in the estimated HR was not negligible for both non-PH patterns.

Although no statistical method can confirm that a between-group difference is indeed a PH, statistical testing can help detect a violation of the PH assumption when an investigator ultimately wants to report the HR. For example, testing based on cumulative sums of martingale residual [15] can be implemented easily using the *ASSESS* statement in SAS's PHREG procedure, *stphtest* command in Stata, *gof* package in R, and so on.

Even when the difference between two groups is a PH, there are other limitations to the HR [1]. For example, it can be challenging to determine whether the observed HR indicates a clinically meaningful benefit (or harm), because there is no reference value from the control group. Given that, careful considerations should be made at the design stage in specifying a primary summary measure. Choosing a statistically robust and clinically interpretable summary measure is critically important to the clinicians who must interpret comparative clinical trials and make treatment recommendations to patients. Unless there is strong justification for using HR as a summary for the treatment effect, it would not be recommended. We recommend using alternative measures that do not rely on any modeling assumptions, such as the difference or ratio of median survival times, *t*-year event rates, and restricted mean survival times [1–4, 6, 16–19].

In summary, the HR estimate from the Cox procedure cannot be interpreted as a simple or meaningful weighted average of the time-specific HR in non-PH cases. Decisions based on HR estimates may be misleading when the PH assumption is violated. Caution should be exercised when investigators report, and readers interpret, HR as the magnitude of the between-group difference.

Acknowledgments

The work was supported by institutional funds of Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute (Dana Funds). These funds were made possible by a Dana Foundation donation.

Disclosures

The authors indicated no financial relationships.

References

1. Uno H, Claggett B, Tian L et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. J Clin Oncol 2014;32:2380–2385.

2. Uno H, Wittes J, Fu H et al. Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies. Ann Intern Med 2015;163:127–134.

3. A'Hern RP. Restricted mean survival time: An obligatory end point for time-to-event analysis in cancer trials? J Clin Oncol 2016;34:3474–3476.

4. Chappell R, Zhu X. Describing differences in survival curves. JAMA Oncol 2016;2:906–907.

5. Péron J, Roy P, Ozenne B et al. The net chance of a longer survival as a patient-oriented measure of treatment benefit in randomized clinical trials. JAMA Oncol 2016;2:901–905.

6. Trinquart L, Jacot J, Conner SC et al. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. J Clin Oncol 2016;34:1813–1819.

7. Paridaens RJ, Dirix LY, Beex LV et al. Phase III study comparing exemestane with tamoxifen as first-line hormonal treatment of metastatic breast cancer in postmenopausal women: The European Organisation for Research and Treatment of Cancer Breast Cancer Cooperative Group. J Clin Oncol 2008;26:4883–4890.

8. Wilke H, Muro K, Van Cutsem E et al. Ramucirumab plus paclitaxel versus placebo plus paclitaxel in patients with previously treated advanced gastric or gastro-oesophageal junction adenocarcinoma (RAINBOW): A double-blind, randomised phase 3 trial. Lancet Oncol 2014;15:1224–1235.

9. Borghaei H, Paz-Ares L, Horn L et al. Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. N Engl J Med 2015;373:1627–1639.

10. Cox DR. Regression models and life-tables. J R Stat Soc Series B Stat Methodol 1972;34:187–220.

11. Ferris RL, Blumenschein GJ, Fayette J et al. Nivolumab for recurrent squamous-cell carcinoma of the head and neck. N Engl J Med 2016;375:1856–1867.

12. Allison PD. Survival Analysis Using SAS: A Practical Guide. Cary, NC: SAS Institute; 2010.

13. Gilbert PB, Wei LJ, Kosorok MR, et al. Simultaneous inferences on the contrast of two hazard functions with censored observations. Biometrics 2002;58:773–780.

14. Guyot P, Ades AE, Ouwens MJ et al. Enhanced secondary analysis of survival data: Reconstructing the data from published Kaplan-Meier survival curves. BMC Med Res Methodol 2012;12:9.

15. Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika 1993;80:557–572.

16. Royston P, Parmar MK. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. Stat Med 2011;30:2409–2421.

17. Royston P, Parmar MK. Restricted mean survival time: An alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med Res Methodol 2013;13:152.

18. Tian L, Fu H, Ruberg SJ et al. Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations. Biometrics 2018;74:694–702.

19. Zhao L, Claggett B, Tian L et al. On the restricted mean survival time curve in survival analysis. Biometrics 2016;72:215–221.

See http://www.TheOncologist.com for supplemental material available online.