Circulation

PERSPECTIVE

Using the Restricted Mean Survival Time Difference as an Alternative to the Hazard Ratio for Analyzing Clinical Cardiovascular Studies

Zachary R. McCaw, PhD
Guosheng Yin, PhD
Lee-Jen Wei, PhD

Cardiovascular trials often use the time to a clinical event as the primary end point when evaluating a new treatment, versus control, via clinical and statistical significance criteria. For the past 50 years, the hazard ratio (HR) has been routinely used for quantifying the treatment effect. However, it is difficult to interpret clinical significance using a ratio measure, such as HR, when there is no reference hazard available from the control arm. Moreover, valid HR analysis requires the proportional hazards (PH) assumption: that the ratio of hazard curves is constant over time. This assumption is hardly plausible in practice. When PH is not met, HR may lack statistical power to detect a true treatment effect. Furthermore, without PH, the estimated HR is not a simple average of HRs over time, and is even more difficult to interpret.[1,2] In this article, we discuss the advantages of an alternative analytical procedure based on the restricted mean survival time (RMST)[1,2] via 3 examples.

The first example comes from the CHARM-Overall (Candesartan in Heart Failure–Assessment of Mortality and Morbidity–Overall) trial, which evaluated the effect of candesartan on all-cause mortality in patients with chronic heart failure.[3] The HR (candesartan vs placebo) of 0.91 (95% CI, 0.83–1.00; *P*=0.055) did not provide strong statistical evidence of a benefit from candesartan. Moreover, it is unclear whether a HR of 0.91 is clinically significant. This does not suggest that candesartan reduced the risk of mortality by 9% versus placebo, because the hazard is not a probability measure like risk. To explore whether HR analysis was appropriate, we reconstructed patient-level survival data from the Kaplan–Meier curves presented in the CHARM-Overall article.[3] Here, the reconstructed Kaplan–Meier curves are presented in Figure, A. Because the Kaplan–Meier curves separate initially but remain parallel after 0.5 years, the PH assumption was not met, as was confirmed via the Schoenfeld goodness-of-fit test (*P*=0.0005).

An alternative is to use the difference in RMSTs to quantify the treatment effect. RMST is the average time-to-event over a fixed follow-up period (for example, 3.5-years in Figure, A). Graphically, RMST corresponds to the area under the survival curve. The higher the curve, the greater the RMST. As presented in Figure, B, the area under the curve for candesartan was 3.07 years. That is, across 3.5 years of follow-up, patients treated with candesartan survived for 3.07 years on average. The corresponding RMST for placebo was 3.00 years. The difference of 0.07 years (95% CI, 0.03–0.11; *P*=0.0016) was highly statistically significant in favor of candesartan. This example demonstrates that RMST can be more statistically efficient than HR when PH is not met. After passing the statistical significance hurdle, the clinical utility of candesartan can be evaluated using individual RMSTs (3.07 and 3.00 years). For the present case, it is debatable whether the 0.07-year (3.6-week) survival gain is clinically significant. A limitation of RMST-based analysis is the need to specify a truncation time. Ideally, this would be prespecified via clinical considerations during study design. Otherwise, one can choose the last observed event or

Key Words: coronary artery disease ■ follow-up studies ■ heart failure ■ probability

https://www.ahajournals.org/journal/circ

Circulation. 2019;140:1366–1368. DOI: 10.1161/CIRCULATIONAHA.119.040680

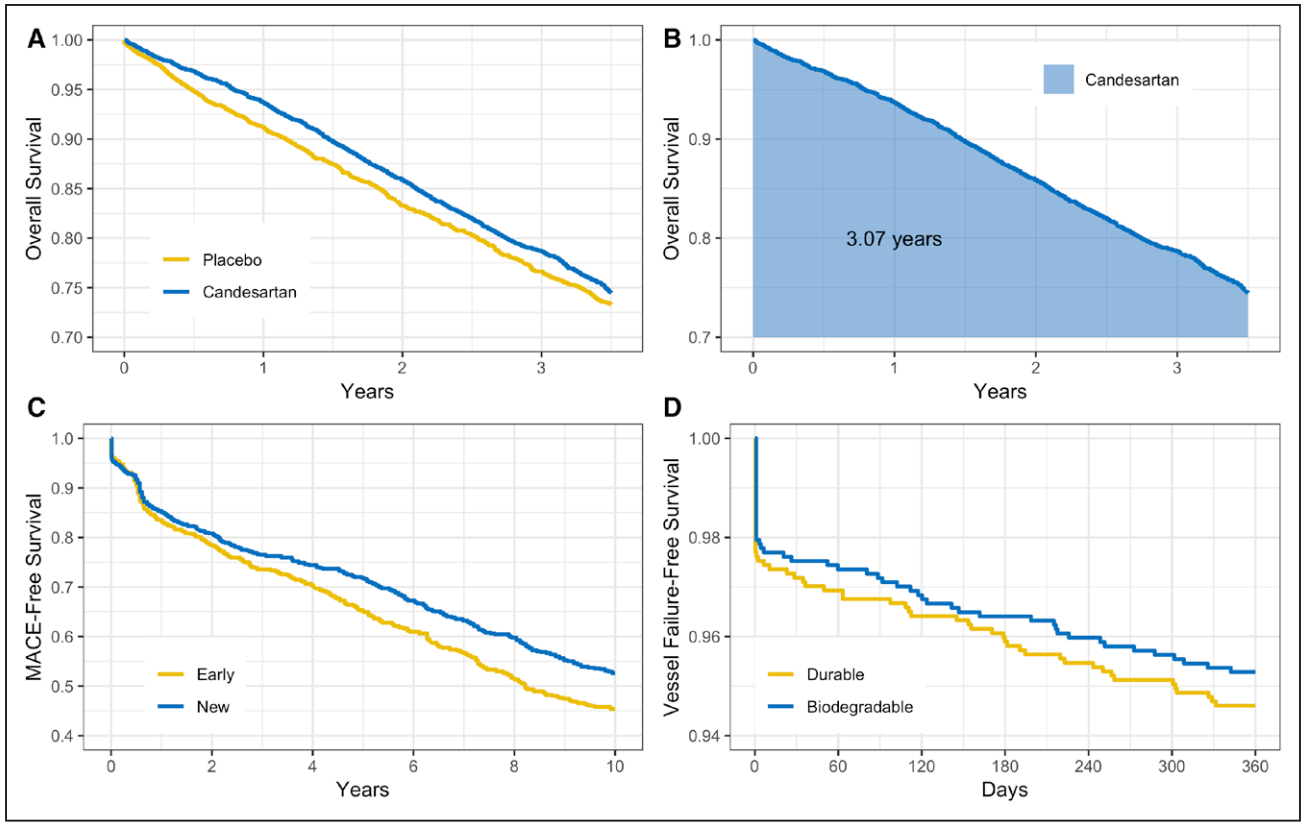

Figure. Reconstructed Kaplan–Meier curves.
A, Overall survival comparing candesartan with placebo from the study by Pfeffer and colleagues.[3] **B**, RMST as the area under the Kaplan–Meier curve up to 3.5 years of follow-up for patients receiving candesartan. **C**, MACE-free survival comparing new and early generation stents from the study by Kufner and colleagues.[4] **D**, Survival without target vessel failure comparing biodegradable and durable stents from the study by von Birgelen and colleagues.[5] MACE indicates major adverse cardiac event.

censoring time, which here was at 3.5 years. With this choice, RMST incorporates all available information.

For the second example, we show that if HR identifies a statistically significant treatment effect, then so too does RMST. Consider a recent study[4] comparing new and early generation sirolimus-eluting stents among patients with coronary artery disease. The primary outcome was time to a major adverse cardiac event. The HR was 0.82 (95% CI, 0.71-0.93), which significantly favored new generation stents. Using the reconstructed time-to-event data in the Figure (C), the 10-year RMSTs were 7.07 years and 6.57 years for new and early generation stents, respectively. The difference of 0.50 years (95% CI, 0.13–0.86; *P*=0.01) also significantly favored new generation stents. In general, RMST procedures do not result in any power loss versus HR, and beyond the statistical evaluation, provide valuable insight on the clinical utility of the new stents.

For the third example, we demonstrate the utility of RMST for noninferiority studies. It is known that the HR's precision depends on the number of observed events, but not on patients' exposure times. Consequently, for noninferiority trials where the event rate is low, HR-based designs may require large numbers of patients. Consider a recent noninferiority study[5] comparing 3 types of stents with time to target vessel failure as the primary end point. For evaluating biodegradable everolimus-eluting stents versus durable zotarolimus-eluting stents (n=2345), the HR (biodegradable vs durable) was 0.87 (95% CI, 0.61–1.25). The CI upper bound of 1.25 suggests that biodegradable stents may actually increase the hazard of target vessel failure by 25%. Using the reconstructed Kaplan–Meier curves in the Figure (D), the 12-month RMSTs were 11.41 and 11.36 months for biodegradable and durable stents, respectively. The difference was 1.80 days (95% CI, –3.40 to 6.99). Thus, in the worst case, biodegradable stents may shorten the time-to-failure by 3.40 days. This time-scale quantification is much easier to interpret than HR. For this noninferiority study, the important consideration is whether 12-month follow-up is sufficiently long to evaluate the performance of the stents. In contrast to evaluating superiority, the number of observed events is not crucial for assessing noninferiority.[2] To show that RMST may allow one to reduce the study size without losing much precision, we repeatedly drew random subsets with 50% of the data (n=1173) and calculated CIs for the HR and RMST difference. With 100 random subsamples, the average upper bound for the HR increased to 1.49, which seems much higher than 1.25.

Thus, a smaller size study may not be justifiable when using HR. On the other hand, with the reduced study size, biodegradable stents would in the worst case shorten the time-to-failure by 5.66 days, which is only slightly higher than 3.40 days. This result suggests that a smaller study may have been justified for assessing noninferiority. When designing a noninferiority study with an appropriate, prespecified exposure time, using RMST to set the noninferiority margin is efficient and heuristically easy to justify.

In conclusion, RMST is a powerful, robust, and interpretable tool for the design and analysis of clinical studies. All analyses presented herein can be implemented via the survRM2 package in R, or the RMSTREG proc in SAS.

ARTICLE INFORMATION

Correspondence

Lee-Jen Wei, PhD, 655 Huntington Ave, Boston, MA 02115. Email wei@hsph.harvard.edu

Affiliations

Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA (Z.R.M., L.J.W.). Department of Statistics and Actuarial Science, The University of Hong Kong, China (G.Y.).

Disclosures

None.

REFERENCES

1. Kim DH, Uno H, Wei LJ. Restricted mean survival time as a measure to interpret clinical trial results. *JAMA Cardiol.* 2017;2:1179–1180. doi: 10.1001/jamacardio.2017.2922
2. Uno H, Wittes J, Fu H, Solomon SD, Claggett B, Tian L, Cai T, Pfeffer MA, Evans SR, Wei LJ. Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies. *Ann Intern Med.* 2015;163:127–134. doi: 10.7326/M14-1741
3. Pfeffer MA, Swedberg K, Granger CB, Held P, McMurray JJ, Michelson EL, Olofsson B, Ostergren J, Yusuf S, Pocock S; CHARM Investigators and Committees. Effects of candesartan on mortality and morbidity in patients with chronic heart failure: the CHARM-Overall programme. *Lancet.* 2003;362:759–766. doi: 10.1016/s0140-6736(03)14282-1
4. Kufner S, Joner M, Thannheimer A, Hoppmann P, Ibrahim T, Mayer K, Cassese S, Laugwitz KL, Schunkert H, Kastrati A, et al; ISAR-TEST 4 (Intracoronary Stenting and Angiographic Results: Test Efficacy of 3 Limus-Eluting Stents) Investigators. Ten-year clinical outcomes from a trial of three limus-eluting stents with different polymer coatings in patients with coronary artery disease. *Circulation.* 2019;139:325–333. doi: 10.1161/CIRCULATIONAHA.118.038065
5. von Birgelen C, Kok MM, van der Heijden LC, Danse PW, Schotborgh CE, Scholte M, Gin RMTJ, Somi S, van Houwelingen KG, Stoel MG, et al. Very thin strut biodegradable polymer everolimus-eluting and sirolimus-eluting stents versus durable polymer zotarolimus-eluting stents in allcomers with coronary artery disease (BIO-RESORT): a three-arm, randomised, noninferiority trial. *Lancet.* 2016;388:2607–2617. doi: 10.1016/S0140-6736(16)31920-1