OXFORD

JNCI Cancer Spectrum (2019) 3(4): pkz058

doi: 10.1093/jncics/pkz058
First published online August 1, 2019
Commentary

COMMENTARY

Assessing Clinical Equivalence in Oncology Biosimilar Trials With Time-to-Event Outcomes

Hajime Uno, Deborah Schrag, Dae Hyun Kim, Dejun Tang, Lu Tian, Hope S. Rugo, Lee-Jen Wei

See the Notes section for the full list of authors' affiliations.
Correspondence to: Lee-Jen Wei, PhD, Department of Biostatistics, Harvard University, 655 Huntington Ave, Boston, MA 02115 (e-mail: wei@hsph.harvard.edu).

Abstract

A typical biosimilar study in oncology uses the overall response evaluated at a specific time point as the primary endpoint, which is generally acceptable regulatorily, to assess clinical equivalence between a biosimilar and its reference product. The standard primary endpoint for evaluating an anticancer therapy, progression-free or overall survival would be a secondary endpoint in a biosimilar trial. With a conventional analytic procedure via, for example, hazard ratio to quantify the group difference, it is difficult and challenging to assess clinical equivalence with respect to progression-free or overall survival because the study generally has a limited number of clinical events observed in the study. In this article, we show that an alternative procedure based on the restricted mean survival time, which has been discussed extensively for design and analysis of a general equivalence study, is readily applicable to a biosimilar trial. Unlike the hazard ratio, this procedure provides a clinically interpretable estimate for assessing equivalence. Using the restricted mean survival time as a summary measure of the survival curve will enhance better treatment decision making in adopting a biosimilar product over the reference product.

To evaluate whether a biosimilar product is “equivalent” to its reference product, an extensive assessment of analytic and functional characteristics of the products are important (1). In addition, studies for evaluating clinical equivalence are also essential. To expedite regulatory approval, a biosimilar oncology study generally uses the overall response rate (ORR) at a specific time point as the primary efficacy measure, rather than conventional progression-free survival (PFS) or overall survival (OS) used for approval of the reference product (1,2). A core tenet of this approach is that the ORR is a valid surrogate for PFS and/or OS. It is unclear, however, whether demonstration of “equivalent ORR” suffices to establish clinical equivalence with respect to PFS and/or OS (3). For example, HERITAGE, a multicenter, double-blind, randomized, phase III equivalence study, was conducted to assess whether a trastuzumab biosimilar plus a taxane was clinically equivalent to trastuzumab plus a taxane in patients with metastatic breast cancer. A total of 500 patients were enrolled from December 2012 to August 2015 and randomly assigned to either a biosimilar group or trastuzumab group. Forty-two patients were excluded from the primary intention-to-treat analysis population, leaving 458 patients for the final analysis (230 in the biosimilar group and 228 in the trastuzumab group). The primary endpoint was 24-week ORR with prespecified equivalence bounds. Specifically, it was specified that equivalence would be claimed if the two-sided 90% confidence interval (CI) for the ratio of the 24-week ORRs was completely within the range of 0.81 to 1.24. For the biosimilar study, it is required to show that the biosimilar product is not inferior or superior to the reference product. The ORR was 69.6% in the biosimilar group and 64.0% in the trastuzumab group (biosimilar vs trastuzumab: ORR = 1.09, 90% CI = 0.974 to 1.211), which supported the equivalence claim based on the above prespecified bounds.

Secondary endpoints for HERITAGE included PFS and OS times during a 48-week follow-up. For the biosimilar and trastuzumab arms, PFS had 145.7 and 141.6 person-years of follow-up, respectively. The corresponding observed number of progression or death events was 102 for each group (biosimilar

Received: May 7, 2019; **Revised:** July 5, 2019; **Accepted:** July 29, 2019

vs trastuzumab: hazard ratio [HR] = 0.97, 95% CI = 0.74 to 1.28). For OS, the person-years of follow-up were 188.3 in biosimilar and 181.8 trastuzumab. The numbers of deaths in the biosimilar and trastuzumab groups were 25 and 34 deaths, respectively (biosimilar vs trastuzumab: HR = 0.67, 95% CI = 0.40 to 1.13) (2). These wide confidence intervals may be interpreted as insufficient information regarding PFS and OS. It seems difficult to claim these two groups were similar with such a large range of possible hazard ratio values for the PFS or OS endpoint. For PFS, the hazard from the biosimilar group might be 28% higher than that from the reference product. Consequently, even if the biosimilar product is approved by the regulatory agencies, clinicians and patients may not be convinced to adopt the biosimilar as an alternative to the reference product. Appropriate and efficient procedures for assessing clinical equivalence for PFS and OS are needed for expeditious evaluation and assurance of a biosimilarity claim with clinically meaningful interpretations.

The HERITAGE investigators also reported difference in 48-week event rates for both treatment groups with PFS and OS endpoints. At 48 weeks, the PFS rate was 44.3% in the biosimilar group vs 44.7% in the trastuzumab group. The difference was −0.4% (95% CI = −9.4% to 8.7%). For OS, it was 89.1% in the biosimilar group vs 85.1% in the trastuzumab group. The difference was 4.0% (95% CI = −2.1% to 10.3%). These confidence intervals also appear to be quite large. Moreover, such an event rate difference at a specific time point does not include the information about the temporal profile of PFS or OS before week 48.

In this paper we consider a general comparative clinical trial with the time to a specific event as the endpoint to evaluate whether a new treatment is equivalent to the standard care. The biosimilar trial is a special case under this general setting. The standard analytic procedure to make an inference about the treatment effect is to use the hazard ratio under the proportional hazards assumption that the hazard functions between the two arms are proportional to each other over the entire study duration (4). The estimation procedure for the hazard ratio elegantly proposed by Cox has been used routinely since the 1970s. On the other hand, there are several issues and concerns with using this procedure that have been extensively discussed in the statistical and medical literature (5–8). For instance, the hazard ratio by itself without a reference hazard value from the control arm is difficult to interpret. A hazard ratio of, for example, 0.67, for OS (biosimilar vs reference product) in HERITAGE may not be a clinically meaningful improvement if the hazard from the control is low. Moreover, a 33% hazard reduction should not be interpreted as the risk reduction because the hazard is not a probability measure, such as risk. The hazard was referred to as the “force of mortality.” When the proportional hazards model assumption is not met, the resulting hazard ratio estimate is not a simple average of hazard ratios over time. In fact, this empirical hazard ratio estimates a population quantity, which depends on underlying study-specific censoring time distributions. That is, if we conducted two studies under the same setting, but with different follow-up time distributions, then the resulting hazard ratios can be quite different beyond sampling variation from study to study (8). This is undesirable because we are interested only in the underlying distributions of the time-to-event observations. Another issue of using hazard ratio is that it tends to require a large sample size or longer follow-up time to observe a large number of events to obtain a desirable precision for the hazard ratio estimate. However, when we deal with an equivalence or noninferiority study, the number of events would not play as important a role for a superiority study. In this paper, we present an alternative approach evaluating whether the treatment is equivalent or noninferior to the control without the need for a large number of events in the study, provided that the patients’ exposure times are long enough from clinical considerations. The usual event-driven equivalence or noninferiority study via hazard ratio may require an unnecessarily larger sample size and result in a waste of valuable resources.

Methods

An alternative approach to hazard ratio is to use t-year mean survival time (t-MST) or the restricted mean survival time (RMST)-based statistics to design and analyze time-to-event data (5–7,9–13). The summary measures based on RMST can be interpreted heuristically and do not need any model assumption to quantify the treatment effect. More important, because the inference based on RMST-related statistics takes the patients’ exposure times into consideration, the study size can be substantially reduced for an equivalence trial. In this paper, we use PFS and OS data from the HERITAGE trial to illustrate how to apply this procedure to assess whether a biosimilar product is acceptable compared with a reference product with respect to clinical endpoints. With this alternative tool, one may consider PFS or OS as the primary endpoint for a biosimilar oncology study instead of using ORR.

Results

Using a Single RMST-based Statistic for Assessment of Equivalence

We use data from the HERITAGE trial to illustrate RMST procedures. To this end, we reconstructed PFS and OS patient-level data from the US Food and Drug Administration briefing document for HERITAGE (14,15). The corresponding Kaplan-Meier curves up to week 48 for OS and PFS are presented in Figure 1, A and B . The RMST or t-MST is the average OS time within t-year of follow-up. That is, a survival time beyond t-year would be truncated at year t in calculating RMST. The RMST is the area under the survival curve over this time period: The higher the curve, the greater the area. Thus, the RMST is heuristically interpretable and would be appreciated by clinicians and patients as a summary of the survival profile. Figure 2 shows the estimated RMST (ie, the area under the Kaplan-Meier curve) of OS (Figure 2A) and PFS (Figure 2B) in the biosimilar group. The 48-week RMST of OS for the biosimilar was 45.9 weeks. That is, patients would survive, on average, 45.9 weeks of the 48-week scheduled follow-up. The counterpart from the trastuzumab group was 45.2 weeks. The differences in RMSTs between the biosimilar and trastuzumab can then be used to evaluate the treatment difference for superiority, noninferiority, or equivalence studies (6). The RMST difference for OS (biosimilar minus trastuzumab) was 0.7 weeks (95% CI = −0.7 to 2.1 weeks). That is, the largest possible difference in RMST for OS between the two arms would be 2.1 weeks (which is less than 5% of 45.2 weeks from the reference product). The RMST difference for PFS (38.8 vs 37.5 weeks) was 1.3 weeks (95% CI = −1.3 to 3.8 weeks). The largest plausible difference of 3.8 weeks is only 10.1% of 37.5 weeks from the reference group. Unlike the results from the hazard ratio, these time-scaled small differences in OS and PFS, coupled with the reference RMST values from the trastuzumab group, provide a clinically meaningful way to evaluate whether the biosimilar is clinically equivalent to trastuzumab

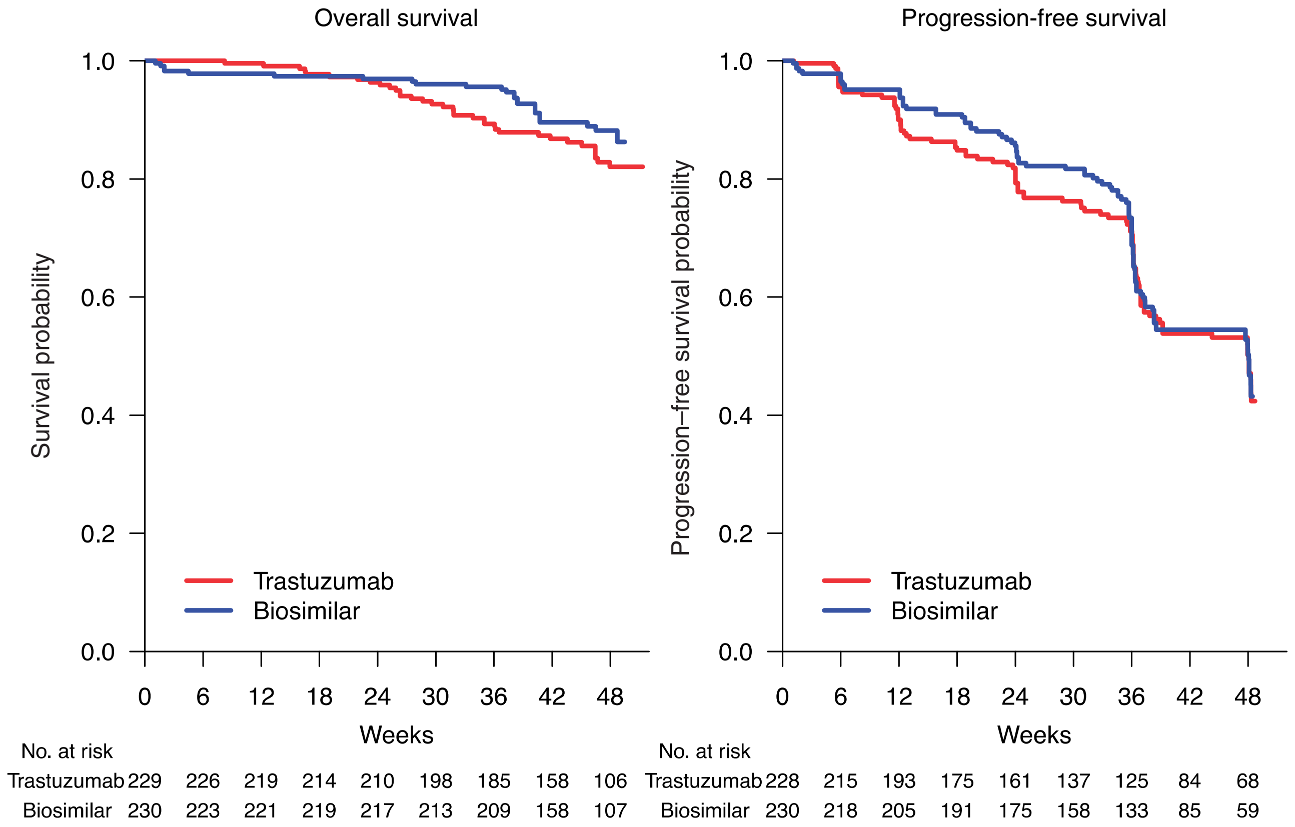

Figure 1. The Kaplan-Meier curves based on the reconstructed data for overall survival and progression-free survival for patients taking trastuzumab vs biosimilar.

over a 48-week time window. The RMST-based analysis can be implemented via standard statistical software such as RMSTREG procedure (SAS version 9.4; SAS Corporation, Cary, NC), Stata (Stata Corp LLC, College Station, TX), and survRM2 package (R version 3.5.2, The R Foundation, Vienna, Austria). Furthermore, designing a trial with the RMST-based approach can also be implemented by R (SSRMST package).

Simultaneously Evaluating RMST Differences Over Time

Using the above single measure RMST evaluated at week 48 to summarize the survival curve may not be sufficient to assess the equivalence between the two arms. It is possible that the survival profiles from the two arms are quite different over the study period, but the two RMSTs at week 48 are similar. To address this issue, one may consider the RMST or the t-MST as a function of time t. That is, for each time point t, we calculate the RMST up to t. We then evaluate whether the t-MST curves between the two arms are equivalent. Such a generalization of RMST was studied by Zhao et al. (16). Specifically, they considered comparing t-MSTs over a given time window simultaneously. Figure 3 shows RMST curves up to 48 weeks by treatment group, and Figure 4 shows the difference between two RMST curves and a 95% simultaneous confidence (equal precision) band from 1.1 to 48 weeks for OS (Figure 4A) and PFS (Figure 4B). With a high probability (ie, 95%), the true curve of the underlying RMST difference between the two treatment groups would be within the simultaneous confidence band. For example, at 12, 24, 36, and 48 weeks, the true differences of two RMST curves for OS are likely to be within the intervals (−0.4, 0.4), (−0.9, 0.3), (−1.0, 1.3), and (−1.2, 2.5) weeks, simultaneously. For PFS, the true differences at those time points would be likely within the intervals (−0.4, 0.4), (−0.6, 1.8), (−1.1, 3.5), and (−2.1, 4.6) weeks, respectively. Because each of these intervals is for a difference in RMST, it can easily be interpreted clinically. The simultaneous confidence band for RMST difference would be a useful addition when we address the equivalence or noninferiority question about time-to-event outcomes. The R code implementing this method is available with the paper (16) at the Biometrics website on Wiley Online Library (https://onlinelibrary.wiley.com/doi/10.1111/biom.12384).

Discussion

Expediting regulatory approval for biosimilars has the potential to make lifesaving treatments globally available at lower costs (17). However, appropriate choices of clinical endpoints and efficient quantitative procedures for evaluating clinical equivalence are essential to ensure that physicians, patients, and the public have confidence in the biosimilar products. Advantages of RMST-based methods over hazard ratio–based methods have been discussed extensively in general equivalence or noninferiority studies (6,18–21). Unfortunately, biosimilar trials of the oncology products have not taken advantage of these efficient procedures yet. The confidence intervals of hazard ratios, as shown in HERITAGE, may not be narrow enough to conclude clinical equivalence. Such insufficient precision of hazard ratios depends on the total number of observed events instead of patient exposure times (6). To obtain a desirable estimation precision for clinical equivalence claim using hazard ratio via an

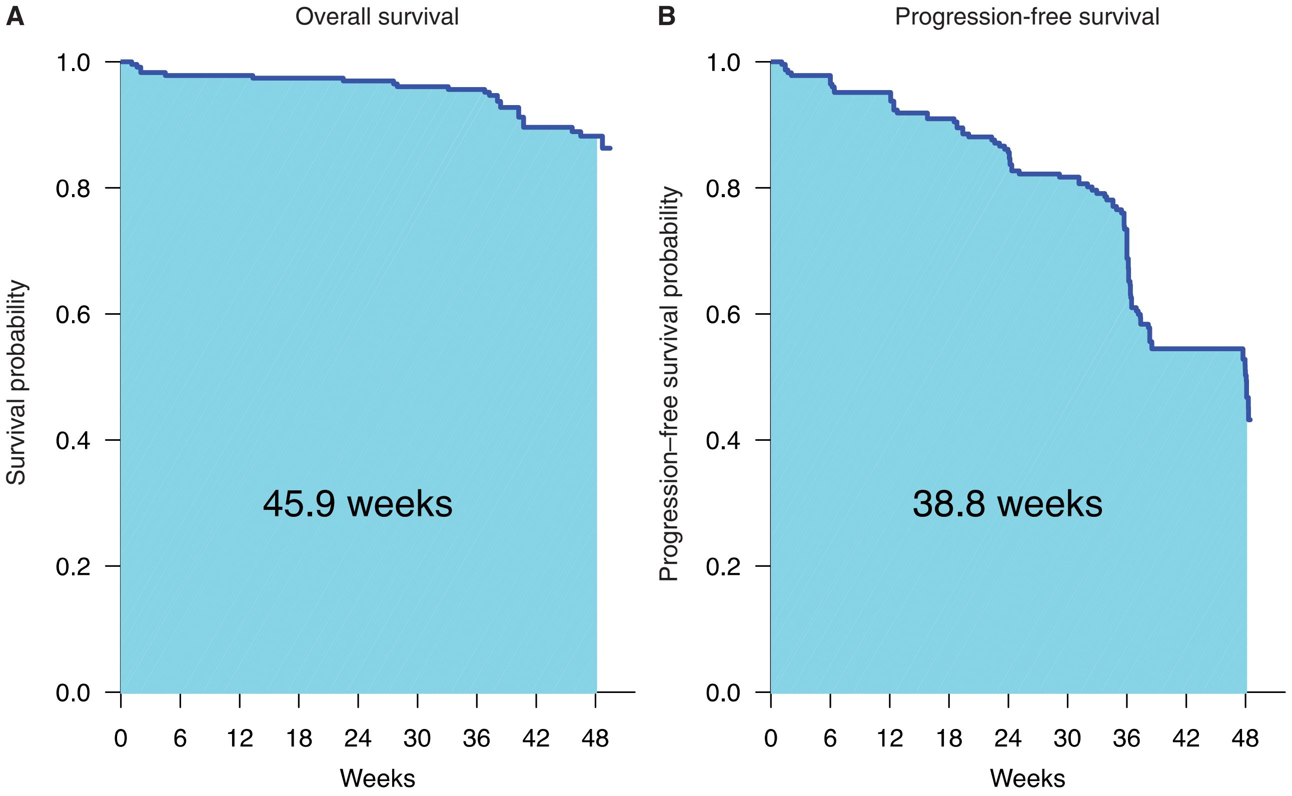

Figure 2. Estimated restricted mean survival time for patients taking biosimilar with 48-week follow-up. **A**) Overall survival. **B**) Progression-free survival.

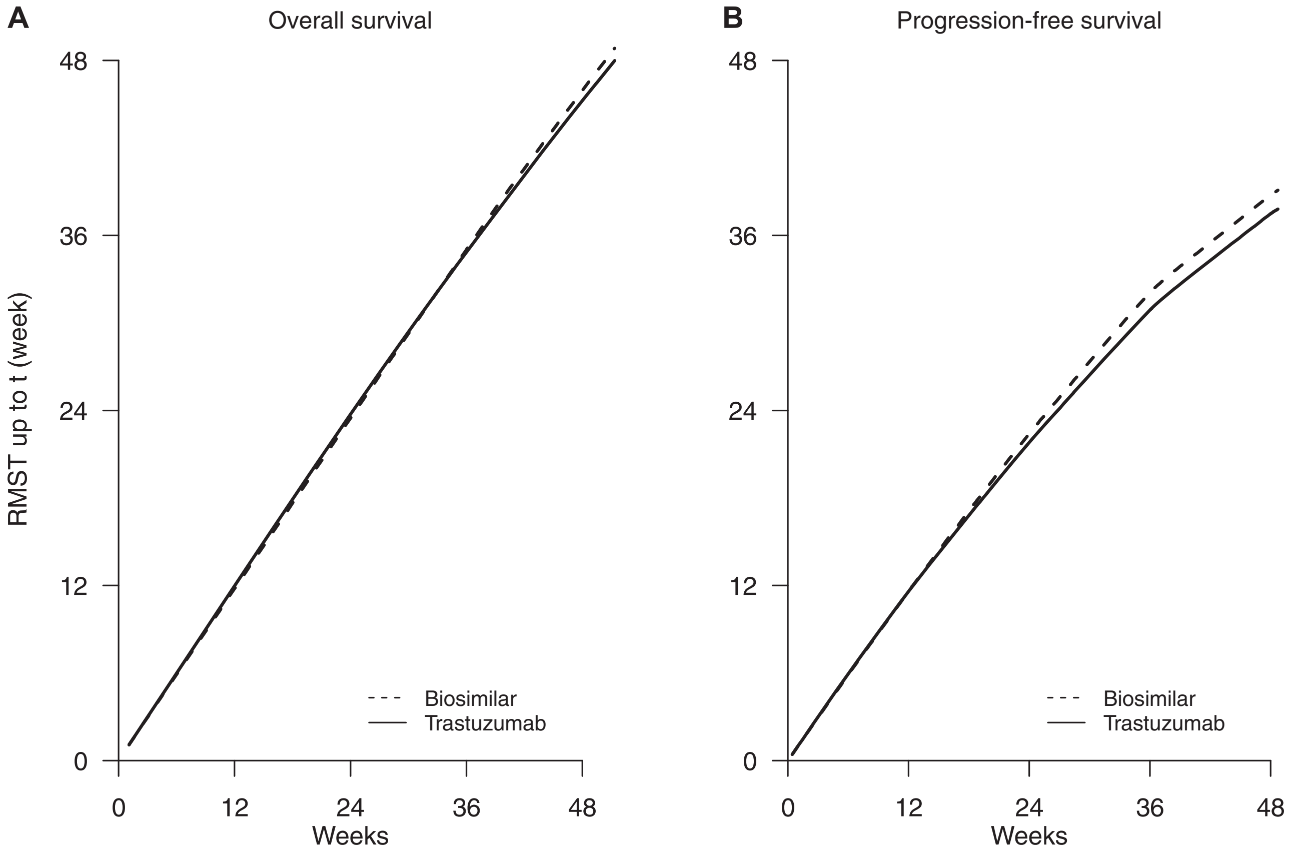

Figure 3. Restricted mean survival time (RMST) curves up to 48 weeks for biosimilar group and trastuzumab group. **A**) Overall survival. **B**) Progression-free survival.

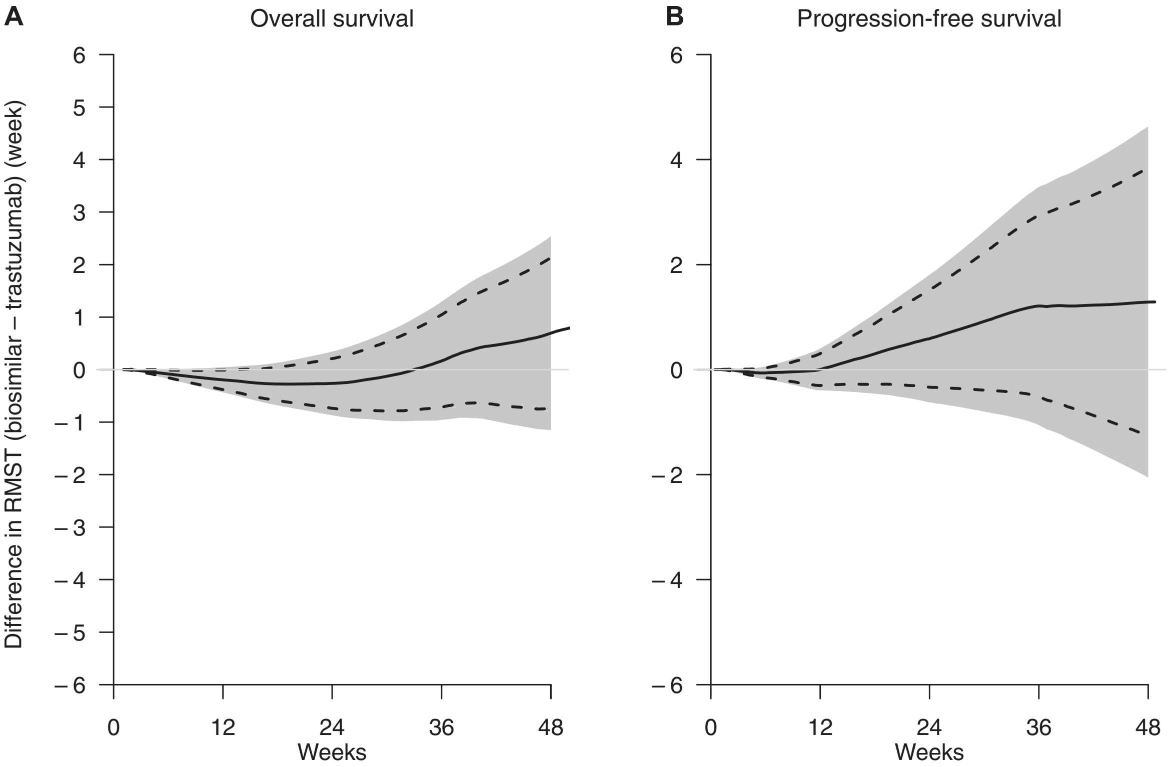

Figure 4. Difference in restricted mean survival time (RMST) curves up to 48 weeks between biosimilar group and trastuzumab group. **A**) Overall survival. **B**) Progression-free survival. **Solid line** = point estimate; **dashed line** = 95% pointwise confidence interval; **shaded area** = 95% simultaneous confidence interval.

event-driven design setting, an impractically large sample size would be needed (eg, 1607 events are needed for an upper equivalence boundary for OS/PFS of 1.15 for HR). Moreover, it is difficult to choose hazard ratio–based equivalence boundaries because of the lack of clinical interpretability of the hazard ratio (6). In contrast, RMST-based methods provide a clinically interpretable between-group difference with feasible study size and study duration compared with the hazard ratio–based approach.

For this equivalence study, the important consideration is whether 48-week follow-up is long enough to evaluate the performance of the biosimilar. To show that RMST may allow us to reduce the study size without losing much precision, we repeatedly drew random subsets with 50% of the OS data (n = 229) and calculated confidence intervals for the hazard ratio and RMST difference. With 500 random subsamples, the average upper bound for the hazard ratio increased to 1.43, which seems much higher than 1.13. Therefore, a smaller study would not be justifiable when using hazard ratio. On the other hand, with the reduced study size, the biosimilar, in the worst case, might shorten the OS time by 1.3 weeks, which is only slightly worse than 0.7 weeks. This suggests a smaller study may be justifiable for assessing equivalence over 48 weeks. When designing an equivalence study with an appropriate, prespecified exposure time, using RMST to set the noninferiority margin is efficient and heuristically easy to justify. The RMST-based method is one of the quantitative procedures that are well suited to a biosimilar oncology trial with PFS or OS as a study endpoint even under the current biosimilar trial setting with the ORR as the primary endpoint.

Naturally, for most current biosimilar trials, one may argue that the patients' exposure times might be too short to assess PFS and OS and long-term safety. Development of postmarket evidence is also an essential component for biosimilar products (22). We may need to follow the study participants longer to collect more information on PFS and OS even after a biosimilar product is approved regulatorily to convince clinical practitioners and patients that the biosimilar is equally safe and efficacious to the reference product. The RMST-based method is also useful to assess the postmarketing data regarding PFS and OS efficiently.

Notes

Affiliations of authors: Division of Population Sciences, Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA (HU, DS); Marcus Institute for Aging Research, Hebrew Senior Life, Boston, MA (DHK); Biostatistics & Clinical Submission Management, Sandoz Biopharmaceuticals, Holzkirchen, Germany (DT); Department of Biomedical Data Science, Stanford University School of Medicine, Palo Alto, CA (LT); Helen Diller Family Comprehensive Cancer Center, University of California-San Francisco, San Francisco, CA (HSR); Department of Biostatistics, Harvard University, Boston, MA (L-JW).

Hope S. Rugo and Lee-Jen Wei contributed equally as senior authors.

Dejun Tang is an employee of Novartis. The other authors have declared no conflicts of interest.

References

1. US Department of Health and Human Services; US Food and Drug Administration; Center for Drug Evaluation and Research; Center for Biologics Evaluation and Research. Scientific considerations in

demonstrating biosimilarity to a reference product. 2015. https://www.fda.gov/media/82647/download. Accessed August 5, 2019.

2. Rugo HS, Barve A, Waller CF, et al. Effect of a proposed trastuzumab biosimilar compared with trastuzumab on overall response rate in patients with ERBB2 (HER2)-positive metastatic breast cancer: a randomized clinical trial. *JAMA*. 2017;317(1):37–47.
3. Schleicher SM, Seidman AD. An important step forward for biosimilars in cancer treatment. *JAMA Oncol*. 2017;3(7):989–990.
4. Cox DR. Regression models and life-tables. *J R Stat Soc Series B Methodol*. 1972; 34(2):187–202.
5. Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol*. 2014;32(22): 2380–2385.
6. 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(2):127–134.
7. Chappell R, Zhu X. Describing differences in survival curves. *JAMA Oncol*. 2016;2(7):906–907.
8. Horiguchi M, Hassett MJ, Uno H. How do the accrual pattern and follow-up duration affect the hazard ratio estimate when the proportional hazards assumption is violated? *Oncologist*. 2019;24(7):867–871.
9. Royston P, Parmar M. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol*. 2013;13(1):152.
10. Royston P, Parmar M. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat Med*. 2011;30(19):2409–2421.
11. A'Hern RP. Restricted mean survival time: an obligatory end point for time-to-event analysis in cancer trials? *J Clin Oncol*. 2016;34(28):3474–3476.
12. Péron J, Roy P, Ozenne B, Roche L, Buyse M. The net chance of a longer survival as a patient-oriented measure of treatment benefit in randomized clinical trials. *JAMA Oncol*. 2016;2(7):901–905.
13. Liang F, Zhang S, Wang Q, Li W. Treatment effects measured by restricted mean survival time in trials of immune checkpoint inhibitors for cancer. *Ann Oncol*. 2018;29(5):1320–1324.
14. US Food and Drug Administration. Oncologic Drugs Advisory Committee meeting: BLA 761074 MYL-1401O, a proposed biosimilar to herceptin (trastuzumab). 2017: 40. https://www.fda.gov/media/106566/download. Accessed August 5, 2019.
15. Guyot P, Ades AE, Ouwens M, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2012;12(1):9.
16. Zhao L, Claggett B, Tian L, et al. On the restricted mean survival time curve in survival analysis. *Biometrics*. 2016;72(1):215–221.
17. Burstein HJ, Schrag D. Biosimilar therapy for ERBB2 (HER2)-positive breast cancer: close enough? *JAMA*. 2017;317(1):30–32.
18. Hasegawa T, Uno H, Wei L-J. How to summarize the safety profile of epoetin alfa versus best standard of care in anemic patients with metastatic breast cancer receiving standard chemotherapy? *J Clin Oncol*. 2016;34(31):3818.
19. Hasegawa T, Uno H, Wei L-J. Safety study of salmeterol in asthma in adults. *N Engl J Med*. 2016;375(11):1097.
20. Uno H, Hassett MJ, Wei L-J. Axillary vs sentinel lymph node dissection in women with invasive breast cancer. *JAMA*. 2018;319(3):306.
21. Horiguchi M, Uno H, Wei L-J. Evaluating noninferiority with clinically interpretable statistics for the PROSELICA study to assess treatment efficacy of a reduced dose of cabazitaxel for treating metastatic prostate cancer. *J Clin Oncol*. 2018;36(8):825–826.
22. Lyman GH, Balaban E, Diaz M, et al. American Society of Clinical Oncology statement: biosimilars in oncology. *J Clin Oncol*. 2018;36(12):1260–1265.