Clinical Infectious Diseases

BRIEF REPORT

Survival Analysis of Treatment Efficacy in Comparative Coronavirus Disease 2019 Studies

Zachary R. McCaw,[1] Lu Tian,[2] Dae Hyun Kim,[3] A. Russell Localio,[4] and Lee-Jen Wei[5]

[1]Google, Mountain View, California, USA, [2]Department of Biomedical Data Science, Stanford University, Stanford, California, USA, [3]Hinda and Arthur Marcus Institute for Aging Research, Hebrew SeniorLife, Harvard Medical School, Boston, Massachusetts, USA, [4]Division of Biostatistics, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA, and [5]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA

For survival analysis in comparative coronavirus disease 2019 trials, the routinely used hazard ratio may not provide a meaningful summary of the treatment effect. The mean survival time difference/ratio is an intuitive, assumption-free alternative. However, for short-term studies, landmark mortality rate differences/ratios are more clinically relevant and should be formally analyzed and reported.

Keywords. mean survival time; hazard ratio; survival rate; critical care; dexamethasone.

In comparing a new therapy with standard care with respect to patient survival, 2 endpoints are routinely considered. The first is a binary indicator of whether the patient has survived across a specific time window, such as 28 days, and is utilized to estimate, for example, the difference in 28-day mortality rates. The second, the observed survival time, can be used to quantify how much the new treatment is expected to prolong a patient's survival across the 28 days of follow-up. These 2 approaches address different questions, and their statistical and/or clinical implications regarding the treatment effect may not be exchangeable. For short-term studies in critical care medicine, analyzing 28-day mortality rates seems more relevant.

As an example, in the recent Randomized Evaluation of COVID-19 Therapy (RECOVERY) trial, 2104 and 4321 patients were randomly assigned to dexamethasone and standard care [1]. The primary goal was to investigate whether patients would benefit from dexamethasone with respect to 28-day mortality. The study size was determined to provide 90% power, at $\alpha = .01$, for detecting a 4% decrease in mortality from 20% for standard care to 16% for dexamethasone. However, in the published report [1], the effect of treatment on survival among all participants was assessed using the hazard ratio (HR) only. The observed age-adjusted HR was 0.83 (95% confidence interval [CI], .75–.93]; $P = .0009$). The 28-day mortality rates reported in Figure 3 of [1] were 22.9% (482/2104) and 25.7% (1110/4321), but no formal assessment of the mortality rate difference among all patients was provided. The question is whether an HR alone provides sufficient clinical and statistical evidence to conclude that dexamethasone improved 28-day survival.

An HR of 0.83 is difficult to interpret clinically since hazard, which is the "force of mortality," is not a probability measure like risk. One cannot claim, for example, that dexamethasone reduced the risk of death by 17% across the study period. Moreover, it is not clear how to assess the survival benefit from a ratio alone; that is, without a reference hazard curve for standard care. Last, the validity of using the HR to assess the treatment effect depends on a strong proportional hazards assumption: that the ratio of the hazards from the dexamethasone and standard care groups is constant across time. Although for very short-term studies the HR may approximate the mortality rate ratio at the end of follow-up, the accuracy of such an approximation is not guaranteed, and this approximation is definitely not needed given that the mortality rate ratio may be directly estimated using survival rates from the Kaplan-Meier curves.

To explore whether additional survival time analysis would assist in interpreting the treatment effect, we scanned the survival curves in Figure 2A of the original paper [1] to reconstruct [2] the individual patient-level survival times. The corresponding Kaplan-Meier curves are presented in Figure 1A. The upper survival curve for dexamethasone is above the lower curve for standard care across the entire 28 days of follow-up, visually indicating that dexamethasone was superior to standard care.

The HR for Figure 1A, not adjusted for age [1], was 0.87 (95% CI, .78–.97; $P = .0089$). Although we cannot evaluate the original data, a standard Schoenfeld residual lack-of-fit test applied to the reconstructed data suggests that the proportional hazards assumption may not have held. Consequently, the HR does not have obvious clinical meaning for quantifying the treatment effect [3]. On the other hand, the higher the Kaplan-Meier curve, the better the treatment. Thus, the area under the Kaplan-Meier curve in Figure 1B or 1C is a reasonable summary of the survival profile over time, with a larger area indicating a more effective treatment. Moreover, the areas under the curve in Figures 1B and 1C have clinically meaningful interpretation as the mean survival times across the 28 days of follow-up [4, 5]. These were 23.8 and 23.1 days for dexamethasone and

Received 24 August 2020; editorial decision 9 October 2020; published online 14 October 2020.
Correspondence: L,-J. Wei, Department of Biostatistics, Harvard University, 655 Huntington Ave, Boston, MA 02115 (wei@hsph.harvard.edu).

Clinical Infectious Diseases® 2020;XX(XX):1–3

DOI: 10.1093/cid/ciaa1563

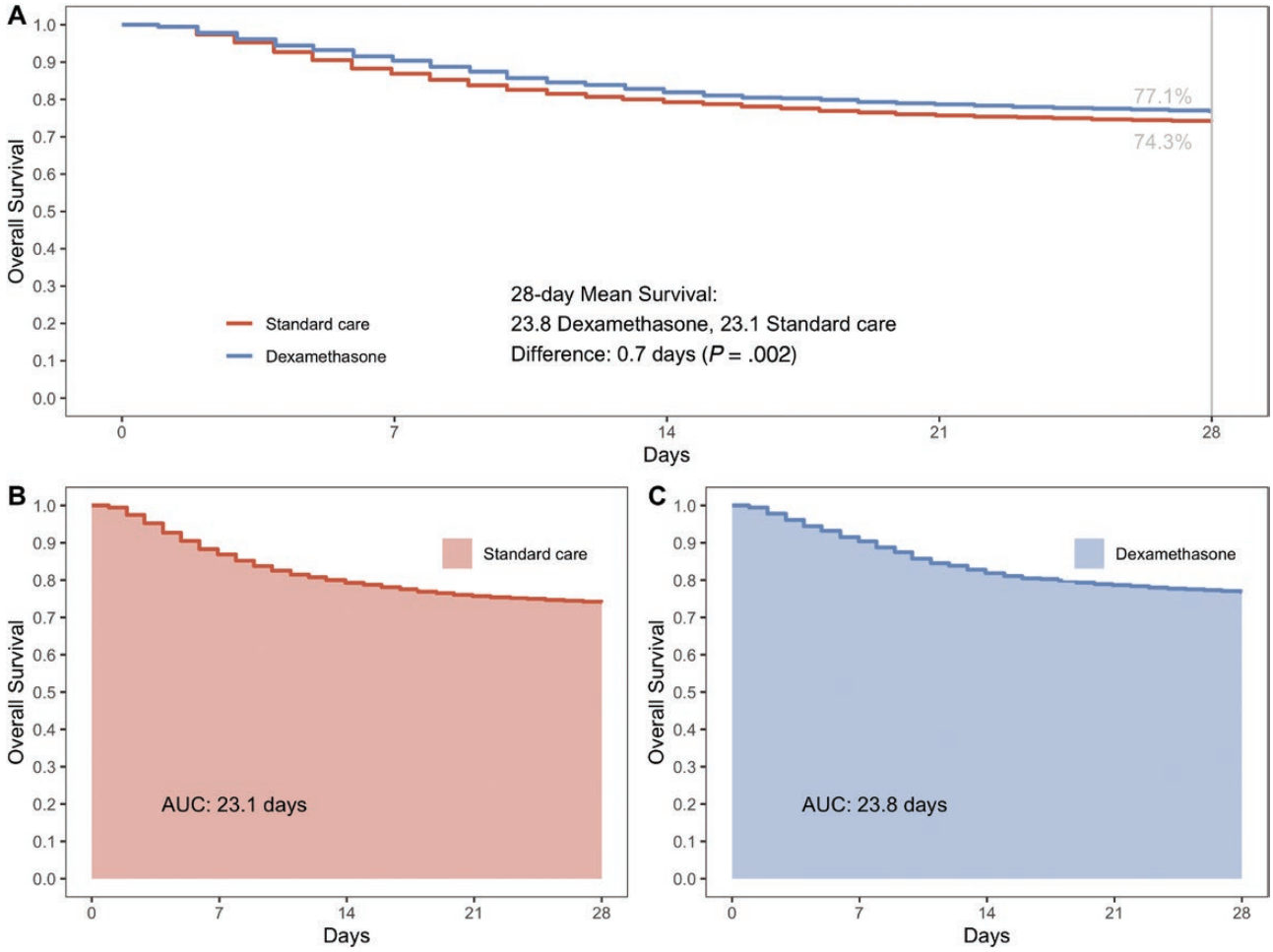

Figure 1. *A*, Reconstructed overall survival curves from the Randomized Evaluation of COVID-19 Therapy (RECOVERY) trial among all patients. *B* and *C*, 28-day mean survival times in the standard care and dexamethasone arms as the area under the Kaplan-Meier curve (AUC).

standard care. The difference of 0.7 days (95% CI, .3–1.2 days; $P = .002$) significantly favored dexamethasone. This time-scale summary of the treatment difference is more interpretable than the HR and requires no modeling assumptions for its validity. However, it is unclear whether the highly statistically significant gain of 0.7 days in mean survival time with dexamethasone is clinically meaningful, or adds any relevant information beyond the 28-day mortality rate difference of 2.8 percentage points (95% CI, .6%–5%; $P = .02$), which was not reported in the publication [1]. Note that the 28-day mortality rates of 22.9% for dexamethasone and 25.7% for standard care can easily be estimated from the corresponding Kaplan-Meier curves. Statistical inferences regarding the mortality rate difference/ratio can be performed by obtaining the variance estimates directly from the Kaplan-Meier curves, and without relying on either parametric or semiparametric assumptions, as in Cox regression.

As another example from RECOVERY, with reconstructed data, among patients requiring mechanical ventilation, the HR, unadjusted for age, was 0.67 (95% CI, .54–.84; $P < .001$), which is again difficult to interpret. The 28-day mean survival times were 22.7 and 20.7 days for dexamethasone and standard care, respectively, with a highly significant difference of 2.0 days (95% CI, .8–3.2 days; $P < .001$) in favor of dexamethasone. In this case, the mortality rates on day 28 were 29.3% and 41.4%, with an absolute difference of 12.1 percentage points (95% CI, 5.9–18; $P < .001$). Although all 3 comparisons are statistically significant, it is unclear whether the 33% relative reduction in hazard or the 2.0-day delay in mortality add any clinically relevant information beyond the clearly important 12.1 percentage point reduction in 28-day mortality. This example demonstrates that survival time analysis may obscure or underemphasize a clinically meaningful 28-day mortality rate benefit.

The HR is routinely used for assessing the treatment difference in survival analysis generally. Using the mean survival time difference across a specific time window may assist us to interpret the treatment difference in a more intuitive manner. One of the reasons for using survival time as an endpoint is to increase the statistical power for detecting a treatment benefit. However, the power gain is not always guaranteed [6], and an overall

difference between the 2 survival curves does not necessarily imply a mortality rate difference at a clinically meaningful time point. For short-term studies in critical care medicine, using the mortality difference at a specified time seems more clinically relevant. While the test based on mean survival time could be more powerful than those based on mortality difference/ratio at a single time point, the mean survival time difference in a short-term study tends to be very small. If the treatment only improves survival time without reducing mortality, then the small gain in mean survival time may not be clinically important. Therefore, the mortality rate difference/ratio is a preferred summary for the treatment effect in short-term studies, regardless of the power of the associated tests. RECOVERY is not unique in relying on the HR for quantifying the survival benefit. For instance, Adaptive COVID-19 Treatment Trial 1 (ACTT-1) quantified 14-day mortality using the HR only, with no comparisons of mortality rates presented in the published report [7]. It seems there is a misconception that the HR and the risk ratio are exchangeable for short-term studies.

In conclusion, for short-term mortality studies, reporting formal statistical analysis of the mortality rate difference/ratio, either at a prespecified timepoint or at the end of the study period, is warranted. It is not sufficient to claim statistical or clinical evidence of a survival benefit on the basis of, for example, an HR or mean survival time difference alone. Since there is no single summary measure that can capture the entire survival profile, various summaries of the treatment effect should be considered simultaneously, with the emphasis on directly comparing mortality rates. More importantly, any summary measure, such as the difference/ratio of mortality rates, needs to be accompanied by the individual summaries of the patients' survival in each treatment arm. A single group contrast, such as the HR alone, does not provide sufficient information to evaluate the treatment difference; the hazard rates from the individual treatment arms are needed for context.

Notes

Acknowledgments. The authors are grateful to the 2 reviewers for their insightful comments that have helped to improve the presentation of this manuscript.

Financial support. This work was supported by the National Institutes of Health (R21AG060227 to D. H. K. and R01 HL089778 to L. T.).

Potential conflicts of interest. The authors: No reported conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest.

References

1. RECOVERY Collaborative Group. Dexamethasone in hospitalized patients with Covid-19—preliminary report [manuscript published online ahead of print 17 July 2020]. New Engl J Med **2020**. doi:10.1056/NEJMoa2021436.
2. Guyot P, Ades AE, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Med Res Methodol **2012**; 12:9.
3. 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**; 32:2380–5.
4. Kim DH, Uno H, Wei LJ. Restricted mean survival time as a measure to interpret clinical trial results. JAMA Cardiol **2017**; 2:1179–80.
5. Pak K, Uno H, Kim DH, et al. Interpretability of cancer clinical trial results using restricted mean survival time as an alternative to the hazard ratio. JAMA Oncol **2017**; 3:1692–6.
6. Lin DY, Zeng DL, Eron JJ. Evaluating the efficacy of therapies in COVID-19 patients. Clin Infect Dis **2020**. doi:10.1093/cid/ciaa1231.
7. Beigel JH, Tomashek KM, Dodd LE, et al. Remdesivir for the treatment of Covid-19—a preliminary report. New Engl J Med **2020**. doi:10.1056/NEJMoa2007764.