Perspectives

Caution About Truncation-By-Death in Clinical Trial Statistical Analysis: A Lesson from Remdesivir

Yuhao Deng[1]; Xiao-Hua Zhou[2,3,4,#]

In an effort to combat coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), clinicians across the globe have been working tirelessly to find effective treatments. In 2020, inspiring drug trials have focused on treating COVID-19 with the antiviral drug remdesivir. Beigel et al. conducted a well-designed multicenter randomized trial, where 541 patients were assigned to receive remdesivir, and 522 patients were assigned to receive placebo treatment (*1*). They reported that those who received remdesivir had a median recovery time of 11 days [95% confidence interval (CI): 9–12], as compared with 15 days (95% CI: 13–19) in those who received a placebo (*P*<0.001).

Another multicenter randomized trial was conducted by Wang et al. at 10 hospitals in Hubei Province, China (*2*). They reported that remdesivir use had a positive but insignificant effect compared with standard care in the time to clinical improvement [hazard ratio (HR)=1.23, 95% CI: 0.87–1.75]. Due to early suspension of the trial because of adverse events, this study was underpowered, and the findings were deemed to be inconclusive (*3*). Discrepant findings between these two studies show that a small sample size may fail to achieve the predetermined power or make an expected conclusion. Despite the small sample size, the remdesivir studies for COVID-19 can still also provide a lot of valuable information under more careful statistical analysis.

First, we must be cautious whether remdesivir is safer than placebo. Beigel et al. found serious adverse events among 114 of the 541 (21.1%) patients in the remdesivir group and 141 of the 522 (27%) patients in the placebo group, while Wang et al. reported adverse events in 102 of 155 (66%) remdesivir recipients versus 50 of 78 (64%) placebo recipients. This difference in adverse events may be attributable to underlying medical diseases among patients included in the studies. As Wang et al. described in their article, the remdesivir group included more patients with hypertension, diabetes, or coronary artery disease than the placebo group, which led to an imbalance between the two treatment groups. Although this study was randomized at baseline, randomization alone does not guarantee balance between the treatment and placebo groups. The consequence of clinical trials relying on pure randomization has been discussed in some statistics literature: the treatment effect estimate may be far from the true value if the sample size is not large enough (*4*–*6*). Thus, the causal effect of underlying medical diseases on recovery rate or safety outcomes may be confounded by the severity of underlying diseases. By comparing these two studies, we have reason to think that remdesivir has different effects on populations with different baseline status.

Another topic related to treatment imbalance is the statistical analysis for the truncated-by-death individuals. Beigel et al. reported that the Kaplan-Meier estimates of mortality by 14 days were 7.1% with remdesivir and 11.9% with placebo (HR for death=0.70, 95% CI: 0.47–1.04). An analysis with adjustment for baseline ordinal score as a stratification variable showed a HR for death of 0.74, 95% CI: 0.50–1.10. Wang et al. also found insignificant differences in mortality between the remdesivir group and the placebo group. Similar with adverse events, the possibility that death was associated with underlying diseases cannot be excluded. If patients with underlying diseases were more likely to develop adverse events or die, then the treatment effect of remdesivir versus placebo would have been underestimated (*7*).

It is worth noting that handling the truncation-by-death problem is different from censoring. That is to say, the Kaplan-Meier approach, which is commonly adopted in survival analysis with censoring, should be used with great care if the target of a study is the time to clinical improvement or recovery instead of mortality. Censoring can be understood as a missing data problem: the time to clinical improvement or recovery does exist but is longer than the study period. For example, if a patient recovered at Day 35 but the study ended at Day 28, then the time to recovery was censored. In the methodology, partial likelihood is calculated for at-risk individuals at each time point.

However, truncation-by-death is a completely different issue. If a patient was truncated by death, then his/her outcome (time to clinical improvement or recovery) is undefined. In the classical survival analysis, every individual would experience the failure event at some day. But by definition, a patient that dies at Day 21, for example, should not be treated as censored at any day because he/she has lost the ability to experience the failure event (referring to clinical improvement or recovery here), and the failure event would never occur no matter how long the follow-up is. To be brief, treating death as censoring would confuse different types of outcomes. Different statistical procedures should be adopted in dealing with the truncation-by-death problem.

In fact, comparing treatment effects is a question about causal inference. Under the potential outcome framework, each individual has two potential outcomes: one under the treatment, the other under the control. The treatment effect is the difference of these two potential outcomes. By principal stratification, one can divide the whole population into four strata (*8*):

(1) LL, always survivor, alive either if treated or untreated.

(2) LD, protected, alive if treated but dead if untreated.

(3) DL, harmed, dead if treated but alive if untreated.

(4) DD, doomed, dead either if treated or untreated.

The fundamental problem in causal inference is that one can only observe one of these two potential outcomes, since a patient can either be treated or untreated, but not both. Thus, the observed alive individuals at Day 28 come from mixed strata: LL and LD for the treatment group and LL and DL for the placebo group. However, the treatment effect is only meaningful in the LL stratum, since the pair of potential outcomes, clinical improvement, or recovery, are only well defined in the LL stratum.

Since common clinical analyses considered death as right censoring at the endpoint, we cannot conclude whether the estimates represent a meaningful parameter. In order to identify the LL stratum, a substitutional variable for survival is needed (*9–10*). Since the information of baseline covariates and COVID-19 is still insufficient, we do not know whether a qualified substitutional variable exists or not. Analysis based on observed survivors may underestimate the true treatment effect due to the positive correlation between the severity of underlying diseases and death, so principal stratification or baseline adjustment for underlying diseases is recommended for better statistical analysis.

To address the truncation-by-death problem, we use generated simulation data (see Supplementary Material for simulation details, available at weekly.chinacdc.cn) that mimics the findings of Beigel et al. to show the grave consequence of considering death as right censoring. Suppose that 500 patients are enrolled into the treatment group and 500 patients are enrolled into the placebo group. The probability of possessing underlying disease is 0.4 in the treatment group and 0.3 in the placebo group. The probability of death is 0.3 with underlying diseases and 0.1 with no underlying diseases. Thus, by assuming that death is independent of treatment course, the probability of being alive is 0.82 in the treatment group and 0.84 in the placebo group. Suppose the time to recovery (or clinical improvement) follows an exponential distribution with mean 11 days if receiving treatment and 15 days if receiving placebo, so the true HR is 1.36. Recovery time of more than 30 days is considered as right censored. We simulate for 500 runs and use the Cox proportional hazard model to analyze the data. The procedures and codes are listed in the Appendix.

(1) If the dead individuals are regarded as right censored at Day 30, the average estimated HR is 1.27 (s.e.=0.09, average *P*-value=0.012).

(2) If conditioning on the alive individuals, the estimated HR is 1.37 (s.e.=0.10, average *P*-value 0.002).

(3) If conditioning on the alive subsample and weighting the alive individuals by the survival probability in each group, the estimated HR is 1.37 (s.e.=0.10, average *P*-value=0.001).

(4) If dividing the sample into 2 subsamples of possessing underlying diseases and not possessing underlying diseases, and regarding the dead individuals as right censored at Day 30, the estimated HR is 1.27 (s.e.=0.13, average *P*-value=0.125) in the former subsample and 1.33 (s.e.=0.11, average *P*-value=0.017) in the latter subsample.

(5) If dividing the sample into 2 subsamples of possessing underlying diseases and not possessing underlying diseases and conditioning on the alive individuals, the estimated HR is 1.38 (s.e.=0.19, average *P*-value=0.099) in the former subsample and 1.38 (s.e.=0.13, average *P*-value=0.013) in the latter subsample.

One can see that regarding death as right censoring would underestimate the treatment effect, even if stratifying the severity of underlying diseases. The second and third approaches yield similar estimates and are close to the true value, because the survival probability is similar in the two groups. The fourth and fifth approaches have larger standard errors and *P*-values due to the decline of sample size by dividing the observed sample.

At the very least, to make the assumption of truncation-by-death at random (i.e., death is independent of treatment) more convincing, baseline covariates such as disease severity (baseline score) and underlying diseases should be adjusted if the imbalance in enrollment is obvious, as Beigel and his colleagues did in their work. Furthermore, a better experimental design at the design phase would allow for a more hassle-free analysis at the analysis phase. It is true that randomization for recruitment is a commonly adopted approach to eliminate the effects of confounding. Still, there are some approaches to improve the randomization at the design phase that can minimize the impact of confounding and treatment imbalance if a few critical covariates exist. For example, rerandomization can be used to balance the covariates between the treatment and placebo groups (*11*). By iteratively trying to randomize the assignment, only the assignment that satisfies some criterion (for example, the distance of covariates between the treatment and the placebo groups is lower than a threshold) can enter into the experiment. It is encouraging that statistical inference under rerandomization is still valid with a little adjustment (*12*). Therefore, in future clinical studies, we suggest that greater attention should be given to the design phase, so that problems that may occur in the analysis phase can be avoided.

Funding: The Bill and Melinda Gates Foundation (INV-016826).

doi: 10.46234/ccdcw2021.139

[#] Corresponding author: Xiao-Hua Zhou, azhou@math.pku.edu.cn.

[1] School of Mathematical Sciences, Peking University, Beijing, China; [2] Beijing International Center for Mathematical Research, Peking University, Beijing, China; [3] Department of Biostatistics, School of Public Health, Peking University, Beijing, China; [4] National Engineering Lab for Big Data Analysis and Applications, Peking University, Beijing, China.

Submitted: June 10, 2021; Accepted: June 11, 2021

REFERENCES

1. Beigel JH, Tomashek KM, Dodd LE. Remdesivir for the treatment of covid-19—preliminary report. reply. N Engl J Med 2020;383(10):994. http://dx.doi.org/10.1056/NEJMc2022236.
2. Wang YM, Zhang DY, Du GH, Du RH, Zhao JP, Jin Y, et al. Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. Lancet 2020;395(10236): 1569 – 78. http://dx.doi.org/10.1016/S0140-6736(20)31022-9.
3. Norrie JD. Remdesivir for COVID-19: challenges of underpowered studies. Lancet 2020;395(10236):1525 – 27. http://dx.doi.org/10.1016/S0140-6736(20)31023-0.
4. Urbach P. Randomization and the design of experiments. Philos Sci 1985;52(2):256 – 73. http://dx.doi.org/10.1086/289243.
5. Rubin DB. Comment: the design and analysis of gold standard randomized experiments. J Am Stat Assoc 2008;103(484):1350 – 53. http://dx.doi.org/10.1198/016214508000001011.
6. Worrall J. Evidence: philosophy of science meets medicine. J Eval Clin Pract 2010;16(2):356 – 62. http://dx.doi.org/10.1111/j.1365-2753.2010.01400.x.
7. Rücker G, Schumacher M. Simpson's paradox visualized: the example of the rosiglitazone meta-analysis. BMC Med Res Methodol 2008;8:34. http://dx.doi.org/10.1186/1471-2288-8-34.
8. Rubin DB. Causal inference through potential outcomes and principal stratification: application to studies with "censoring" due to death. Stat Sci 2006;21(3):299 – 309. http://dx.doi.org/10.1214/088342306000000114.
9. Ding P, Geng Z, Yan W, Zhou XH. Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. J Am Stat Assoc 2011;106(496):1578 – 91. http://dx.doi.org/10.1198/jasa.2011.tm10265.
10. Wang LB, Zhou XH, Richardson TS. Identification and estimation of causal effects with outcomes truncated by death. Biometrika 2017;104(3):597 – 612. http://dx.doi.org/10.1093/biomet/asx034.
11. Morgan KL, Rubin DB. Rerandomization to improve covariate balance in experiments. Ann Stat 2012;40(2):1263 – 82.
12. Li XR, Ding P, Rubin DB. Asymptotic theory of rerandomization in treatment–control experiments. Proc Natl Acad Sci USA 2018;115(37):9157 – 62. http://dx.doi.org/10.1073/pnas.1808191115.

Supplementary Material

We conduct a simulation study to show truncation-by-death is different from censoring. The simulation is conducted using R statistical software (version 3.6.1; The R Foundation for Statistical Computing, Vienna, Austria). The following R code lists the procedure of model fit in each iteration.

(1) Generate the recovery time with right censoring of treatment group and placebo group.

```
n1=500
n0=500
t1=round(rexp(n1,1/11),0)
t0=round(rexp(n0,1/15),0)
t1[t1>30]=30
t0[t0>30]=30
c1=as.numeric(t1<30)
c0=as.numeric(t0<30)
```

(2) Consider two strata: with and without underlying medical diseases. Generate the survival status.

```
s1=rbinom(n1,1,0.4)
s0=rbinom(n0,1,0.3)
d1=–(s1×rbinom(n1,1,0.3)+(1–s1)×rbinom(n1,1,0.1))
d0=1–(s0×rbinom(n0,1,0.3)+(1–s0)×rbinom(n0,1,0.1))
```

(3) The observed recovery time T, observability R, underlying diseases S, survival status D, treatment X.

```
T1=t1×d1+30×(1–d1)
T0=t0×d0+30×(1–d0)
R1=apply(rbind(c1,c1),2,min)
R0=apply(rbind(c0,c0),2,min)
X=c(rep(1,n1),rep(0,n0))
R=c(R1,R0)
T=c(T1,T0)
S=c(s1,s0)
D=c(d1,d0)
W=1/c(rep(0.82,n1),rep(0.84,n0))
```

(4) Fitting the Cox proportional hazard model.

```
library(survival)
res.cox1<–coxph(Surv(T,R)~X)
res.cox2<–coxph(Surv(T,R)~X, subset=(D==1))
res.cox3<–coxph(Surv(T,R)~X, weights=W, subset=(D==1))
res.cox41<–coxph(Surv(T,R)~X, subset=(S==1))
res.cox42<–coxph(Surv(T,R)~X, subset=(S==0))
res.cox51<–coxph(Surv(T,R)~X, subset=(D==1&S==1))
res.cox52<–coxph(Surv(T,R)~X, subset=(D==1&S==0))
```