Replacing *P* values with confidence intervals may not achieve anything

Seo Young Park, PhD

P values from null hypothesis significance testing (NHST) have been the standard way to report the result of medical research, but it is facing increasing criticism today. The medical science community has begun to make efforts to move away from the *P* value, one of which is the set of new guidelines recently announced by the *New England Journal of Medicine*. These guidelines can be summarized as the importance of setting up the statistical analysis plan ahead of the study and sticking to it, discouraging use of the *P* value unless the corresponding test was prespecified and its multiplicity is properly controlled, and emphasizing the reporting of confidence intervals (CIs). In particular, the guideline says if no method to adjust multiplicity was specified in the protocol or statistical analysis plan, then the report of all secondary and exploratory outcomes should be done by point estimates and 95% CIs of the effect of interest, not by the *P* values. Although this is a meaningful step toward a "world beyond $P < .05$,"[1] I want to point out that replacing *P* values with CIs might not make any real change in how medical research is being conducted and how results are understood. Because of their duality, the *P* value and CI deliver essentially the same information—the compatibility of data with the model. Indeed, CIs put more emphasis on estimation compared with hypothesis testing, and they provide clues about the precision of the estimate. Still, the location or width of the CI does not translate to clinical significance, and we all know that the simplistic use of the CI that dichotomizes the result into "success" versus "failure" by checking whether the CI includes the null value (usually 0 or 1) will persist. Moreover, interpretation of the CI is not straightforward. Despite common misconception, the 95% CI of a parameter does not tell us that the parameter of interest falls within such a CI with probability of .95. What it really means is that if we take repeat random samples of the current sample size from the same population to obtain the 95% CI many times, 95% of such calculated CIs would include the true value of the parameter. Often times this is not how the investigators think of 95% confidence intervals.

Seo Young Park, PhD

CENTRAL MESSAGE

Replacing *P* values with confidence intervals might not achieve the desired goal; using *P* values and testing of more clinically relevant hypotheses rather than null hypotheses may more accurately reflect clinical significance.

This Invited Expert Opinion provides a perspective on the following paper: *N Engl J Med*. 2019;381(3):285-286. https://doi.org/10.1056/NEJMe1906559.

I believe that NHST still has its place in medical research, when used appropriately for the prespecified comparisons carefully planned before the study. Routine reporting of *P* values for all variables should be discouraged. In particular, I want to encourage reporting of the standardized mean difference instead of *P* value for group comparisons in the baseline characteristics tables in the *Journal*, because the main purpose of such tables is to describe their study sample, not to make inferences about true difference between

From the Department of Clinical Epidemiology and Biostatistics, Asan Medical Center, University of Ulsan College of Medicine, Seoul, South Korea.

Received for publication March 1, 2020; revisions received April 1, 2020; accepted for publication April 4, 2020; available ahead of print July 22, 2020.

Address for reprints: Seo Young Park, PhD, Department of Clinical Epidemiology and Biostatistics, Asan Medical Center, University of Ulsan College of Medicine, Seoul, South Korea (E-mail: biostat81@amc.seoul.kr or biostat81@gmail.com).

J Thorac Cardiovasc Surg 2021;161:1379-80
0022-5223/$36.00

https://doi.org/10.1016/j.jtcvs.2020.04.139

the groups. The use of *P* values may give the illusion that groups were comparable, when actually large *P* values were merely the result of a small sample size. In confirmatory studies with the goal of investigating clinical significance of treatment/exposure effect, instead of *P* value from NHST, it seems more reasonable to report the *P* value from a test of a prespecified minimal important effect size, which was suggested by Amrhein and colleagues[2] and Greenland.[3] For example, rather than testing whether the odds ratio (OR) of a treatment is 1 or something different, one may test whether OR is <0.7 when it has been predetermined that the OR of the treatment needs to be at most 0.7 to be clinically important. They also suggested transforming such *P* values into *S* values, so named for C. E. Shannon, the father of information theory. The *S* value is negative base-2 log of a *P* value [$s = -\log_2(p)$]. This represents the bits (binary digits) to encode the information against the hypothesis being tested. For example, a *P* value of 0.05 transforms to an *S* value of $-\log_2(0.05) = 4.3$, which means 4.3 bits of information against the hypothesis being tested. This also means that under the hypothesis being tested, the observed data are a little more surprising than seeing all heads in 4 fair coin tosses but less surprising than seeing 5 heads in 5 fair coin tosses. The *S* value has a nice interpretation and also moves users away from the trap of an arbitrary cutoff of .05.

Statisticians have suggested numerous other alternatives to NHST *P* values, some of which have been nicely summarized in the special issue of *The American Statistician* containing the articles referenced herein. I believe that many more new methods will be proposed, and that the discussion on the use of statistical inference in scientific research may never end. I want to encourage the authors, reviewers, and the editorial board to rethink the current practice in research and to be open-minded to alternative ways to make inferences from data, to make a fair and robust judgments about the clinical significance of the observed effects, and to improve the quality and reproducibility of research.

Conflict of Interest Statement

The author reported no conflicts of interest.

The *Journal* policy requires editors and reviewers to disclose conflicts of interest and to decline handling or reviewing manuscripts for which they may have a conflict of interest. The editors and reviewers of this article have no conflicts of interest.

References

1. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "$P < 0.05$". *Am Stat*. 2009;73(Supp1):1-19.
2. Amrhein V, Trafimow D, Greenland S. Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *Am Stat*. 2009;73(Supp1):262-70.
3. Greenland S. Valid *P* values behave exactly as they should: some misleading criticisms of *P*-values and their resolution with *S*-values. *Am Stat*. 2009;73(Supp1):106-14.