Statistical significance: Is there a way out of it?

Adin-Cristian Andrei, PhD

Investigating "statistical significance" in biomedical research studies can constitute a serious source of anguish and anxiety for all parties involved. For some medical investigators, it might be one of those treacherous publish-or-perish situations leaning rather toward a missed encounter with success. For those conducting data analyses, it might be yet another instance in which one needs to convey "unpleasant *P* value news" to their collaborators. In reality, in most situations, such artificial crises originate in the very nature of statistical and clinical significance and are avoidable for the most part with adequate planning.

Statistical significance came into existence to address a simple quantitative question arising after concluding an experiment: are the data obtained consistent with the study hypothesis? Presented in this way, it is clear what statistical significance does not do: it does not confirm or refute the hypothesis. Instead, significance is a statement pertaining to the data, under the assumption that the hypothesis is true. The *P* value is an amalgamation of the data distilled into a single value. A small *P* value suggests that the data are inconsistent with the hypothesis. Logically, this may prompt a reevaluation of the hypothesis. Conversely, a large *P* value does not validate the hypothesis. It merely indicates that the data collected are not in an obvious contradiction with the experiment hypothesis.

Adin-Cristian Andrei, PhD

CENTRAL MESSAGE

Statistical testing remains a fundamental scientific tool, despite its limitations. The focus should be on its adequate and judicious use.

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.

Statistical Testing in a Crisis

Why are *P* value–driven conclusions and studies facing an evident crisis? The answer is not simple, given the numerous contributing factors. An evident root cause is foundational: hypothesis testing was not designed to unequivocally establish the irrefutable veracity of a hypothesis. Because such decisions rely on data that in turn are subject to randomness, the *P* value was born out of a need to quantify how implausible the observed data are, should the hypothesis in question be assumed to be true. This is how the "statistical significance" label came into existence. Deeply engrained in the psyche of any student of statistics these days is fact that a "*P* value $< .5$" is a threshold that separates findings that matter (significant) from those that are not important (not significant). It is exactly this type of dichotomization that serves as an inexhaustible source of contradictions and motivates the intense focus on reproducibility and replicability in the scientific community. We should resist the dogma of classifying statistical results into either significant or not significant; instead, we should convey the relevant data summaries, *P* values included, and allow the readership to weigh the evidence and its strength and form an opinion. After all, in all fairness, the practical value of a study resides primarily in its clinical significance and its potential for outcome improvement. For example, in a study comparing mortality, a highly statistically significant hazard ratio of 1.0001 is likely clinically irrelevant. That is why it is important to decide and declare—before data analysis—what constitutes clinical relevance. Doing so would limit the number of statistical tests performed, thus reducing the possibility of type I errors.

Are There Ways to Perform Less Statistical Testing?

An efficient way to reduce the number of statistical hypotheses tested is to focus on the most meaningful questions. For example, a comparison of low-mortality cohorts—say, less than 1% in each—is unlikely to detect a statistically significant difference unless group sizes are very large. Therefore, precious type I error is easily saved

From the Division of Biostatistics, Department of Preventive Medicine, Northwestern University, Chicago, Ill.

Received for publication April 15, 2020; revisions received April 15, 2020; accepted for publication April 16, 2020; available ahead of print June 4, 2020.

Address for reprints: Adin-Cristian Andrei, PhD, Division of Biostatistics, Department of Preventive Medicine, Northwestern University, 680 N Lake Shore Drive, Suite 1400, Chicago, IL 60611 (E-mail: a-andrei@northwestern.edu).

J Thorac Cardiovasc Surg 2021;161:1377-8
0022-5223/$36.00

https://doi.org/10.1016/j.jtcvs.2020.04.138

by avoiding statistical testing that would not permit a definitive conclusion anyway. The same principle applies to other studies, for example, those in which rare events are evaluated and in which the lack of statistical power to produce meaningful statements should be fairly obvious. Such studies are best to remain purely descriptive. Testing in small-scale studies also should be performed judiciously; even if higher proportions of participants experience the event of interest, that may still produce insufficiently reliable conclusions because of inadequate scaling. When statistical testing is inopportune, irrelevant, or potentially inconclusive, it is probably advisable to avoid it altogether and save precious type I error but instead, before encountering the data, selectively focus on hypotheses and questions that have a decent potential to produce reliable (and not necessarily positive) findings.

Can Experimental Design Enhance *P* Value Credibility?

The answer is, evidently, affirmative. For example, when the goal is to investigate whether surgical procedures A and B produce comparable results, one is best served by using tests for equivalence and not for superiority. With properly chosen, clinically meaningful equivalence margins, such tests could achieve the best of both worlds (clinical and statistical) simultaneously, while remaining scientifically sound. Other experimental design features are also potentially useful. Choices that lead to less biased or less variable estimates clearly impact *P* values in a positive way, rendering them more credible. Adequately selecting the statistical testing methodology is also critical. To illustrate, a study comparing overall survival in groups A and B is best to use the log-rank test if the underlying hazards are proportional. Otherwise, a test relying on the difference in (truncated) Kaplan–Meier curves might be more highly powered, yielding a *P* value that would permit a more reliable conclusion.

Is It Necessary to Avoid Statistical Testing?

Statistical testing remains a critical component in the quest for valid scientific conclusions. It should be used judiciously, under a well-described analytical plan specified before encountering the data. It should be focused on important questions and avoided in peripheral or inconsequential ones. It should also be performed by those who are adequately trained to do so and who have the necessary knowledge to make valid study design choices that enhance the usefulness of statistical testing.

Returning to the title question, it is fairly clear that statistical testing is here to stay, despite its limitations. Avoiding it altogether would be unwise and just as counterproductive as overusing it. There is no inconsequential way out of statistical testing, but there are numerous and reliable venues to put it to good use and create valuable and, most importantly, valid knowledge.

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.