JNCI Cancer Spectrum (2020) 4(2): pkaa012

doi: 10.1093/jncics/pkaa012
First published online February 26, 2020
Solicited Commentary

In Defense of *P* Values

Colin B. Begg , PhD*

Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

*Correspondence to: Colin B. Begg, PhD, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY 10021, USA (e-mail: beggc@mskcc.org).

Abstract

Recently, a controversy has erupted regarding the use of statistical significance tests and the associated *P* values. Prominent academic statisticians have recommended that the use of statistical tests be discouraged or not used at all. This has naturally led to a lot of confusion among research investigators about the support in the academic statistical community for statistical methods in general. In fact, the controversy surrounding the use of *P* values has a long history. Critics of *P* values argue that their use encourages bad scientific practice, leading to the publication of far more false-positive and false-negative findings than the methodology would imply. The thesis of this commentary is that the problem is really human nature, the natural proclivity of scientists to believe their own theories and present data in the most favorable light. This is strongly encouraged by a celebrity culture that is fueled by academic institutions, the scientific journals, and the media. The importance of the truth-seeking tradition of the scientific method needs to be reinforced, and this is being helped by current initiatives to improve transparency in science and to encourage reproducible and replicable research. Statistical testing, used correctly, has an important and valuable place in the scientific tradition.

In the past year, a furor has erupted over the use of statistical testing in medical research, stimulated by a widely reported article in the journal *Nature* by Amrhein, Greenland, and McShane (1). The article was a polemic describing how research investigators misuse and misinterpret the results of statistical tests. This article added fuel to concerns about statistical testing that had led earlier to the unusual step of the American Statistic Association issuing a policy statement on *P* values, authored by a number of prominent statisticians, that was broadly critical of their widespread use (2). For many scientists and oncologists, this was news, and it certainly has caused some widespread befuddlement at the perceived notion that academic statisticians don't approve of statistical methods. However, although this was indeed news, it wasn't really new. Acrimonious debates about the merits of *P* values, statistical testing, and indeed the foundations for conducting statistical analyses have been going on for decades. A notable early critic of *P* values was Rothman, who argued in the *New England Journal of Medicine* in 1978 that statistical significance tests should be replaced by the use of confidence intervals and who later went on to ban the use of *P* values in the journal he subsequently edited, *Epidemiology* (3). In the more recent past, the editor of the journal *Basic and Applied Social Psychology* created a stir by making a similar policy decision for that journal (4). Countless other prominent commentators have weighed in over the years to try to influence users of statistical tests to be less reliant on them, including Gardner and Altman in their influential series on statistics in medical research in the *British Medical Journal* (5). As will be explained later, the basic thesis of these commentators is that the use of statistical tests hampers the scientific enterprise by encouraging bad science.

So what's going on here? Is statistical testing really a major fly in the ointment hampering the progress of science? Or is this just one of those debates that takes place all the time in the ivory tower that can be safely ignored by the wider scientific community? My goal in this commentary is to frame the issue regarding statistical testing in the context of the use of statistical methods in general. I will argue that the attempts to eliminate the use of statistical tests are deeply misguided and that, instead, statisticians and others concerned about the role of data analysis in the quality of scientific research should focus on the issues that really matter, some of which are also receiving much contemporary attention through the focus on reproducibility, replicability, and transparency in medical research.

Why Are *P* Values so Popular With Statistical "Users"?

The construction of the statistical test and its summary by a single statistic, the *P* value, provides an extraordinarily versatile

Received: February 5, 2020; **Revised:** February 10, 2020; **Accepted:** February 14, 2020

and simple framework for making a statistical inference. As a result, thousands of different types of statistical tests are in widespread use, from the basic *t* test to countless specialized tests that have been developed for specific technical settings, such as the log-rank test for comparing survival curves, the McNemar test for comparing paired binary outcomes, and so on. All such tests can be summarized by the *P* value, which has a simple, unified interpretation. Its use is ubiquitous throughout science. Moreover, the interpretation is straightforward and, I would argue, well understood by users. In technical terms, the *P* value is the probability that the test statistic, or a more extreme value, is observed when the hypothesis being tested (usually referred to as the "null" hypothesis) is true. Thus, small values suggest the hypothesis is false. Although many users may not be fully familiar with this technical definition, in my opinion they do understand its essence—you challenge a hypothesis and declare the hypothesis false if the *P* value meets some criterion, usually *P* less than .05. Academic critics try to make the case that users don't actually understand the meaning of the test, claiming that users often interpret the *P* value as the probability that the hypothesis is true (2) or that a nonstatistically significant finding means that the hypothesis is true (1), or more esoterically, that the hypothesis can never be true from a literal standpoint (6). But these concerns miss the point. The statistical test is primarily designed for and used as a litmus test, and users understand this at a core level. In other words, the statistical test is enormously popular because it is exceptionally simple and remarkably versatile. Indeed, in an earlier phase of this controversy 20 years ago, Weinberg spoke eloquently to the many settings in which statistical testing provides a valuable and arguably irreplaceable strategy for scientific inferences (7). It is for these reasons that the statistical test and associated *P* value have become so embedded for so long in the scientific infrastructure.

Why Are *P* Values so Unpopular With Some Statistical Theorists?

The current *P* value debate in many ways resembles that of a book or a movie that is wildly beloved by the public but scorned by professional critics in the media. The criticisms about statistical tests are largely coming from academic statistical methodologists while users in the scientific community vote with their feet. I have already explained why statistical tests are popular with users. I am not going to dwell further on the technical understanding (or lack thereof) of statistical tests by users, because I don't think such concerns are important, as indicated above. However, I will focus on the much more substantive concern that the use of statistical testing leads to bad scientific practice and, by implication, use of alternative methods will improve scientific practice.

The conventional use of a statistical test leads to a "binary" decision: either the test is significant, whereby we conclude that the hypothesis being tested is false, or it is not significant, whereby we cannot conclude that the hypothesis is false. This "dichotomization" of the statistical inference is indeed one of the main concerns of critics, in that it oversimplifies what should be a judgment about strength of evidence rather than an either/or conclusion (8). The significance level (eg, 5%) represents, in a valid test setting, the probability of a false-positive result. The probability of a false-negative result is not embedded in the *P* value itself, because it depends on the unknown magnitude of the true positive effect under investigation. Critics of significance testing argue that the way tests are used encourages both false-positive and false-negative findings, with the result that the strengths of evidence underlying lots of scientific findings reported in the literature are grossly overstated. The false-positive probability can only be reliably inferred in a very carefully designed and executed experiment, such as a randomized trial, where the protocol is followed to the letter. However, there are many ways that the legitimacy of the statistical testing framework can be undermined even in the setting of a randomized trial, such as by changing the primary endpoint or its definition, selectively excluding patients from the analysis because of perceived eligibility concerns, and so forth. In the much broader setting of observational research, there often is no defined primary endpoint, the selection of cases to the groups being compared is not random, and statistical analyses often have a much more exploratory flavor at the outset. Indeed, much research published is not protocol-driven at all. All of these factors do not necessarily inflate the false-positive rate inherently. However, data analysts know that achieving a statistically significant result provides a benchmark that will give legitimacy to the scientific theses they are investigating, and so human nature dictates that investigators will be inclined to select and present those analyses that best support their preferred scientific theses. The point I'm trying to make here is that it is not the statistical testing framework itself that is flawed, it is the human factor driving the statistical analysis toward a preferred outcome.

Another concern is that significance testing also encourages false-negative interpretations of statistical tests. In fact, this concern was a major focus of the *Nature* article cited at the beginning of this commentary (1). To understand this, consider a clinical trial comparing the effects of two drugs. The probability that a test will turn out to be significant (the statistical power) depends on how truly different the effects of the drugs are. If one drug is much better than the other, it is highly likely the result of the trial will be significant, but if the difference is modest, it is much more likely the results will not be statistically significant, even though the hypothesis being tested—that the drugs are equivalent—is not true. One should never draw the conclusion of equivalence merely from a nonsignificant finding. Yet, many authors seem to do this. Interestingly, the establishment of equivalence in the clinical trials setting can be addressed through so-called equivalence trials, whereby the acceptable degree of similarity (admittedly subjective) of drug effect is built into the design (9).

In summary, the premise of critics of significance testing that studies reported in the literature are far more often false-positives than the theory would imply (ie, that, at most, 5% of tests conducted are false-positives) is undoubtedly correct. It is also true that investigators frequently draw inappropriate conclusions of equivalence following findings of nonsignificance. That is, statistical tests are frequently misapplied and misinterpreted by users. But does this mean that we will all be better off if investigators cease to use statistical testing?

The Wider Landscape of Data and Scientific Inferences From Data

The interpretation of data lies at the heart of scientific investigation. Everyone who engages in research knows that they must conduct studies to produce data to support theories that they develop. Despite this, serious concerns have been raised, legitimately, by thought leaders such as the director of the

National Institutes of Health, that medical science is plagued by poor study design and inappropriate use of statistical methods (10). This opinion is driven by a perception that many published discoveries turn out to be false. The proposed remedy is better training in principles of study design and statistical analysis. However, the reality is much more complex than this picture.

The Ideal Experiment and the Reality

The ideal scientific experiment is protocol-driven, as has been well recognized in the clinical oncology community for decades (11). The randomized trial is frequently cited as representing this ideal. However, it is far more than simply the use of randomization that is needed to ensure a fully credible test of a hypothesis under investigation. It is crucial that key details of the study be prespecified and followed. These key details include, most importantly, precise specification of the primary endpoint and how it is ascertained. Investigators that modify the endpoint in the light of emerging data can undermine the validity of the statistical analysis. Analyses of selected subgroups also have reduced credibility. The major reason that published reports of trials that deviate from the protocol have lower validity is not because protocol deviations inherently increase the risk of false-positive findings. It is because such deviations are typically elective, chosen selectively by the investigators to reflect a positive (ie, statistically significant) result. Investigators don't choose to alter protocol details to make the results less convincing; they alter them to make the results more convincing. It is for this reason that there have been initiatives by journal editors to link reports of trials to their protocols during the review process to discourage such elective modifications, although it has been shown that registered protocols are often not the original a priori version, instead being versions amended during the course of the trial (12).

Protocolization of research is, however, largely limited to the clinical trials setting. The multitude of other types of investigations that appear in scientific journals tend to have a more haphazard, exploratory origin. In these circumstances, decisions about how to focus the analysis, or indeed what hypotheses to test, tend to occur in a free-flowing fashion. In these circumstances, it would be difficult, if not impossible, to reconstruct the tests actually presented in a published article in the context of a study plan. An example of an interesting exception is genome-wide association studies that seek to correlate individual single nucleotide polymorphisms with disease incidence. In these studies, hundreds of thousands of statistical tests are performed, and there are elaborate techniques available to understand the precise statistical implications of performing so many statistical tests and choosing to report the most significant ones (13).

Perverse Incentives in the Modern Medico-scientific Complex

The selective use and misinterpretation of statistical tests have to be understood in the context of the current scientific milieu. In discussing the replicability crisis in scientific research, Collins and Tabak (10) cited poor training in study design as a cause of the frequent failure of scientific studies to be replicated. They also pointed to a modern culture in which scientific findings tend to be exaggerated. This culture is encouraged by a strong feedback loop in which academic institutions and journals play a prominent role (see Figure 1). Investigators perceive

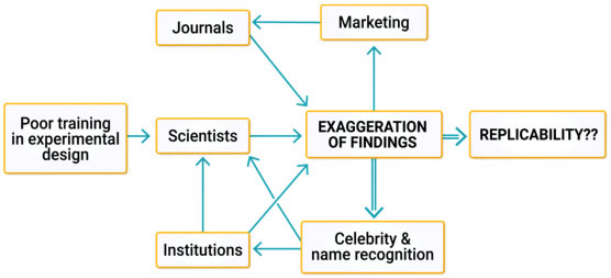

Figure 1. Pressures and incentives on investigators.

that their research is more likely to be published in major journals if the results are emphatic. They are encouraged indirectly in this by public affairs departments in leading academic institutions that seek publicity for fundraising. Resulting lay publicity improves name recognition, which in turn can improve the prospects for academic promotion. Prominent journals are also culpable, seeking preferentially to publish articles that will attract the attention of the lay press. All of these forces operate in a feedback loop that thrives in the modern celebrity culture. What's lost in this milieu is scientific rigor and accountability. Is the use of statistical testing a key ingredient in all this? Well, the culture certainly incentivizes the exaggeration of scientific findings through the selective reporting of statistically significant tests. But in my opinion, there are countless other ways to use data to exaggerate conclusions, and the banning of statistical testing would be unlikely to have much impact on the forces reflected in Figure 1. Scientists are certainly sufficiently creative to find other statistical tools to place their theories in the most favorable light.

The Private Sector, Drug Approvals, and the Food and Drug Administration

Rigid adherence to statistical significance has historically played an especially prominent role in the Food and Drug Administration (FDA) rules governing approvals of new drugs. Traditionally, to be approved for marketing, a drug has to be shown to be significantly superior to a placebo in two pivotal randomized trials, although in recent years, especially in oncology, this standard has been relaxed substantially (14). Having such a rigid, arbitrary rule for something as important as determining whether a new drug should be allowed on the market has troubled many commentators. However, there is a reason that the FDA chose to employ a standard of this nature: for such an important decision, where the sponsor conducts the trials and where the financial ramifications could run into the billions, the pressure to create and present data in the most favorable light is overwhelming. The only realistic way to counter this pressure is to insist on a protocol-driven strategy where the commercial sponsor has no options to manipulate the data, endpoints, and analysis plan in ways that selectively advantage the inferences in favor of drug approval. In short, the purpose of rigid protocolization of the research is to offset the human factor, rather than as a preference for any specific style of statistical analysis.

Reproducibility, Replication, and Transparency

Over the past several years, a broad impetus has developed to try to create a framework for improving the chances that

scientific findings will stand the test of time. Both of the terms *reproducibility* and *replicability* have been used to characterize the notion that the findings of reported studies are likely to be durable (15). If we follow the terminology proposed by the National Academy of Sciences, the term *replicability* should be used for this purpose, and there is broad agreement that it is enhanced by the kind of careful, protocol-driven approaches that I have been discussing that adhere to the scientific method. *Reproducibility*, on the other hand, refers to the ability of an interested scientist who has access to the data to obtain the same conclusions as presented in a published study. A sine qua non of reproducibility is transparency, that is, the availability of the data (and other key details of the methods used). The pursuit of transparency in science has been developing now for a generation and includes initiatives for registering clinical trials (16), guidelines for details needed when reporting clinical (17) and animal studies (18), and, increasingly, requirements by journals that authors publish their raw data either as supplementary material or on publicly accessible databases (19). Ultimately, these initiatives have a strong potential to influence the research environment by discouraging authors from engaging in selective reporting of scientific findings, either by focusing on selective aspects of the data or by selective use of statistical tests or other methods that fail to reflect objectively the strength of evidence in the data.

In 1963, in a celebrated lecture entitled "Is the Scientific Paper a Fraud?" the Nobel laureate Sir Peter Medawar ridiculed the credibility of the style in which research findings are communicated in scientific journals (20). He was not implying that scientists are dishonest. He was recognizing that the style of presenting findings, as a logical, linear narrative, does not recognize the process of discovery, which is a disjointed process of false starts, mistakes, and evolution of opinion as the research progresses. Consequently, the data presented often reflect a highly sanitized version of the data as they truly emerged. Looking back more than half a century, his article appears like a plea for transparency and an early recognition of the reproducibility and replicability crisis. Many factors influence this crisis, but my primary message in this commentary is that the human factor is paramount. The tradition of scientific inquiry is truth seeking, but our instincts tend to lead us astray from this idealized goal in the highly competitive, modern research environment, especially when we are incentivized to present research findings with less dispassionate restraint than we should. Statistical methods play an important role in the scientific process. Ideally, they provide a quantitative framework for characterizing the strength of evidence behind research findings. They need to be used wisely and interpreted from the truth-seeking tradition of science, a task that is more easily said than done. The need for a dispassionate, truth-seeking approach to statistical analyses and their interpretation applies to all available statistical methods, including the use of statistical tests and *P* values.

Notes

Disclaimer: The author has no conflicts of interest.

I am grateful to Dave DeMets and Andrew Vickers for feedback on an early draft of this work.

References

1. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature*. 2019;567(7748):305–307.
2. Wasserstein RL, Lazar NA. The ASA statement on p values: context, process and purpose. *Am Stat*. 2016;70(2):129–133.
3. Rothman KJ. A show of confidence. *N Engl J Med*. 1978;299(24):1362–1363.
4. Trafimow D., Marks M. Editorial. *Basic Appl Soc Psychol*. 2015;37(1):1–2.
5. Gardner MJ, Altman DG. Statistics in medicine: confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J*. 1986; 292(6522):746–750.
6. Trafimow D. Descriptive versus inferential cheating. *Front Psychol*. 2013;4:627.
7. Weinberg CR. It's time to rehabilitate the P value. *Epidemiology*. 2001;12(3): 288–290.
8. McShane BB, Gal D. Statistical significance and the dichotomization of evidence. *J Am Stat Assoc*. 2017;112(519):885–908.
9. Lesaffre E. Superiority, equivalence and non-inferiority trials. *Bull NYU Hosp Jt Dis*. 2008;66(2):150–154.
10. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. *Nature*. 2014;505(7485):612–613.
11. Cook JA, Fergusson DA, Ford I, et al. There is still a place for significance testing in clinical trials. *Clin Trials*. 2019;16(3):223–224.
12. Spence O, Hong K, Uba RO, Doshi P. Availability of study protocols for randomized trials published in high-impact medical journals: a cross-sectional analysis. *Clin Trials*. 2020;17(1):99–105.
13. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B*. 1995;57(1):289–300.
14. Woodcock J. Expediting drug development for serious illness: trade-offs between patient access and certainty. *Clin Trials*. 2018;15(3):230–234.
15. Bollen K. *Reproducibility, Replicability and Generalization in the Social, Behavioral and Economic Sciences*. Alexandria, VA: National Science Foundation; 2015. https://www.nsf.gov/sbe/SBE_Spring_2015_AC_Meeting_Presentations/Bollen_Report_on_Replicability_SubcommitteeMay_2015.pdf. Accessed March 12, 2020.
16. De Angelis C, Drazen JM, Frizelle FA, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *N Engl J Med*. 2004;351(12):1250–1251.
17. Begg CB, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT Statement. *J Am Med Assoc*. 1996; 276(8):637–639.
18. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol*. 2010;8(6):e1000412.
19. PLoSONE. Information for authors. https://journals.plos.org/plosone/s/data-availability. Last updated December 5, 2019, Accessed March 12, 2020.
20. Medawar P. *Is the Scientific Paper a Fraud?* (BBC broadcast) *Listener*;1963. http://www.weizmann.ac.il/mcb/UriAlon/sites/mcb.UriAlon/files/uploads/medawar.pdf. Accessed March 12, 2020.