Clin Orthop Relat Res / DOI 10.1007/s11999-016-4814-3

Clinical Orthopaedics and Related Research®
A Publication of The Association of Bone and Joint Surgeons®

Not the Last Word

Not the Last Word: Inigo Montoya and Statistical Significance

Joseph Bernstein MD

In Rob Reiner's classic film, *The Princess Bride*, Vezzini the Sicilian, Fezzik the giant, and Inigo Montoya have abducted the Princess, but the Man in Black is in hot pursuit and making great gains. Each time Inigo reports to Vezzini that the Man in Black is getting closer, Vezzini replies, "Inconceivable!" Finally, after about the fifth exchange between the two, Inigo replies in exasperation, "You keep using that word—I do not think it means what you think it means."

Sometimes I wish I had Inigo Montoya at my side when I read the results of some scientific studies (or, especially, news accounts thereof). I would love to hear Inigo say, "You keep using that phrase, 'statistical significance'—I do not think it means what you think it means."

Despite its broad use in both the scientific and lay press, or perhaps because of it, the term "statistical significance" is often misused.

Statistical significance is not a statement about clinical importance. A study that shows, for example, that patients with autograft ACL reconstructions return to sports 2 days sooner on average than patients with allografts very well may show statistical significance (especially if there are many subjects in the study). However, shortening a months-long recovery time by a mean of two days has scant practical value, even for an elite athlete.

Most people realize that the term "statistical significance" refers to probabilities and the influence of random chance. Yet even with that understanding, there is room to get it wrong. The p value is not the conditional probability that the null hypothesis is true given what was seen; it is the conditional probability of seeing results at least as extreme as those observed, assuming that the null hypothesis was not false. This distinction is subtle, but important.

An elegant example suggested by Faye Flam [6] (which I have modified a bit here) illustrates the difference. Imagine a swindler wants to sell you a magic bracelet that guarantees you win at poker. Needless to say, this swindler's bracelet is a sham, but like any good con man, the seller says, "Don't take my word for it; let me demonstrate it." This demonstration, is, in a sense, an experiment testing the hypothesis that "the bracelet is magic," with a null hypothesis of "the bracelet has no effect."

The crook deals you five cards. If these five cards contain at most one pair, you might not be impressed. The p value of this observation (the probability of obtaining one pair in a five-card hand without a magic bracelet) is about 0.42. On the other hand, if he

Note from the Editor-in-Chief: *We are pleased to present to readers of Clinical Orthopaedics and Related Research® the next* Not the Last Word. *The goal of this section is to explore timely and controversial issues that affect how orthopaedic surgery is taught, learned, and practiced. We welcome reader feedback on all of our columns and articles; please send your comments to eic@clinorthop.org.*

The author certifies that he, or any member of his immediate family, has no funding or commercial associations (eg, consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article.
All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research®* editors and board members are on file with the publication and can be viewed on request.

J. Bernstein MD (✉)
Department of Orthopaedic Surgery, University of Pennsylvania, 424 Stemmler Hall, Philadelphia, PA 19104, USA
e-mail: orthodoc@uphs.upenn.edu

Published online: 06 April 2016

Not the Last Word

were to deal you three-of-a-kind, you may take notice, for the probability of that event is only about 0.02.

Of course a p value of 0.02 does not mean that the probability is only two out of 100 that the statement, "the bracelet has no effect" is true. Indeed, we have stipulated that the bracelet is a sham, and therefore the probability of that statement is 100%. All the low p value tells you is that you have witnessed an unlikely event. If the event is sufficiently unlikely, one can reasonably reject random chance alone as the cause. The low p value does not mean that all other explanations besides your null should be rejected as well. As Flam points out, if the dealer were to repeatedly deliver a royal flush, each with a p value of about 0.0000016, a rational person probably would draw conclusions from that observation—but "the bracelet is magic" is not apt to be one of them [6].

It is clear by Flam's example that p values can be informative, but they can misinform as well [2]. For that, and related reasons, the journal *Basic and Applied Social Psychology* has banned their use [12].

If I were in charge, I would not ban the p value, but I would ban the word "significant." I would make p-value reporting optional, but allowed only if introduced with the following wording: "The probability of seeing this result on the basis of chance alone is…." This mode of presentation removes the false choice of "it's the null hypothesis or our hypothesis," and liberates the writer and the reader alike from the procrustean tyranny of the 0.05 threshold.

And as long as I am imposing the use of certain phrases in the manuscript, I would like to insist that the Methods section also contain the following statement: "The sample size employed in this study was chosen on the following assumptions…." The authors would use that as a springboard to state the effect size they considered important, the expected variance in the data and in turn the necessary sample size for the study. This statement could only be made if a sample size calculation and power analysis were undertaken—and that is the point. (I should note that, in the main, my prescriptions are already followed here at *CORR*® implicitly; though I should also note that as a deputy editor, I neither establish nor enforce journal policies).

In recent years, there has been more awareness of sample size and power, especially regarding studies that did not find statistically significant differences. But the concept is also relevant to studies that do find such differences. That is because if a study has too few subjects, it is likely to attain statistical significance only if the effect found is, by chance, larger than the true difference: The so-called winner's curse [4]. The "curse" is that findings from studies that lack power, yet attain significance, likely overstate the magnitude of the effect (even when they reach the correct conclusion about the existence of some, albeit smaller, effect).

Consider, for example, a study that attempts to measure the incidence of infection after immediate versus delayed closure of open fractures. Let's stipulate that the yet-undiscovered truth is that delayed closure has a 15% lower incidence of infection. In the case where the study is powered (that is, had a sample size large enough) to detect only a 25% difference or larger, the discovery of a 15% difference—the truth!—will likely fail to attain "significance"; and owing to the lack of power, the official result of the investigation must be "we found no differences."

One can easily imagine that if 10 researchers were to measure the incidence of infection, each with underpowered studies, there might be nine studies that find rate differences in the range of 10–20%. Because of the low power, the p values of these otherwise positive results exceed 0.05. Given this failure to find significant differences, the researchers may elect to discard their findings, or if they do submit them to a journal, they are apt

Not the Last Word

to face rejection, as there can be a "positive-outcome bias" present during peer review [5].

A tenth study that randomly found a 30% difference, indeed might cross the threshold of significance (because of the large effect size). This 30% effect will be enshrined in the literature. My guess is that casual readers will hone in on this stated 30% difference, and not the smallest possible difference that would attain significance–the gap between the edges of the confidence intervals. Because of this tendency, casual readers will walk away thinking that delayed closure has a lower rate of infection, which is true; but they'd also think that the difference is 30%, which it is not.

Given the tendency of underpowered studies to overstate the true magnitude of the outcome, and given that many studies are underpowered [1], replication—repeating the experiment to see if the same result will be found—often fails [3]. These replication studies issue an implicit rebuke to the original scientists: "Your findings—I don't think they mean what you think they mean." One good way to be spared that rebuke is to make sure that the sample size is adequate. More generally, we should stop thinking of "$p < 0.05$" as a binary truth detector. As Flam teaches, the p value is a useful heuristic, but little more.

Andrew Gelman PhD

Professor, Department of Statistics and Department of Political Science

Columbia University

I agree with just about everything in this article. As I have written elsewhere [7], ultimately the problem is not with p values, but with null-hypothesis significance testing, that parody of falsificationism in which straw-man null hypothesis A is rejected and this is taken as evidence in favor of preferred alternative B.

My only experience with orthopaedics comes from a couple of broken bones a few years ago—but it is my impression that null-hypothesis significance testing has particular problems in medical research because of the typical combination of small effects and small sample sizes. Effects are small, not because treatments are so ineffective, but rather because medicine works so well. It is unethical to compare a new treatment against a placebo if some effective alternative exists, and in a mature research field, improvements will typically be incremental. Dramatic new treatments do appear from time to time, but most of the hundreds of thousands of medical research papers published every year will necessarily be testing small effects. And sample sizes will be small too, for reasons of time, cost, and practicality.

Nonreplication of statistically significant findings and overestimation of effect sizes (the statistical significance filter, or what Bernstein calls "the winner's curse," or what my colleagues and I have called Type M ("magnitude") errors [8]; are prevalent in studies with small effects, small samples, and highly variable outcomes—which are par for the course in medical research.

Moving forward, it seems like a good idea for researchers to publish all their raw data on successful and unsuccessful studies, with the aim of future analysis of combined datasets. Lowering the bar to publication would, we hope, reduce the incentives for researchers to search for statistical significance in their data and reduce the selection bias that is currently inherent in any inferences, meta- or otherwise, from published data.

David Trafimow

Department of Psychology, New Mexico State University

Editor, *Basic and Applied Social Psychology*

Not the Last Word

I agree with much of what Dr. Bernstein says; p values do not give the probability of the null hypothesis given the finding. It also does not give the probability of replication, the extent to which the finding can be generalized, the usefulness or reliability of the data, the size of the effect, how well the data represent the population, or practically anything else of value to researchers. Using p values to conclude anything other than the tautology that the finding is unlikely given the null hypothesis, constitutes misuse of p values. I disagree with both Dr. Bernstein's recommendation that reporting p values should be optional and his wording in that direction: "The probability of seeing this result on the basis of chance alone is…." But p values are not the probability of obtaining the finding by chance. The computation of the probability of the finding by chance depends on the unknown population parameter to be estimated by the obtained sample statistic; therefore, there generally is no way to compute the probability of the finding due to chance. Unfortunately, p values only provide a hypothetical value under an unlikely null hypothesis. The null hypothesis is almost never exactly true because it specifies a particular value, and there is an infinitude of possible values. The difference between population means in an experimental and control condition could be 0.0000000 00001, −0.1012334, or anything else and it is unlikely to be exactly zero. As the null hypothesis is almost always false a priori, p values will not represent the probability of the finding by chance.

Now, let us consider Bernstein's broader point. He states that despite the problems with using p values, it would be better to let its use remain optional, rather than banning p values as we did in *Basic and Applied Social Psychology* [10]. However, I already tried this [11]. Before instituting the ban, I discouraged the use of p values, but chose to keep its use optional. The result was that every empirical first submission I received in the following year contained p values! Clearly, an optional approach did not work, thereby necessitating stronger medicine, namely the ban.

Many believe that the ban went too far, even though no one can state a useful conclusion that p values allow researchers to reach in a logically valid way. For those who think the ban went too far, consider a simple question. Suppose that I had not instituted the ban. Would researchers in different areas of science, including Bernstein in orthopedics, even be discussing whether the use of p values are justified? If you believe that the answer is "no," as I do, then it is difficult to avoid concluding that the ban was a good thing.

Alex Reinhart BSc

PhD student in statistics at Carnegie Mellon University

Author of *Statistics Done Wrong*

Dr. Bernstein is right to emphasize the importance of adequate sample sizes. Underpowered studies with inadequate sample sizes are often overestimates when statistically significant and meaningless when insignificant. But we cannot deny one of the root causes of this problem—getting an adequate sample size can be very difficult.

Consider, for example, a study evaluating a new surgical technique aimed at reducing side effect rates for a complicated procedure. The current gold standard procedure has roughly a 20% side effect rate, and we hope the new technique will reduce that to 10%—a 50% improvement. We propose a randomized controlled trial with two groups, one receiving the new treatment and one receiving the old treatment, and need to know the necessary sample size. How many patients, and how many procedures, must we obtain if we are to have an 80% chance of getting a statistically significant result?

The answer is 400. We need two hundred patients to undergo each procedure, which may take years and cost millions of dollars. Smaller studies would be problematic. If we used only

Not the Last Word

100 patients, we would only achieve statistical significance about one-third of the time, and our significant results would, on average, overestimate the benefits of the new treatment by 40%.

Of course, we are not always able to afford adequate sample sizes, and we are not always lucky enough to get statistically significant results. Sometimes we obtain $p > 0.05$ and admit defeat. There is no evidence that the treatment has an effect. But this is not evidence that the treatment has no effect! The treatment may have a dramatic effect that we had inadequate power to detect.

One example is legal right-turn-on-red in the United States [9]. Right turn on red was not widespread until the 1970s, when the oil crisis and moves to improve efficiency led states to consider it. In Virginia, the Department of Highways and Transportation hired a consultant to do a before-and-after test of 20 intersections to determine if the change caused an increase in accidents. Accidents did increase, but the consultant found no statistically significant change, and the change was deemed safe.

Unfortunately, statistical insignificance did not mean there was no difference. Later research, with larger sample sizes, found that right-turn crashes increased by about 23%. Dr. Bernstein noted that statistical significance does not imply practical value, and the reverse is true as well. Statistical insignificance cannot be turned into practical insignificance if the sample size is inadequate.

Physicians and statisticians must keep this problem in mind when interpreting negative results. As Dr. Bernstein pointed out, many medical trials are underpowered, and we must remember that in underpowered studies, a negative result may also be consistent with a large effect. We should be more concerned with the confidence interval—the range of effects consistent with our data—instead of narrowly focusing on statistical significance. Significance is a tool to aid our research, not its end goal.

References

1. Abdullah L, Davis DE, Fabricant PD, Baldwin K, Namdari S. Is there truly "no significant difference"? Underpowered randomized controlled trials in the orthopaedic literature. *J Bone Joint Surg Am.* 2015;97: 2068–2073.
2. American Statistical Association. American Statistical Association releases statement on statistical significance and p-values. Available at: https://www.amstat.org/newsroom/pressreleases/P-ValueStatement.pdf. Accessed March 24, 2016.
3. Baker M. First results from psychology's largest reproducibility test. Available at: http://www.nature.com/news/first-results-from-psychology-s-largest-reproducibility-test-1.17433. Accessed March 3, 2013.
4. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR. Power failure: Why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013;14:365–376.
5. Emerson GB, Warme WJ, Wolf FM, Heckman JD, Brand RA, Leopold SS. Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial. *Arch Intern Med*. 2010;170: 1934–1939.
6. Flam F. Lies, damned lies and physics. Available at: http://www.bloombergview.com/articles/2015-12-30/lies-damned-lies-and-physics. Accessed March 3, 2016.
7. Gelman A. Statistical Modeling, Causal Inference, and Social Science. Confirmationist and falsificationist paradigms of science. Available at: http://andrewgelman.com/2014/09/05/confirmationist-falsificationist-paradigms-science/. Accessed March 22, 2016.
8. Gelman A, Carlin JB. Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspect Psychol Sci*. 2014;9:641–651.
9. Hauer E. The harm done by tests of significance. *Accid Anal Prev*. 2004; 36:495–500.
10. Trafimow D. Editorial. *Basic Appl Soc Psych*. 2014;36:1–2.
11. Trafimow D, Marks M. Editorial. *Basic Appl Soc Psych*. 2015;38:1–2.
12. Woolston C. Psychology journal bans P values. Available at: http://www.nature.com/news/psychology-journal-bans-p-values-1.17001. Accessed March 3, 2016.