Global Spine Journal
2019, Vol. 9(3) 357-359
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2192568219838538
journals.sagepub.com/home/gsj

# *P*-Value Worship: Is the Idol Significant?

**Joseph R. Dettori, PhD[1], Daniel C. Norvell, PhD[1], and Jens R. Chapman, MD[2]**

> We think that idols are bad things, but that is almost never the case. The greater the good, the more likely we are to expect that it can satisfy our deepest needs and hopes. Anything can serve as a counterfeit god, especially the very best things in life.
>
> —Timothy Keller[1]

## Introduction

We have observed the tendency to overemphasize the *P*-value among beginning and seasoned clinical researchers in the spine community. It has become too important at the expense of other factors necessary to derive sound scientific inferences such as evidential justification of the research hypothesis, study design, quality of measurements, study integrity, and validity of assumptions that underlie the data analysis. As a result, clinical researchers, rather than looking at their data critically, tend to focus on statistical testing and *P*-values. It is what satisfies their research needs and hopes. Comments too frequently heard include ones similar to the following: "Here is the data I collected. What are the *P*-values?" "Is there significance in my results?" "These results ($P = .003$) look highly significant!" "My *P*-value of .06 shows a trend toward statistical significance." These comments betray the misunderstanding of and overdependence on the almighty *P*-value among clinical researchers.

However, it is not just clinical researchers who are guilty of *P*-value worship; journal reviewers and editors are culpable as well. Consider some partial data describing baseline characteristics of 2 groups receiving spine fusion (Table 1). One group received BMP (bone morphogenetic protein) and the other did not. The purpose of this type of table is to describe the sample in a way that indicates the similarities and imbalances between groups. It is not used to make inference to a population.[2] In this example, one can see that the groups are similar in terms of age and sex, but different in terms of site of surgery. The data are clear; the table accomplishes its purpose. Yet the journal reviewers and editors did not accept the table as presented, but insisted on including *P*-values for resubmission. Again, a misunderstanding of and overdependence on the *P*-value.

## Understanding the Idol

How is this idol defined in the first place? What is a *P*-value? The American Statistical Association (ASA) defines it this way: "The probability under a specified statistical model that a statistical summary of the data (eg, the sample mean difference between the 2 compared groups) would be equal to or more extreme than its observed value."[3] A more simple way of restating the definition is the following: the *P*-value is the likelihood of obtaining one's data (or more extreme) if the null hypothesis and all other assumptions used to compute the *P*-value are true.[4,5] Table 2 summarizes select principles by the ASA with respect to the *P*-value.

## The Hazards of *P*-Value Worship

> **Hazard 1:** Believing that a statistically significant result indicates a clinically important finding.

It is important to realize that the *P*-value is not the same as the effect size. The effect size is the clinical or practical difference between groups, which is what we are most interested in. A *P*-value is particularly sensitive to the sample size such that a very small difference between groups can be statistically significant when the study is large, yet without clinical importance.[6]

Consider the following data (Table 3) from the National Inpatient Sample database comparing the mean age of octogenarians and nonagenarians receiving decompression, fusion, or discectomy.[7] There is a statistically significant difference in age among the 3 groups, *P*-value <.001. This significance is driven by the very large sample size and not the small mean

[1] Spectrum Research, Inc, Steilacoom, WA, USA
[2] Swedish Medical Center, Seattle, WA, USA

**Corresponding Author:**
Joseph R. Dettori, Spectrum Research, Inc, Steilacoom, WA 98388, USA.
Email: joe@specri.com

**Table 1.** Baseline Characteristics Describing Similarities and Imbalances in Patients Receiving Spine Fusion With and Without Bone Morphogenetic Protein (BMP).

| Characteristics | No BMP | BMP |
|---|---|---|
| Age at surgery (years), mean ± SD | 55.4 ± 13.4 | 55.4 ± 13.3 |
| Sex | | |
| Men | 42.4% | 42.3% |
| Site of surgery | | |
| Cervical | 13.6% | 56.2% |
| Dorsolumbar | 2.5% | 2.4% |
| Lumbosacral | 83.8% | 41.3% |
| Unspecified | 0.2% | 0.2% |

**Table 2.** The 2016 Statement by the American Statistical Association on Statistical Significance and *P*-Values.

1. *P*-values can indicate how incompatible the data are with a specified statistical model
2. *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone
3. Scientific conclusions and business or policy decisions should not be based only on whether a *P*-value passes a specific threshold
4. Proper inference requires full reporting and transparency
5. A *P*-value, or statistical significance, does not measure the size of an effect or the importance of a result
6. By itself, a *P*-value does not provide a good measure of evidence regarding a model or hypothesis

**Table 3.** Example of the Small but Statistically Significant Difference in Mean Age From 3 Treatments Using a Large Sample From the National Inpatient Sample Database.

| Variables | Decompression (n = 113 267) | Fusion (n = 60 345) | Discectomy (n = 50 740) | *P*-Value |
|---|---|---|---|---|
| Age, mean (SE) | 83.2 (0.02) | 82.8 (0.03) | 83.0 (0.03) | <.001 |

differences in age. Although these are statistically different with respect to age, the difference is not clinically important. Practically, these 3 populations have the same mean age.

**Hazard 2:** Trusting that a nonsignificant test (*P*-value >.05) means that the 2 groups are the same and the null hypothesis should be accepted.

This is a common error from gazing too longingly at the *P*-value alone. The truth of the matter is that the *P*-value may be inflated as a result of a small sample size or a large random error. As a result, whether the 2 groups are the same may still be unknown.

**Hazard 3:** Supposing that a "highly significant" *P*-value indicates a large effect size.

By itself, a *P*-value does not inform us about the size of the difference between groups. Consider 2 studies of different sample sizes each comparing nonunion risk following lumbar fusion (Table 4). Both studies evaluate the same 2 surgical techniques. One study evaluates 100 patients, the other 2000. The results of the 2 studies produce essentially the same *P*-value, both statistically significant. Yet the difference in the nonunion risk is much higher in Study 1. The *P*-value does not tell us the size of the difference.

**Table 4.** Example of Nonunion Assessed in 2 Different Studies With Varying Treatment Differences (Effect Size) but Equal *P*-Values.

| | Surgery A | Surgery B | Difference | *P*-Value |
|---|---|---|---|---|
| Study 1 (n = 100) | 4/50 (8%) | 14/50 (28%) | 20% | .017 |
| Study 2 (n = 2000) | 80/920 (8%) | 112/888 (11%) | 3% | .015 |

**Hazard 4:** Clinging to the idea that a *P*-value close to but not quite statistically significant (eg, .06) supports a "trend" toward statistical significance.

The *Oxford English Dictionary* defines the noun trend as "a general direction in which something is changing or developing."[8] This definition is used in both scientific and nonscientific literature. Applying the term *trend* to a single *P*-value that is close to but >.05 is inappropriate and betrays a misunderstanding of the *P*-value. It also is as logical as describing a *P*-value close to but <.05 (eg, .04) as supporting a trend toward nonsignificance.[9,10]

## Suggestions on How to Replace *P*-Value Worship

1. When describing the baseline balances or imbalances between groups, use numerical differences and sample summary statistics, not *P*-values. Focus on differences of clinical importance.
2. Present and interpret effect sizes and the associated confidence intervals rather than focusing on the *P*-value. Remember that *P*-values are most meaningful in the context of clear, a priori hypotheses that support the main conclusions of a manuscript.
3. Define and interpret effect measures that are clinically relevant. When possible, report the absolute differences rather than the relative differences.

## Summary

- *P*-value worship is not uncommon among clinical researchers in the spine community. It has become too important at the expense of other factors necessary to derive sound scientific inferences.
- There are a number of hazards surrounding *P*-value worship that may lead to spurious interpretations of research data.
- Clinical researchers should focus on presenting and interpreting clinically relevant effect sizes and their

associated confidence intervals, using *P*-values only as one of many tools in the statistical toolbox.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Support for this work was provided by Spectrum Research, Inc, with funding from AOSpine.

## References

1. Keller T. *Counterfeit Gods: The Empty Promises of Money, Sex, and Power, and the Only Hope That Matters*. New York, NY: Penguin Books; 2009.
2. Cummings P, Rivara FP. Reporting statistical information in medical journal articles. *Arch Pediatr Adolesc Med*. 2003;157:321-324.
3. Wasserstein RL, Lazar NA. The ASA's statement on *P*-values: context, process, and purpose. *Am Stat*. 2016;70:129-133.
4. Karpen SC. *P* value problems. *Am J Pharm Educ*. 2017;81:6570.
5. Tanha K, Mohammadi N, Janani L. *P*-value: what is and what is not. *Med J Islam Repub Iran*. 2017;31:65.
6. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31:337-350.
7. Drazin D, Lagman C, Bhargava S, Nuno M, Kim TT, Johnson JP. National trends following decompression, discectomy, and fusion in octogenarians and nonagenarians. *Acta Neurochir (Wien)*. 2017;159:517-525.
8. *Oxford English Dictionary*. Trend, https://en.oxforddictionaries.com/definition/trend. Accessed February 11, 2019.
9. Mansfield L. The reading, writing, and arithmetic of the medical literature, part 2: critical evaluation of statistical reporting. *Ann Allergy Asthma Immunol*. 2005;95:315-321.
10. Gibbs NM, Gibbs SV. Misuse of "trend" to describe "almost significant" differences in anaesthesia research. *Br J Anaesth*. 2015;115:337-339.