Published in final edited form as:

Am J Robot Surg. 2015 December ; 2(1): 53–56. doi:10.1166/ajrs.2015.1017.

The Value of the *P* Value

Dinesh Vyas^{*}, Archana Balakrishnan, and Arpita Vyas

College of Human Medicine, Michigan State University, East Lansing, MI 48824, USA

Abstract

Recently, the discussion on the implications of irreproducibility in the sciences has been brought into the spotlight. This topic has been discussed for years in the literature. A multitude of reasons have been attributed to this issue; one commonly labeled culprit is the overuse of the *p* value as a determinant of significance by the scientific community. Both scientists and statisticians have questioned the use of null hypothesis testing as the basis of scientific analysis. This survey of the current issues at hand in irreproducibility in research emphasizes potential causes of the issue, impacts that this can have for drug development and efforts been taken to increase transparency of findings in research.

Keywords

p-Value; Data Analysis; Statistics; *P* Value; Statistical Modeling; Irreproducibility in Research

THE VALUE OF *P* VALUE

Researchers use a variety of statistical analytic methods to better quantify the validity and reproducibility of their experiments. A popular method of determining statistical significance is through the use of the *p*-value, a measure of how well experimental data rejects the null hypothesis. For years, the debate over the value of *p*-value has been discussed in the literature [1–4]. Critics argue that *p*-value creates an artificial metric of establishing significance in a data set. Recently, Leek and Peng published a commentary in *Nature* that discusses limitations of null hypothesis testing.⁵ Rather than focusing their criticism on *p*-value, the authors argue that data analysis, as a whole, must be more stringently regulated. The null hypothesis is widely used in both the basic sciences and in clinical/translational research, a field in which experimental findings strongly influence the treatment of patients. Therefore, it is vital to understand how experimental design, data analysis, and statistical testing may be optimized to generate studies that more effectively improve human health.

^{*} Author to whom correspondence should be addressed. vyasd@msu.edu.

Author Contributions

Dinesh Vyas and Archana collected the material and wrote the manuscript. Dinesh Vyas and Arpita discussed the topic, and Dinesh Vyas supervised the publication of this commentary.

Core Tip

This article provides an overview of the current issues surrounding irreproducibility in research and discusses initiatives being done to rectify this. In this article, we are expanding upon a recently published commentary in *Nature*.

Typical Research Methodology

In most research studies, there is a classical progression of experimental steps (Fig. 1). First, a problem is identified through analysis of evidence seen in prior experimentation and literature. A hypothesis is proposed to address the problem, and experiments are devised to test the hypothesis.⁶ After the generation of data and appropriate statistical filtering, a p -value test is applied to suggest whether the findings are “significant” or “not significant.” P -value is popularly used as a measure of significance because it represents the probability of an event occurring solely by chance or random error. Therefore, a high p -value indicates that there is a higher probability that the event could be explained by chance. In other words, a higher p -value means there is a higher probability that the null hypothesis should be accepted instead of rejected.⁷ Cut off values are arbitrarily assigned for evaluation of significance of the p -value. By definition, a p -value larger than the cutoff value indicates that the event is more likely to have occurred due to chance.⁸ Unfortunately, many scientists misinterpret the p -value as a simple method to determine significance while ignoring other critical factors of a data set that can drastically impact the p -value such as the magnitude of the association and the sample size. These components must be recognized and evaluated for a data set to ensure that the p -value is an accurate measure of deviation from the null hypothesis. The greatest impact on a data set is not the calculation of p value, but rather the manners in which the raw data is altered through cleaning and statistical modeling.⁹ Typically, data is purged of outliers using various statistical algorithms to generate a “clean” dataset. At this stage, the use of null hypothesis testing may generate a p -value that may falsely represent the significance of the study.^{10–12}

Implications of Misuse of P Value

An underappreciated but critical aspect of analysis of research findings is the correct interpretation of a calculated p value in the context of the sample set of data. There are many factors that define the reliability of a sample set including potential bias and sample size.¹³ For example, having a smaller sample size is less likely to represent the population as a whole. This makes the p value generated from this data less likely to represent the population and therefore less likely to be replicated if the experiment was repeated. Additionally, p value is limited in that it does not correlate to or provide information on the magnitude of an effect.¹⁴

In basic science research where the use of animal models is common, careful planning of experiments is essential to minimize the number of animals required to generate usable data. However, this factor must be balanced with the risk of drawing incorrect conclusions due to an inadequate sample size. Careful selection of animal models and the proper use of controls are critical to picking the appropriate type of data analysis.¹⁵ Another issue that impacts statistical analysis is the inclusion of replicates. There are two types of experimental replicates: biologic and technical. A technical replicate assays a single sample multiple times and serves as an internal control for the experiment. Whereas a biological replicate is when different samples are assayed.¹⁶ Replicates are not independent tests of the hypothesis, and therefore should not be used to calculate p values. However, they are often incorrectly included in statistical calculations.¹⁷

In the clinical sciences, the results of studies often guide clinicians in treatment and can have a more immediate impact on human health. All clinical studies have some type of bias or effects that can alter conclusions away from the truth.¹⁸ It is also important to look at the statistical power of the study. For example, pharmaceutical companies may include hundreds of patients in a clinical trial; this leads to the ability to detect minor differences that are very statistically significant that may not represent a clinical difference.^{19, 20}

A Journal that Publishes Without p Value

The use of p value has been present in literature for years and is a central tenant of evidence-based scientific methods.²¹ The raging debate over the usage of the p -value in science encouraged the journal *Basic and Applied Social Psychology* to ban the use of null hypothesis significance testing analysis in its future publications.²² This journal aims to publish research that integrates basic science research with practical real world applications. Declaring that null hypothesis testing is invalid, the journal now calls for more descriptive statistics to be used like effect size, a measure of the strength of a phenomenon.²³ This is the first journal to institute a ban of this type. This decision has been met with applause from some and criticism from others in the scientific community who believe that this bold move will have little effect on the quality of scientific literature.

Impacts to Drug Development

A common result of the systematic tailoring of data to achieve a p value that demonstrates “significant” is that many studies are irreproducible.²⁴ Validation of scientific data through reproduction of experiments is fundamental to the scientific method. Pharmaceutical companies constantly scan the literature to identify new molecular targets or compounds that show high potential to further study. These targets are often identified and published by academic labs. Amgen, a biotechnology company based in California, identified fifty-three promising cancer studies over the course of ten years from high profile journals. However, only 6 of these studies, a mere 11%, were both reproducible and demonstrated continued promise as a novel target.²⁵ Other pharmaceutical companies have also reported difficulties with validating promising studies, a problem that drastically hinders the progress of therapeutic development.²⁶ This lack of reproducibility has been attributed to a variety of factors, including the chronic misuse and lack of understanding of what the p value represents.²⁷

Future Directions

To help eliminate some of the issues derived from poor experimental design, students-in-training along with experienced research scientists must learn to better critically scrutinize experimental design and identify flaws in published studies. The National Institutes of Health (NIH) is currently developing a training module for postdoctoral students that emphasize the importance of good experimental design.²⁸ The development of similar initiatives in research centers and universities will be positive steps to improve reproducibility of experiments. Publishers of peer-reviewed research also can play a role in improving the reliability and quality of research by enforcing stricter statistical guidelines. Already, high impact journals such as *Nature* and *Science* have made changes by eliminating a word count for the methods section and calling for increased transparency of data

handling.^{29, 30} These steps are merely the beginning, but provide hope that the scientific community will begin to recognize the dire need for an improvement in the reliability and reproducibility of research.

The scientific community has urged the push for transparency in research presentation. For example should be a continued push to eliminate the use of phrases such as “representative data” and “data not shown,” but rather show the primary data and create a culture of transparency.³¹ Reporting of other statistical values such as confidence interval can help provide a more complete understanding of the statistics by going beyond the scope of a p -value to provide the magnitude and accuracy of an effect.³²

The p value is inherently a valuable statistical measure. However, the manner in which the use of the p value has morphed into a sole determinant of significance has led to an epidemic of irreproducible, vastly over-interpreted research. A complete reform of the experimental process is needed to accelerate the production of scientific research that can more directly and expediently be applied to the treatment and improvement of human health. The actions of the NIH and high profile journals help push in the direction of achieving more research transparency. However, this is only the beginning. Reform of the experimental process must take place at all levels of the scientific ladder, from those students who are first learning about the scientific method to experienced researchers currently training future generations of scientists.

REFERENCES

1. Evans SJ, Mills P, Dawson J. The end of the p value? Br. Heart J. 1988; 60(3):177–180. [PMID: 3052552 PMCID: Pmc1216550]. [PubMed: 3052552]
2. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann. Intern. Med. 1999; 130(12):995–1004. [PMID: 10383371]. [PubMed: 10383371]
3. Dixon P. The p -value fallacy and how to avoid it. Can. J. Exp. Psychol. 2003; 57(3):189–202. [PMID: 14596477]. [PubMed: 14596477]
4. Mogie M. In support of null hypothesis significance testing. Proc. Biol. Sci. 2004; 271(Suppl 3):S82–S84. [PMID: 15101426 PMCID: Pmc1810001]. [PubMed: 15101426]
5. Leek JT, Peng RD. Statistics: P values are just the tip of the iceberg. Nature. 2015; 520(7549):612. [PMID: 25925460]. [PubMed: 25925460]
6. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians: 1. Hypothesis testing. Cmaj. 1995; 152(1):27–32. [PMID: 7804919 PMCID: Pmc1337490]. [PubMed: 7804919]
7. Dorey F. In brief: The P value: What is it and what does it tell you? Clin. Orthop. Relat. Res. :2297–2298.
8. Biau DJ, Jolles BM, Porcher R. P value and the theory of hypothesis testing: An explanation for new researchers. Clin. Orthop. Relat. Res. 2010; 468(3):885–892. [PMID: 19921345 PMCID: Pmc2816758]. [PubMed: 19921345]
9. Motulsky HJ. Common misconceptions about data analysis and statistics. Pharmacol. Res. Perspect. 2015; 3(1):e00093. [PMID: 25692012 PMCID: Pmc4317225]. [PubMed: 25692012]
10. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p -hacking in science. PLoS Biol. 2015; 13(3):e1002106. [PMID: 25768323 PMCID: Pmc4359000]. [PubMed: 25768323]
11. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 2011; 22(11):1359–1366. [PMID: 22006061]. [PubMed: 22006061]

12. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p -values. *R Soc. Open. Sci.* 2014; 1(3):140216. [PMID: 26064558 PMCID: Pmc4448847]. [PubMed: 26064558]
13. Krzywinski M, Altman N. Points of significance: Importance of being uncertain. *Nat. Methods.* 2013; 10(9):809–810. [PMID: 24143821]. [PubMed: 24143821]
14. Nuzzo R. Scientific method: Statistical errors. *Nature. England.* 2014:150–152.
15. Johnson PD, Besselsen DG. Practical aspects of experimental design in animal research. *Ilar j.* 2002; 43(4):202–206. [PMID: 12391395]. [PubMed: 12391395]
16. Blainey P, Krzywinski M, Altman N. Points of significance: Replication. *Nat Methods.* 2014; 11(9):879–880. [PMID: 25317452]. [PubMed: 25317452]
17. Vaux DL, Fidler F, Cumming G. Replicates and repeats—what is the difference and is it significant? A brief discussion of statistics and experimental design. *EMBO Rep.* 2012; 13(4):291–296. [PMID: 22421999 PMCID: Pmc3321166]. [PubMed: 22421999]
18. Lambert J. Statistics in brief: How to assess bias in clinical studies? *Clin. Orthop. Relat. Res.* 2011; 469(6):1794–1796. [PMID: 20809163 PMCID: Pmc3094617]. [PubMed: 20809163]
19. Hochster HS. The power of “ p ”: On overpowered clinical trials and “positive” results. *Gastrointest Cancer Res.* 2008; 2(2):108–109. [PMID: 19259305 PMCID: Pmc2630828]. [PubMed: 19259305]
20. Fernandes-Taylor S, Hyun JK, Reeder RN, Harris AH. Common statistical and research design problems in manuscripts submitted to high-impact medical journals. *BMC Res. Notes.* 2011; 4:304. [PMID: 21854631 PMCID: Pmc3224575]. [PubMed: 21854631]
21. Greenhalgh T, Howick J, Maskrey N. Evidence based medicine: A movement in crisis? *Bmj.* 2014; 348:g3725. [PMID: 24927763 PMCID: Pmc4056639]. [PubMed: 24927763]
22. Trafimow D, Marks M. Editorial. *Basic and Applied Social Psychology.* 2015; 37(1)
23. Trafimow D. Editorial. *Basic and Applied Social Psychology.* 2014; 36(1):1–2.
24. Freedman LP, Gibson MC, Ethier SP, Soule HR, Neve RM, Reid YA. Reproducibility: Changing the policies and culture of cell line authentication. *Nat. Methods.* 2015; 12(6):493–497. [PMID: 26020501]. [PubMed: 26020501]
25. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. *Nature.* 2012; 483(7391):531–533. [PMID: 22460880]. [PubMed: 22460880]
26. Prinz F, Schlange T, Asadullah K. Believe it or not: How much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov. England.* 2011:712.
27. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. *Nat Methods.* 2015; 12(3):179–185. [PMID: 25719825]. [PubMed: 25719825]
28. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. *Nature.* 2014; 505(7485): 612–613. [PMID: 24482835 PMCID: Pmc4058759]. [PubMed: 24482835]
29. Nature Publishing Group. Announcement: Reducing our irreproducibility. *Nature.* 2013; 496(7446)
30. McNutt M. Reproducibility. *Science United States.* 2014; 229
31. Pulverer B. Significant statistics. *EMBO Rep. England.* 2012:280.
32. Singh AK, Kelley K, Agarwal R. Interpreting results of clinical trials: A conceptual framework. *Clin. J. Am. Soc. Nephrol. United States.* 2008:1246–1252.

Figure 1.
A pictorial representation of data processing from collection of data to final statistical analysis.