“Petite” p value: A Researchers’ Dream! Readers, Beware of the Pit ...

Sharada Mailankody¹, Jyoti Bajpai², Sudeep Gupta³

ABSTRACT

“What is the magic word for publication?” “The p value $<0.05\%$ ” seems like an automatic answer from the researchers. Is that right? Let us take a quick peek at this seemingly simple gate-pass to publications!

Keywords: Misuse, p value, Pitfalls, Statistics.

Indian Journal of Critical Care Medicine (2020); 10.5005/jp-journals-10071-23399

INTRODUCTION

Although introduced by Pearson in 1900, the concept of p value is central to understanding medical and other scientific literature even today.¹ Though the definition is taught from the days of undergraduate statistics courses, the exact meaning and interpretation remain elusive to many of the doctors. The “statistical methods” part of the research papers tends to intimidate the casual readers, forcing them to read the authors’ conclusions rather than interpreting the results themselves, which may not be an ideal situation, given the current publication scenario.

For a clear understanding of the “ p ” value, we need to remind ourselves that p value is based on the null hypothesis. It is the probability that the given statistical summary (with the specific statistical model) is equal to or more extreme than the observed value.^{2,3} Put simply, the p value represents the probability of accepting the null hypothesis in a given statistical model. Thus, it is a measure of incompatibility of the study principle with the null hypothesis. Hence, the smaller the p value, the less the chance that the null hypothesis is true and thereby more the probability of “statistical significance” of the study hypothesis. The caveat to this is that the statistical model is created based on a set of assumptions, hence the so-called significance of p value holds good only if these assumptions are valid.² The p value does not measure the truth of the null hypothesis or the effect of randomness on the hypothesis, instead pertains to the data relating to the null hypothesis. The actual numerical p value is dependent on the sample size and measurement techniques. Even a very trivial effect may produce a highly “significant” p value when a huge sample size is used or measurements are precise.²

Knowing this alone is not sufficient to critically analyze a paper. We also need to understand the subtle nuances hidden in the p value. Misinterpretation, overtrusting, and misuse of p values are major problems with the use of p value.⁴ Misinterpretation happens when readers try to relate the p value to the study hypothesis rather than the null hypothesis, or the researchers report their findings based on this premise. Smaller p values are automatically trusted to be associated with a more valid study, leading to ‘overtrust’ on p values. In the quest for a perfect p value, fully transparent reporting may be hindered. Statistical significance need not translate always to clinical significance. However, though p value *per se* implies only statistical significance and not the size or clinical import of an effect, many decisions and guidelines are made/

¹Department of Medical Oncology, Kasturba Medical College, Manipal Academy of Higher Education, Manipal, Karnataka, India

^{2,3}Department of Medical Oncology, Tata Memorial Center, Mumbai, Maharashtra, India; Homi Bhabha National Institute (HBNI), Mumbai, Maharashtra, India

Corresponding Author: Jyoti Bajpai, Department of Medical Oncology, Tata Memorial Center, Mumbai, Maharashtra, India; Homi Bhabha National Institute (HBNI), Mumbai, Maharashtra, India, Phone: +91 22 24177827, e-mail: dr_jyotibajpai@yahoo.co.in

How to cite this article: Mailankody S, Bajpai J, Gupta S. “Petite” p value: A Researchers’ Dream! Readers, Beware of the Pit *Indian J Crit Care Med* 2020;24(Suppl 3):S140–S141.

Source of support: Nil

Conflict of interest: None

changed on the basis of the reported p values; causing ‘misuse’ of p value.⁴ Instead the focus should be on the context of the research, taking the study as a whole. It is not practical to expect “yes–no” answers from all the research papers that we read; acknowledging this gray area in medicine will help in clear understanding of the presented data. There may not be an easy answer to some of the questions.

Selective reporting of significant p values is also a problem. Researchers studied 1,000 abstracts and 100 full-text articles from MEDLINE and PMC databases to qualify the use of p value in biomedical literature.⁵ They reported that many abstracts state lower p values than the full text. Also many p values are reported in isolation, with no corresponding confidence intervals (CIs), Bayes factors, or effect sizes that would help in the interpretation of p values. It was also noted 96% of the articles had at least one positive (<0.05 or <0.01) p value reported, implying a publication bias for positive results only.^{5,6} In this era of “publish or perish”, this creates pressure on the researchers to report the “good” p values. Widespread practice of the so-called data dredging, torturing the data, “significance chasing”, selective inference or “ p -hacking” all refer to the excessive presentation of only the promising findings.² Hence, clinician readers should be aware and seek information on the actual hypothesis that was explored during the study, the data collection and analyses performed, and the computed p values before changing practices based on the so-called positive studies. Misguiding declarations of “trends towards significance” when the

p value is close to the threshold for significance further confuse the matters. In the field of oncology, 63 (8.7%) of the 722 publications high-impact oncology journals had statements pertaining to "trend" toward significance, while reporting statistically nonsignificant results.⁷ There is no transparency regarding the analytic paths or the number of analysis performed despite the requirement for preregistration of trials. Researchers still have considerable power and leeway in the method they choose to project their results.⁷

HOW CAN WE OVERCOME THESE ISSUES?

Several suggestions have been put forth for logical interpretation and use of *p* values.⁴ Some journals forbid the trend statements.⁸ Another suggestion to lower the threshold of *p* value, especially for newer findings, was also proposed. Using the absolute numerical *p* value instead of thresholds may be challenging as many articles have already been published with *p* value thresholds. There is an emerging use of Bayesian statistics in oncology journals, involving inductive reasoning instead of the deductive reasoning of classical statistical analysis.⁹ Novel methods like analyzing the "fragility" of a randomized trial may complement the use of *p* value. The fragility index of a randomized trial is the number of events needed to convert the statistical significance of an existing result.¹⁰ Tannock et al. analyzed the key phase 3 randomized oncology trials that led to Food Drug Administration approvals, they reported that nearly half of them had a low fragility index, meaning that these "practice changing results" would have become nonsignificant with change in just a few events.¹¹ This shows that there are other factors to assess the robustness of clinical data than the *p* value.

Reporting of effect sizes and tackling the biases are more difficult solutions which need further progress in the field of reporting biomedical data. Totally abandoning *p* value also may not be practical.⁴

Today we are at the helm of an era of huge computer-generated databases and medical assistance apps; there is massive amounts of available medical literature at our fingertips. Most doctors use a handheld device for quick referencing.¹² Patient outcomes have also been found to improve with such use.¹³ Patient care is evidence based and personalized. So it is of prime importance to sensitize the scientific community about the pitfalls of *p* value. This may require more rigorous statistical training and emphasis from the level of undergraduate medical education curricula. Another solution could be scientific panel-based discussions critically analyzing the important research papers at conferences or meetings among peers.

The *p* value remains a powerful tool when properly utilized and interpreted. Logical thinking and scientific reasoning will remain above any statistical tool.² Clinicians should hone their skills related to the critical appraisal of scientific literature and correctly judge the actual worth of scientific research papers.

Take-home Messages

- *p* value is used as a measure of statistical significance.
- *p* value may not be useful in isolation, it is a powerful tool in the right hands when used in conjunction with other statistical methods.
- Reader should be alert regarding the misuse, misinterpretation, or misguiding statements made regarding the *p* value.
- Critical analysis of the relevant articles is possible only when the clinicians themselves understand the core concept of the statistical methods used in the analysis.
- Suggestions to overcome the pitfalls of *p* value include changing the threshold, reporting exact *p* values, using other methods in conjunction, tackle biases, and most importantly train the clinicians regarding the appropriate use of *p* value.

REFERENCES

1. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos Mag* 1900;50(302):157–175. DOI: 10.1080/14786440009463897.
2. Wasserstein RL, Lazar NA. The ASA's statement on *p*-values: context, process, and purpose. *Am Stat* 2016;70(2):129–133. DOI: 10.1080/00031305.2016.1154108.
3. Nahm FS. What the *p* values really tell us. *Korean J Pain* 2017;30(4):241–242. DOI: 10.3344/kjp.2017.30.4.241.
4. Ioannidis JPA. The proposal to lower *P* value thresholds to .005. *JAMA* 2018;319(14):1429–1430. DOI: 10.1001/jama.2018.1536.
5. Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of reporting *P* values in the biomedical literature, 1990–2015. *JAMA* 2016;315(11):1141–1148. DOI: 10.1001/jama.2016.1952.
6. Fanelli D. Negative results are disappearing from most disciplines and countries. *Scientometrics* 2011;90(3):891–904. DOI: 10.1007/s11192-011-0494-7.
7. Nead KT, Wehner MR, Mitra N. The use of "Trend" statements to describe statistically nonsignificant results in the oncology literature. *JAMA Oncol* 2018;4(12):1778–1779. DOI: 10.1001/jamaoncol.2018.4524.
8. JAMA Oncology, JAMA Network. Instructions for authors. Available at <https://jamanetwork.com/journals/jamaoncology/pages/instructions-for-authors>. Last accessed on 16 March 2020.
9. Adamina M, Tomlinson G, Guller U. Bayesian statistics in oncology. *Cancer* 2009;115(23):5371–5381. DOI: 10.1002/cncr.24628.
10. Walsh M, Srinathan SK, McAuley DF, Mrkobrada M, Levine O, Ribic C, et al. The statistical significance of randomized controlled trial results is frequently fragile: a case for a fragility index. *J Clin Epidemiol* 2014;67(6):622–628. DOI: 10.1016/j.jclinepi.2013.10.019.
11. Del Paggio JC, Tannock IF. The fragility of phase 3 trials supporting FDA-approved anticancer medicine: a retrospective analysis. *Lancet Oncol* 2019;20(8):1065–1069. DOI: 10.1016/S1470-2045(19)30338-9.
12. Ventola CL. Mobile devices and apps for health care professionals: uses and benefits. *P T* 2014;39(5):356–364.
13. Isaac T, Zheng J, Jha A. Use of UpToDate and outcomes in US hospitals. *J Hosp Med* 2012;2(2):85–90. DOI: 10.1002/jhm.944.