LETTER

P Value Problems

To the Editor: Since its invention 90 years ago, the *p* value has become the standard by which most quantitative research is judged; however, it was never intended for this purpose.[1] Indeed, a 2016 joint statement by the American Statistical Association argued, "By itself, a *p* value does not provide a good measure of evidence regarding a model or hypothesis."[2] A lone *p* value is uninformative because it is prone to false positives and says nothing about the magnitude or range of an effect.[3] Additionally, over-reliance on *p* values may even encourage unethical research practices.[4]

Before discussing its shortcomings, it is helpful to define a *p* value. A *p* value is the likelihood of obtaining one's data if the null hypothesis is true. Contrary to popular misconception, it is not the probability that one's results were obtained by chance, the probability that the null hypothesis is true, or the probability of a false positive result.[5] In fact, the false positive rate associated with a *p* value of .05 is usually around 30%, but can be much higher.[6] This discussion may seem pedantic, but accurate false positive rates are a practical matter. A recent study that attempted to replicate 100 psychology experiments was only able to replicate 38.[7] Amgen was only able to replicate six of its 53 landmark cancer studies.[8] Bayer could only replicate 25% of the 67 studies that it attempted.[9] Is there any reason to believe that pharmacy education would fare better?

Even when a *p* value is interpreted correctly, it is silent on the magnitude and range of an effect. Even the most miniscule effect can be statistically significant if the sample size is large enough.[10] For example, a study of the effect of aspirin on myocardial infarction (MI) collected 22,000 subjects over a 5-year period and found that aspirin reduced the risk of MI, $p<.00001$. The risk difference for this study, however, was 0.77% and the R^2 was 0.001. This means that only one-tenth of 1% of the risk of suffering MI could be explained by aspirin.[11] Focusing exclusively on the *p* value when the sample size is large can overstate the practical importance of one's conclusions. In addition to a *p* value, researchers should report effect sizes and R^2 so that readers can properly interpret the magnitude of their findings. Additionally, the *p* value does not specify the range of probable outcomes; hence, researchers should also report confidence intervals.

Not all of the *p* values' shortcomings are mathematical; some are ethical. The publication of accurate and honest results is a moral concern, and overreliance on *p* values can obstruct this goal.[4] Making a *p* value of .05 the sole arbiter of whether or not a manuscript is published can encourage researchers to hunt for small *p* values using ethically questionable research practices such as: attempting a study multiple times but only reporting the study that produced a significant result; hedging their bets by collecting many variables but only reporting the ones that showed significant effects; dropping outliers or changing screening criteria after analysis has begun; splitting, merging, or transforming variables to produce a significant result; conducting significance tests before data collection is complete and terminating collection if a significant result is obtained. All of these practices are ethically dubious, and can harm the replicability of one's results.[4]

Given the *p* value's limitations, it should not be the sole arbiter of publication, and researchers should always report additional information, especially means, standard deviations, confidence intervals, R^2, and effect sizes. These additional statistics do not correct *p* values' shortcomings – some of which were not mentioned here, but they make results sections more informative and help to hold researchers accountable.

Samuel C. Karpen, PhD
Bill Gatton College of Pharmacy, East Tennessee State University, Johnson City, Tennessee

ACKNOWLEDGMENTS

The author would like to thank Adam Welch for his comments on an early draft of this letter.

REFERENCES

1. Fischer RA. *Statistical Methods for Research Workers*. Edinburgh, Scotland: Oliver and Boyd; 1925.
2. Wasserestin RL, Lazar NA. The ASA's statement on *p*-values: context, process, and purpose. *Am Stat*. 2016;70(2):129-133.
3. Kruschke JK. What to believe: Bayesian methods for data analysis. *Trends Cogn Sci.* 2010;14(7):293-300.
4. 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.
5. 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(4):337-350.
6. Sellke T, Bayarri MJ, Berger JO. Calibration of p values for testing precise null hypotheses. *Am Stat*. 2001;55(1):62-71.
7. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*. 2015;349(6251):943-953.
8. Begley CG. Ellis LM. Drug development: raise standards for preclinical cancer research. *Nature.* 2012;483(7391):531-533.
9. 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*. 2011;10(9):712.
10. Sullivan GM, Feinn R. Using effect size – or why the p value is not enough. *J Grad Med Educ*. 2012;4(3):279-282.
11. Bartolucci AA, Tendera M, Howard G. Meta-analysis of multiple primary prevention trials of cardiovascular events using aspirin. *Am J Cardiol*. 2011;107(12):1796-1801.