EDITORIAL

Significant, but not meaningful: an over-reliance on $p<0.05$

Sharmili Vidyadaran[1] and Yong Meng Goh[2]

[1] Department of Pathology, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia
[2] Department of Veterinary Preclinical Sciences, Faculty of Veterinary Medicine, Universiti Putra Malaysia, 43400 Serdang, Selangor Darul Ehsan, Malaysia

For too long, researchers have relied on *p*-values of <0.05 to deem their findings meaningful. A good number of us were told to hunt for that coveted double, or triple asterisks signifying highly significant, or very highly significant findings during hypothesis testing. The realisation that large sample sizes increase the probability for a statistical test to be significant, or lead to false positives (Type I error) has led us to rethink its usage (1).

Recently in *Nature*, Amrhein and colleagues wrote a rather inflammatory commentary on retiring statistical significance to reduce overstated claims or dismissal of important effects in research (2). The article was submitted with more than 800 signatories. In it, they cite the example of two studies that examined the effect of COX-2 inhibitors on atrial fibrillation. One reported that COX-2 inhibitors were associated with increased risk of atrial fibrillation (3), whilst the other reported no association, except for patients with chronic kidney or pulmonary disease (4). The results from the two studies do not differ; the relative risk (RR) for both studies was 1.2. This means that data from ***both*** studies demonstrate a 20% increased risk for COX-2 inhibitors users. It is simply that data from one study had a narrower range with statistical significance whilst the other did not (5). The authors of the second study derived their conclusion of no association ***solely*** based on the lack of statistical significance. *An over-dependence on p<0.05 silences data, and more importantly distorts the literature*.

Statistics was born from a need - studying an entire population is neither feasible nor practical and therefore, a proportion is sampled and used to make inferences for to the population. In March of this year, 43 papers in a special issue of The American Statistician discussed the issue of statistical significance and how to venture beyond it. The editorial in that issue, 'Moving to a World Beyond p<0.05', delivers a critical message – "*statistical significance was never meant to imply scientific importance*" (6). The editorial goes on to state "*As "statistical significance" is used less, statistical thinking will be used more*" (6).

The fault is not in statistical significance, the fault is in our categorical usage of it. Indeed, we need to be mindful that statistical significance is only meaningful if ***all*** the underlying population assumptions are adhered to (7). An increasing number of scientific publications now require the authors to emphasise on optimal sample size to limit Type I error, as well as to include test power. Test power is derived from (1-β), where β is the probability of false negative results (or Type II error) (1). It is a safeguard to prevent researchers from concluding that there was no statistical significance just because the statistical test used, sample size included and the overall conduct of the experiment was not sensitive enough to detect a significant change (8). Test power of a reported study would provide an objective view as when to trust research reports that are based heavily on (serendipitous) $p<0.05$ values but with dubious test power and repeatability.

The power of observation is imperative to science and should not be confounded by poor understanding of experimental conduct and statistical tests to corroborate them. While we race forward with automation, speed and multiplexing, interpretation and thought remain a human privilege. We must not transfer the privilege of data interpretation to the appearance of $p<0.05$ on our computer screens. More importantly, we must remember that there is no cookie-cutter approach to research. The $p<0.05$ approach has become a convenient way of inferring data but is not the only nor the best way. Our dependence on a *p*-value threshold must not blind us from what the data implies. When the effects of sample size and type I error are not reined in, or when we disregard the test power of our experimental and analytical approach to detect a change, over-dependence of $p<0.05$ would only lead to scientific inveracities. A call to retire the term 'statistically significant', however, seems unnecessary and importantly, is not the best solution. What we should call for is the ***thoughtful researcher*** (6) and to re-visit the very fundamentals of the statistical analysis used in our studies. The thoughtful researcher notes that the difference in height between two populations may be significant ($p<0.05$), but also notes that that difference is not necessarily meaningful (a mean difference of 0.02 metres). In science, there are no shortcuts for pondering thoughts.

REFERENCES

1. Sullivan LM, Weinberg J, Keaney JF, Jr. Common Statistical Pitfalls in Basic Science Research. J Am Heart Assoc 2016; 5(10):
2. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature 2019; 567(7748): 305-7
3. Schmidt M, Christiansen CF, Mehnert F, Rothman KJ, Sorensen HT. Non-steroidal anti-inflammatory drug use and risk of atrial fibrillation or flutter: population based case-control study. BMJ 2011; 343: d3450
4. Chao TF, Liu CJ, Chen SJ, Wang KL, Lin YJ, Chang SL, Lo LW, Hu YF, Tuan TC, Wu TJ, Chen TJ, Tsao HM, Chen SA. The association between the use of non-steroidal anti-inflammatory drugs and atrial fibrillation: a nationwide case-control study. Int J Cardiol 2013; 168(1): 312-6
5. Schmidt M, Rothman KJ. Mistaken inference caused by reliance on and misinterpretation of a significance test. Int J Cardiol 2014; 177(3): 1089-90
6. Wasserstein AL, Schirm AL, Lazar NA. Moving to a World Beyond "p<0.05". Am Stat 2019; 73(S1): 1-19
7. Leppink J, O'Sullivan P, Winston K. The bridge between design and analysis. Perspect Med Educ 2017; 6(4): 265-9
8. Dell RB, Holleran S, Ramakrishnan R. Sample size determination. ILAR J 2002; 43(4): 207-13