South African Orthopaedic Journal
http://journal.saoa.org.za

EDITORIAL

The *p*-value: *p* for problem

Leonard Marais

PhD, Editor-in-Chief, *South African Orthopaedic Journal*

The *p*-value was introduced by Fisher as a method to perform null hypothesis testing and has since been used widely in science as an indicator of significance.[1] It can be defined as a measure of strength of evidence against the null hypothesis.[2] In other words, it is the probability of finding an effect at least as or more extreme than the observed findings if the null hypothesis is true. However, the *p*-value is unable to reliably perform this function if the statistical power is not very high. In other words, if the power of a study is low, a repeat study will likely yield a substantially different *p*-value. Beta errors are common in orthopaedic literature, with up to 28% of randomised controlled trials erroneously failing to reject the null hypothesis.[3] Furthermore, the arbitrary cut-off of 0.05 has led to the scientifically unsound practice of regarding so-called 'significant findings' as more valuable, reliable or reproducible.[4] In fact, treating a *p*-value as a dichotomous variable is unfounded.[2] More worrying is the fact that the use of *p*-values may have served as an incentive for the introduction of bias: a practice referred to as '*p*-hacking'. These factors have combined to create serious concerns regarding the validity of many published scientific research findings, culminating in the statement that 'It can be proven that most claimed research findings are false'.[5]

P-hacking, also known as selective reporting or inflation bias, typically involves the misreporting of true effect sizes.[3] It occurs, for example, when researchers selectively employ certain statistical methodologies and/or data eligibility criteria in order to obtain a significant result. Aschwanden has eloquently illustrated, with the aid of an interactive infographic, how simple it can be to manipulate a *p*-value by simply changing a variable.[6] 'Data dredging' and the shotgun approach to data analysis involves bombarding data with statistical tests until something significant is found. The aim should rather be to adhere to a well-designed protocol with an astute research question. Probability testing should be thought of as currency, which should only be used to answer your research question/s or test the hypothesis. While it would remain reasonable to apply statistical methods to illustrate the similarity or dissimilarity of the groups being compared in a cohort study, reporting a *p*-value for every data element in a study depreciates its overall 'value'.

Aside from *p*-hacking, other problems, inherent to the nature of *p*-values, remain. Hypothesis testing involves the calculation of a test statistic (e.g., chi square value) that reflects the magnitude of association and the resulting *p*-value reflecting the extent to which the null hypothesis is compatible with the observed findings (assuming that the test statistic follows a specific probability distribution). It is not, for example, the probability that the null hypothesis is true or that the result is due to chance.[7] As pointed out by Gagnier and Morgenstern: 'Overreliance on significance tests to interpret statistical findings ignores the magnitude of the association, estimation of precision, the consistency and pattern of results, possible bias arising from several sources, previous research findings, and foundational knowledge of relevant biological and clinical phenomena.'[7] In 2016 the American Statistical Association developed a statement that aimed to guide the use and interpretation of *p*-values. Six important principles are highlighted:[7,8]

1. '*P*-values can indicate how compatible the data are with the specified statistical model.' The accuracy of the *p*-value is only as good as the underlying statistical model and the assumptions used to arrive at it.
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.' The *p*-value is therefore 'a statement about data in relation to a specified hypothetical explanation and is not a statement about the explanation itself'. Thus, the *p*-value only informs us of whether the statistical model used to test our assumptions is compatible with the observed data.
3. 'Scientific conclusions should not be based only on whether a *p*-value passes a specific threshold.' In other words, a *p*-value that exceeds some arbitrary threshold does not tell us anything about the clinical importance of the findings.
4. 'Proper inferences require full reporting and transparency.' A nice way of asking us to please avoid *p*-hacking.
5. '… does not measure the size of the effect or the importance of a result.' Therefore, a smaller *p*-value does not imply a stronger association.
6. 'By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.' There are numerous factors that can affect the *p*-value including the power of the study, the statistical model used and various sources of bias.

Gagnier and Morgenstern describe the emphasis on *p*-values in orthopaedic literature as misguided and ask us to move away from the emphasis on *p*-values with statements such as 'statistically significant'.[7] Instead, it is recommended that confidence intervals are reported for tested outcomes in order to convey some information about the magnitude, direction and precision of the association. In addition, interpretation of results requires cognisance of any relevant factor, including all possible confounders and sources of bias; possible measurement errors; the suitability of the statistical

model; and findings from previous studies. As researchers, we have to accept that statistics alone are insufficient to translate our study findings to clinical practice. Finally, hypothesis testing should be employed judiciously in order to maintain the value of our findings.

References

1. Fisher RA. *Statistical methods for research workers*. 1925, London: Oliver & Boyd.
2. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. *Nature* 2015;**12**(3):179-85.
3. Abdullah L, Davies DE, Fabricant PD, Baldwin K, Namdari S. Is there truly 'no significant difference'. *J Bone Joint Surg* 2015;**97**-A(24):2068-73.
4. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of *p*-hacking in science. *PLoS Biol* 2105;**13**(3): e1002106.doi:10.1371/journal.pbio.1002106.
5. Ioannidis JPA. Why most published research findings are false. *PLoS Med* 2015;**2**(8): e124. DOI: 10.1371/journal.pmed.0020124.
6. Aschwanden C. Science isn't broken. FiveThirtyEight 2015. Available from: https://fivethirtyeight.com/features/science-isnt-broken/#part1 (Last accessed: 15 February 2018).
7. Gagnier JJ, Morgenstern H. Misconceptions, misuses, and misinterpretations of p values and significance testing. *J Bone Joint Surg* 2017;**99**-A(18):1598-603.
8. Wasserstein RL, Lazar NA. The ASA's statement on *p*-values: context, process and purpose. *Am Stat* 2016;**70**(2):129-33.