

Misinterpretations of the 'p value': a brief primer for academic sports medicine

Steven D Stovitz,¹ Evert Verhagen,² Ian Shrier³

INTRODUCTION

When comparing treatment groups, the p value is a statistical measure that summarises the chance ('p' for probability) that one would obtain the observed result (or more extreme), if and only if, the treatment is ineffective (ie, under the assumption of the 'null' hypothesis). The p value does not tell us the probability that the null hypothesis is true.¹ This editorial discusses how some common misinterpretations of the p value may impact sports medicine research. Although presented from a treatment standpoint, the same principles hold for causes or prevention.

PROBABILITIES DO NOT TRANSLATE INTO YES OR NO DECISIONS

p Values are probabilities, yet often interpreted based on a categorical cut-off, generally at the level of 0.05 (ie, 5%). Anything below is considered a 'statistically significant difference' and vice versa. However, one would not change a decision to buy a lottery ticket if the chance of winning was 4.9% ($p=0.049$) instead of 5.1% ($p=0.051$). Consider a study where 100 participants who were given an injury prevention programme had six injuries, and 100 participants in the control group had 13 injuries ($p=0.091$). If the prevention group had one fewer injury (ie, 5/100 injuries), the results become statistically significant ($p=0.048$). Is it appropriate to conclude the treatment is ineffective when a single injury changes the results from non-significant to significant? Using a cut-off converts a

continuous variable (the probability) into a categorical variable (yes/no). Presenting actual p values rather than simply >0.05 or <0.05 may help readers understand that any cut-off is arbitrary.

IS THE P VALUE CLINICALLY MEANINGFUL?

With a sufficient number of participants, even a small difference between groups may result in $p<0.05$. Consider an exercise programme for weight loss that lowered the mean body weight of participants from 110 kg (SD=5 kg) to 109.5 kg. The expected p value with 100 participants is 0.43,ⁱⁱ with 1000 participants it is 0.10, and with 10 000 participants it is 5×10^{-7} . Although 'statistically significant' with 10 000 participants, a decrease of 0.5 kg is not clinically meaningful in terms of reduction in obesity-related health problems. Exactly what difference is clinically meaningful will depend on context for particular problems and not p values which incorporate sample size and group differences.

THE POWER OF CHANCE

Sports medicine researchers often investigate multiple outcomes such as pain, weakness and function. If $p<0.05$ for any one outcome, it is often considered a statistically significant difference. Readers may conclude that the treatment was the cause of the difference in the outcome, assuming that the difference was unlikely to be due to chance. This is mathematically incorrect. Consider a 20-sided die with numbers 1–20 on each side. The probability of obtaining a '1' with one roll is 5%. This is equivalent to our p-value of 0.05 for one outcome. The probability of obtaining a '1' at least once on five rolls (equivalent to five outcomes) is actually 22.6%. Similarly, in a study with a treatment that is ineffective for all five outcomes, one would expect to see a p-value <0.05 for at least one outcome ~22.6% of the time.

The problem with the misinterpretation of a low p value as statistically significant

when the study evaluates multiple outcomes was highlighted when a science journalist fooled major media outlets with a study suggesting that eating chocolate helps to lose weight.² The group studied 18 different outcomes, which means there was a 60% probability of having $p<0.05$ for at least one of the outcomes by chance. In this study, the 'significant' outcome happened to be weight loss. If the study were repeated, there would likely be a different outcome that occurred with $p<0.05$. Replication research is essential so that practitioners are not fooled into thinking that differences are due to treatment when they may, in fact, be due to chance.³

SUMMARY AND RECOMMENDATIONS

In summary, the p value is a probability under the assumption that a treatment is ineffective. Sports medicine practitioners must understand that translation of this probability into a categorical decision is questionable. Certainly, a value above the level of 0.05 does not mean that the groups are the same and a value below 0.05 does not mean that the treatment difference is meaningful, or even that the groups are different. The problems associated with misinterpreting p values, as outlined above, may be decreased if authors present the actual estimate differences with confidence intervals.^{4–6} Although beyond the scope of this article, confidence intervals provide readers with an estimated range of true differences that are likely compatible with the observed results and provide much more information than the p value.^{4–6} There is no consensus on the best way to account for multiple comparisons. It is necessary for authors to use some type of statistical correction (and/or alter the p value for rejecting the null hypothesis) or acknowledge the issue of multiple comparisons as a limitation of the study analysis.

Twitter Follow Evert Verhagen @Evertverhagen

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.



To cite Stovitz SD, Verhagen E, Shrier I. *Br J Sports Med* 2017;**51**:1176–1177.

Accepted 7 November 2016
Published Online First 24 December 2016

Br J Sports Med 2017;**51**:1176–1177.
doi:10.1136/bjsports-2016-097072

ⁱp Value calculated via two-tailed χ^2 statistical test.

¹Department of Family Medicine and Community Health, University of Minnesota, Minneapolis, Minnesota, USA; ²Department of Public and Occupational Health Amsterdam, VU University Medical Center, Amsterdam, The Netherlands; ³Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, McGill University, Montreal, Canada

Correspondence to Steven D Stovitz, Department of Family Medicine and Community Health, University of Minnesota, 420 Delaware Street SE, MMC 381, Minneapolis, MN 55455, USA; stovitz@umn.edu

ⁱⁱp Value calculated via two-tailed t-test.

REFERENCES

- 1 Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *Am Stat* 2016;1305:129–33.
- 2 Bohannon J. I fooled millions into thinking that chocolate helps weight loss. Here's how. <http://io9.gizmodo.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800> (accessed 23 Sep 2016).
- 3 Ioannidis JP. Why science is not necessarily self-correcting. *Perspect Psychol Sci* 2012;7:645–54.
- 4 Gardner MJ, Altman DG. Statistics in medicine confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J* 1986;292:746–50.
- 5 Cummings P, Koepsell TD. P values vs estimates of association with confidence intervals. *Arch Pediatr Adolesc Med* 2010;164:193–6.
- 6 Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. *Eur J Epidemiol* 2010;25:225–30.



Misinterpretations of the 'p value': a brief primer for academic sports medicine

Steven D Stovitz, Evert Verhagen and Ian Shrier

Br J Sports Med 2017 51: 1176-1177 originally published online November 24, 2016

doi: 10.1136/bjsports-2016-097072

Updated information and services can be found at:
<http://bjsm.bmj.com/content/51/16/1176>

These include:

References

This article cites 5 articles, 1 of which you can access for free at:
<http://bjsm.bmj.com/content/51/16/1176#BIBL>

Email alerting service

Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
<http://group.bmj.com/group/rights-licensing/permissions>

To order reprints go to:
<http://journals.bmj.com/cgi/reprintform>

To subscribe to BMJ go to:
<http://group.bmj.com/subscribe/>