https://doi.org/10.1016/j.msksp.2021.102351

Editorial

Please refrain from presenting p-values!

Arianne P Verhagen

During our last editorial board meeting we as (associate) editors discussed the use of p-values, significance testing, effect sizes and clinical worthwhile effects in the manuscripts submitted to the journal and how to deal with them. We had a fruitful discussion, and this editorial reflects our consensus, and highlights some guidance for researchers when submitting their manuscript to Musculoskeletal Science & Practice.

Statistical reporting

There have been continuing discussions for at least two decades about the use and misuse of the *p*-value. This prompted the American Statistical Association (ASA) in 2016 to recommend that authors avoid statements on statistical significance and interpretation of outcomes using an arbitrary threshold [ASA 2020, Wasserstein 2016, 2019]. Traditionally, the *p*-value has been used in randomised clinical trials (RCTs) in various ways. Most often it is used in conjunction with testing the null hypothesis to answer study questions related to the effectiveness of interventions by dichotomising results as significant or not significant [Greenland 2016]. Null hypothesis testing has limitations [ASA 2020, Verhagen 2004, Sullivan 2012, Cohen 2016, Herbert 2019]. First and foremost, null hypothesis does not measure the probability of the truth of the null hypothesis. The *p*-value only suggests whether the null hypothesis should be rejected, but nothing about the validity of the null hypothesis [ASA 2020]. Assuming that the null hypothesis is actually correct, meaning there is no difference in effect between two interventions, with a *p*-value set at 0.05 a significant result will be found in about 1 in 20 studies with the same null hypothesis and a comparable set of participants by chance. This one time could, in fact, be in the trial that is being carried out at that moment. Importantly, a **p**-value does not measure the size or magnitude of an effect, and its replicability is poor [ASA 2020, Verhagen 2004, Sullivan 2012, Cohen 2016, Herbert 2019].

Another use of *p*-values in RCTs is when comparing baseline variables between randomized groups. Statistical testing for baseline differences between randomized groups is not recommended as the rationale is that when the randomization procedure is performed well, all differences at baseline are due to chance, and hypothesis testing at baseline is considered inappropriate and illogical [Austin 2010, Harvey 2018].

P-values are also frequently used in other study designs, such as (prospective) cohort studies and cross-sectional studies. Sometimes *p*-values are used to express whether an association or correlations was 'statistically significant'. These significance figures give an indication as to whether the correlation coefficients deviate significantly from 0 (zero) [Schober 2018]. However, this is not something that we need to know. Instead, we want to know how close the correlation coefficients are to 1 or -1.

Reporting guidelines

To increase the reporting quality of research, various guidelines are developed per design [see Equator-network.org]. There are several recommendations in these guidelines regarding the reporting and appropriate use of *p*-values.

For the reporting of RCTs, authors are guided by the CONSORT-statement (Consolidated Standards of Reporting Trials) [Moher 2010]. The CONSORT, for example, recommends authors not report results solely as *p*-values, authors instead are encouraged to use effect estimates and 95% confidence intervals (95% CIs) (item 17a) [Moher 2010]. In contrast to *p*-values, effect estimates demonstrate the strength and the direction of the effect [Herbert 2019, Abbott 2014, McLeod 2019]. In addition, 95% CIs provide a range of values between which the estimated true effect estimate lies [Altman 2006]. For binary outcomes (yes/no answers, like patients are yes/no recovered) it is even more tempting to do statistical significance testing using the Chi-square test (or one of its variants). In this case the CONSORT-statement recommends calculating a relative effect (risk ratios (relative risks) or odds ratio's in rare cases) and an absolute effect (risk differences), all with their 95%CI [Moher 2010]. Unfortunately, this is not the default setting in the most used statistical packages SPSS.

Clinical relevance, or a clinical meaningful or clinical worthwhile effect, is another parameter used to interpret the magnitude of the effect [Kamper 2019a,b,c]. Terms like "minimal clinically important difference", and "smallest worthwhile effect" are used to determine whether the size of the difference between randomised groups is likely to be worthwhile from the patient's perspective [Ferreira 2012, Kamper 2019b].

All advice mentioned above is also recommended in the reporting of systematic reviews of RCTs, as guided by the PRISMA-statement (Preferred Reporting Items for Systematic reviews and Meta-

Analysis) [Moher 2009]. For cohort studies the recommended reporting guideline is the STROBE-guideline (Strengthening The Reporting of OBservational studies in Epidemiology) [von Elm 2014]. The STROBE also strongly encourages presentation of relative and absolute risk as outcome estimates with their 95% CI. Associations between variables can be presented as correlation coefficients (Pearson, Spearman, Rank, Intra-Class), and again these can also be presented with their 95%CI. The interpretation of correlation coefficients is often guided by some cut-off points stating when an association is strong, moderate or weak, although these cut-offs are arbitrary and differ between different authors [Schober 2018].

Meta-research

Apart from endorsing the use of different reporting guidelines, many academic journals separately endorsed the recommendation by the ASA to refrain from using *p*-values when more appropriate measures can be used (not all *p*-values are bad). Nevertheless, authors continue to use *p*-values to conclude whether an intervention is effective and should be used clinically, and reviewers and editors allow this. A large meta-research study evaluating the reporting of *p*-values in biomedical research over 25 years, showed that 96% of all publications presented at least one *p*-value, and this percentage was steady over time [Chavaliaris 2015].

Another meta-research study evaluated the overall quality of methods in physiotherapy RCTs (e.g. randomisation, blinding, analysis) in the PEDro database [Gonzalez 2018]. They found that the methods of these RCTs were of sub-optimal quality, as the PEDro score (ranging from 0-10) was on average 5.3. Another study found that 95% CIs were reported in approximately 29% of physiotherapy trials, with a steady increase in the use over time from 2% in 1986 to 42% in 2016 [Freire 2019].

A very recent, and not yet published, study investigated the use of *p*-values, effect estimates and clinical relevance in physiotherapy RCTs published in 2000 and 2018 in six major (Q1 peer-reviewed) journals [Verhagen, submitted]. In 2018 we found 101 RCTs, and 91.4% of RCTs reported *p*-values for the primary (between-group) analysis, 61.4% performed statistical significance testing for baseline differences and 42.6% of studies did not report the effect estimates. It remains unclear why so many authors choose not to present effect estimates in an RCT, and why reviewers and journal editors permit authors to do so when it is conceivable that a reader may misinterpret the result.

Relevance

Physiotherapy is a profession that strives to work towards an evidence-based model, with numerous initiatives such as the PEDro database to assist consumers of physiotherapy research [Moseley 2019]. Research is one of the pillars of evidence-based practice and plays a fundamental role in guiding treatment selection. When selecting treatments, physiotherapists must be aware that statistical significance does not equate to clinical relevance [Thiese 2016]. Presenting effect estimates and variability of the effect (using 95% CIs) will allow clinicians to consider how much a patient is likely to benefit from a given intervention compared to another intervention. In addition, clinical relevance of outcomes is important when interpreting whether the effects of an intervention are meaningful to patients [Ferreira 2012, Kamper 2019b]. Researchers have an ethical obligation to accurately report findings to allow for evidence-based decision-making [Verhagen 2004, du Prel 2009]. Authors, as well as reviewers and editors, should have been aware of reporting guidelines and been obligated to adhere to these guidelines [du Prel 2009].

Future Directions

As discussed above there are several alternatives for using *p*-values. We will summarise these below and provide some guidance on the reporting of studies in the future:

1. Do not use the phrase "statistically significant", as this way you dichotomise your result although the results are probably on a continuous spectrum [ASA 2020].
2. Refrain from incorrect use of statistical testing, such as baseline comparisons and correlation coefficients.
3. Focus on the size of the effect (effect estimate) for RCTs (and systematic reviews of RCTs) and on the strength of the association in cohort studies.
4. Consider what would be a clinically meaningful effect and formulate your research aim in a way that this is reflected. For instance, "we aim to evaluate whether exercises, compared with manual therapy, affects pain in patients with low back pain". In the method you define 'exercises', 'manual therapy', and 'patients with low back pain'. In addition, you explain how you assess pain and when you consider a difference between randomised groups as clinically meaningful.
5. Accept uncertainty by calculating and interpreting the confidence interval (precision) around the effect estimate. The interpretation can be as follows: assume you found a difference in recovery of 5% between the intervention and control group, and the hypothetical 95% CI ranges from -2% to 12%. The conclusion might be that the 95% CI contains 0% difference, so the true effect could be no effect. This is equivalent to a non-significant result and, like using *p*-values, can still result in a dichotomous interpretation of the findings. Better is to say that

there is probably a real difference in treatment effect that is nearer to 5% than to -2% or 12%, but still there is a chance that patients are deteriorating (not recovering) in the intervention group. Make sure that you do not present separate confidence intervals for the outcome in each group but only for the treatment effect, meaning the between group difference.

6. If you do all the above, and you are not happy with the reviewer comments, challenge editors and reviewers that correct practices and interpretations of research should be followed in papers submitted to their journals.

Arianne Verhagen, PhD, MSc, MT, BSc, PT
Associate Editor
University of Technology Sydney Australia
Arianne.Verhagen@uts.edu.au

References

1. ASA website: https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf. https://www.tandfonline.com/doi/full/10.1080/00031305.2019.1583913 Last visited 27 January 2021.
2. Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician. 2016;70(2):129-133
3. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p < 0.05". The American Statistician. 2019;73(sup1):1-19
4. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European journal of epidemiology. 2016;31(4):337-350
5. Verhagen AP, Ostelo RWJG, Rademaker A. Is the p value really so significant? Australian Journal of Physiotherapy 2004;50:261-2.
6. Sullivan GM, Feinn R. Using Effect Size-or Why the P Value Is Not Enough. J Grad Med Educ. 2012;4:279-282
7. Cohen J. The earth is round (p<. 05). In: What if there were no significance tests?: Routledge; 2016:69-82.
8. Herbert R. Research Note: significance testing and hypothesis testing: meaningless, misleading and mostly unnecessary. Journal of Physiotherapy 2019;65:178-181.
9. Austin PC, Manca A, Zwarenstein M, Juurlink DN, Stanbrook MB. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. J Clin Epidemiol. 2010;63(2):142-53. doi: 10.1016/j.jclinepi.2009.06.002.
10. Harvey LA. Statistical testing for baseline differences between randomised groups is not meaningful. Spinal Cord. 2018;56(10):919. doi: 10.1038/s41393-018-0203-y.

11. Schober, Patrick MD, PhD, MMedStat; Boer, Christa PhD, MSc; Schwarte, Lothar A. MD, PhD, MBA Correlation Coefficients: Appropriate Use and Interpretation, Anesthesia & Analgesia. 2018:126:1763-1768. doi: 10.1213/ANE.0000000000002864
12. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c869.
13. Abbott JH, Schmitt J. Minimum important differences for the patient-specific functional scale, 4 region-specific outcome measures, and the numeric pain rating scale. J Orthop Sports Phys Ther. 2014;44(8):560-564
14. McLeod SA. (2019, June 10). What are confidence intervals in statistics? Simply psychology: https://www.simplypsychology.org/confidence-interval.html. Last visited 21 September 2020
15. Altman DG. Confidence intervals in practice. In: Altman DG, Machin D, Bryant TN, Gardner MJ, eds. Statistics with confidence. 2nd ed. BMJ Books, 2000:6-14.
16. Kamper SJ. Interpreting Outcomes 1-Change and Difference: Linking Evidence to Practice. J Orthop Sports Phys Ther. 2019a;49(5):357-358. doi: 10.2519/jospt.2019.0703.
17. Kamper SJ. Interpreting Outcomes 2-Statistical Significance and Clinical Meaningfulness: Linking Evidence to Practice. J Orthop Sports Phys Ther. 2019b;49(7):559-560. doi: 10.2519/jospt.2019.0704.
18. Kamper SJ. Interpreting Outcomes 3-Clinical Meaningfulness: Linking Evidence to Practice. J Orthop Sports Phys Ther. 2019c;49(9):677-678. doi: 10.2519/jospt.2019.0705.
19. Ferreira ML, Herbert RD, Ferreira PH, Latimer J, Ostelo RW, Nascimento DP, Smeets RJ. A critical review of methods used to determine the smallest worthwhile effect of interventions for low back pain. J Clin Epidemiol. 2012;65(3):253-261. doi:10.1016/j.jclinepi.2011.06.018
20. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol. 2009;62(10):1006-12. doi: 10.1016/j.jclinepi.2009.06.005.
21. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. Int J Surg. 2014;12(12):1495-9. doi: 10.1016/j.ijsu.2014.07.013.
22. Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of Reporting P Values in the Biomedical Literature, 1990-2015. JAMA. 2016;315(11):1141-8. doi: 10.1001/jama.2016.1952.
23. Gonzalez GZ, Moseley AM, Maher CG, Nascimento DP, Costa LDCM, Costa LO. Methodologic Quality and Statistical Reporting of Physical Therapy Randomized Controlled Trials Relevant to Musculoskeletal Conditions. Arch Phys Med Rehabil. 2018;99(1):129-136. doi: 10.1016/j.apmr.2017.08.485.
24. Freire APCF, Elkins MR, Ramos EMC, Moseley AM. Use of 95% confidence intervals in the reporting of between-group differences in randomized controlled trials: analysis of a representative sample of 200 physical therapy trials. Braz J Phys Ther. 2019;23(4):302-310. doi:10.1016/j.bjpt.2018.10.004
25. Verhagen AP, Stubbs PW, Mehta P, Kennedy D, Nasser AM, Quel de Oliveira C, Pate JW, Skinner IW, McCambridge AB. Metaphor-meta-research in physiotherapy trials: trends in the reporting of statistical significance and clinical relevance between 2000 and 2018. Submitted.

26. Moseley AM, Elkins MR, Van der Wees PJ, Pinheiro MB. Using research to guide practice: The Physiotherapy Evidence Database (PEDro). Braz J Phys Ther. 2019;S1413-3555(19)30914-1. doi: 10.1016/j.bjpt.2019.11.002.
27. Thiese MS, Ronna B, Ott U. P value interpretations and considerations. Journal of thoracic disease. 2016;8(9):E928-e931
28. du Prel JB, Hommel G, Rohrig B, Blettner M. Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications. Deutsches Arzteblatt international. 2009;106(19):335-339