Injury, Int. J. Care Injured 50 (2019) 1934–1937

Contents lists available at ScienceDirect

Injury

journal homepage: www.elsevier.com/locate/injury

Effects of a proposal to alter the statistical significance threshold on previously published orthopaedic trauma randomized controlled trials

Austin L. Johnson[a,*], Sheridan Evans[a], Jake X. Checketts[a], Jared T. Scott[a,c], Cole Wayant[a], Mark Johnson[c], Brent Norris[b,c], Matt Vassar[a,c]

[a] Oklahoma State University Center for Health Sciences, Tulsa, OK, United States
[b] Orthopaedic & Trauma Services of Oklahoma, Tulsa, OK, United States
[c] Oklahoma State University Medical Center - Department of Orthopaedics, Tulsa, OK, United States

ARTICLE INFO

Article history:
Accepted 10 August 2019

Keywords:
Statistical significance
P-value
Orthopaedic
Orthopaedic trials
Trauma
Traumatology
Randomized controlled trials
Clinical trials
Analysis
Evidence
Statistics

ABSTRACT

Introduction: A recent proposal suggests changing the threshold for statistical significance from a P value of .05 to .005 to minimize bias and increase reproducibility of future studies. *P* values less than .05 but greater than .005 would be reclassified as “suggestive”, whereas P values less than .005 would be considered significant. The present study explores how lowering the P value threshold would affect the interpretation of previously published orthopaedic trauma randomized controlled trials (RCTs) and whether outcomes from these trials would maintain statistical significance under the proposed P value threshold.

Methods: All RCTs published between January 01, 2016 and January 31, 2018 in the *Journal of Orthopaedic Trauma*, *Injury*, and *Archives of Orthopaedic and Trauma Surgery* were screened by at least 2 authors. Data from included trials were extracted in blinded and duplicate fashion. All P values for primary endpoints were included from each study.

Results: We identified 124 primary endpoints from 48 trials: 39.5% (49/124) of endpoints had a *P* value less than .05 and 60.5% (75/124) had a *P* value greater than .05. Overall, 51.0% (25/49) of statistically significant primary endpoints were less than .005, while 49.0% (24/49) would be reclassified as suggestive.

Conclusion: Based on our results, adopting a lower threshold of significance would heavily alter the significance of orthopaedic trauma RCTs and should be further evaluated and cautiously considered when viewing the effect such a proposal on orthopaedic practice.

© 2019 Elsevier Ltd. All rights reserved.

Introduction

A common misconception in clinical research is that statistical significance equates to clinical significance [1–3]. R.A Fisher’s original idea on the use and interpretation of significance testing remains a long-standing debate because of the increasing reliance on significant *P* (probability) values [4–8]. Relying solely on *P* values for interpreting significant data can be misleading [9–12]; however, their use in research is not necessarily a problem [13]. The primary purpose of using a *P* value is to minimize type I errors — erroneous conclusions made about differences between groups when no such difference truly exists. The type I error rate is often specified *a priori* at 0.05, meaning that there is a 1 in 20 chance — or a 5% risk — that the difference detected is because of chance rather than attributed to the effects of the intervention. If the *P* value lies above this threshold ($P > .05$), the result is not statistically significant [14,15]. Using a *P* value to minimize type I errors can be very beneficial, but misinterpretation of study results can have severe consequences [16–18].

To reduce misinterpretation of study results, a large and influential group of statisticians and research methodologists have advocated for the research community to adopt a new significance level by lowering the threshold from .05 to .005. *P* values ranging from .05 to 0.005 would be reclassified as “suggestive” and interpreted with discretion [19–21]. This proposal is meant to serve as a temporizing measure while long-term strategies are developed, tested, and implemented. Studies claiming statistically significant findings with any inaccuracy or misinterpretation of data could lead to adverse effects on clinical practice and

[*] Corresponding author at: Oklahoma State University Center for Health Sciences, 1111 W 17th St., Tulsa, OK, 74107, United States.
E-mail address: austin.johnson14@okstate.edu (A.L. Johnson).

https://doi.org/10.1016/j.injury.2019.08.012
0020-1383/© 2019 Elsevier Ltd. All rights reserved.

ultimately future studies [16–18]. Therefore, careful consideration should be given to the protective effects of adopting a new threshold of significance.

Because the clinical research community has heavily adopted the use of the *P* value, lowering the threshold may prove beneficial [22]. By lowering the threshold of *P* values from .05 to .005, the proportion of true effects that emerge will be higher, thus doing more good than harm [20]. In particular, RCTs routinely report *P* value for treatment outcomes. RCTs are among the most influential study designs in orthopaedic research, and it is well-known that RCTs heavily influence clinical decision-making [16,23].

Owing to the influential nature of RCTs in orthopaedic surgery, the effects of shifting the *P* value threshold is of particular interest. Our study explores the effect of lowering the *P* value threshold in previously published orthopaedic traumatology RCTs. Here, we evaluate the proportion of primary endpoints reported in orthopaedic traumatology RCTs that would maintain statistical significance under the new proposed *P* value threshold (.005), the proportion that would be reclassified as suggestive, and the proportion that would be nonsignificant.

Methods

All studies published from January 01, 2016 to January 31, 2018 in the *Journal of Orthopaedic Trauma*, *Injury*, and *Archives of Orthopaedic and Trauma Surgery* were analyzed and screened by at least 2 authors (Fig. 1). The following Pubmed search strategy was deployed to locate RCTs: ((("randomized controlled trial"[Publication Type]) OR "clinical trial"[Publication Type])) AND (((("Injury"[Journal]) OR ("Archives of orthopaedic and trauma surgery"[Journal])) OR "Journal of orthopaedic trauma"[Journal])".

Inclusion criteria: The journals selected for inclusion were chosen on the basis of being the only 3 orthopaedic traumatology journals ranked in the top 20 of Google Scholar's "orthopaedic medicine and surgery" category. One of us (BN), a board certified orthopaedic traumatologist, selected these journals because they publish research relevant to orthopaedic traumatology. To ensure an adequate sample size, studies were chosen from January 01, 2016 to January 31, 2018. After this selection process, all studies with a stated primary endpoint and *P* value were included.

Exclusion Criteria: Any trial that did not state a primary endpoint or report *P* values to compare groups were excluded.

Following screening, we extracted *P* value data for each study's primary endpoints since RCTs are more often powered for these endpoints [24]. All *P* values for primary endpoints were included from each trial. Study characteristics were also extracted for each study. Extracted data was performed blinded and in duplicate fashion using a pilot-tested Google form. Following data extraction, all discrepancies were resolved by consensus between the investigators. We used Google Forms for collection of data and

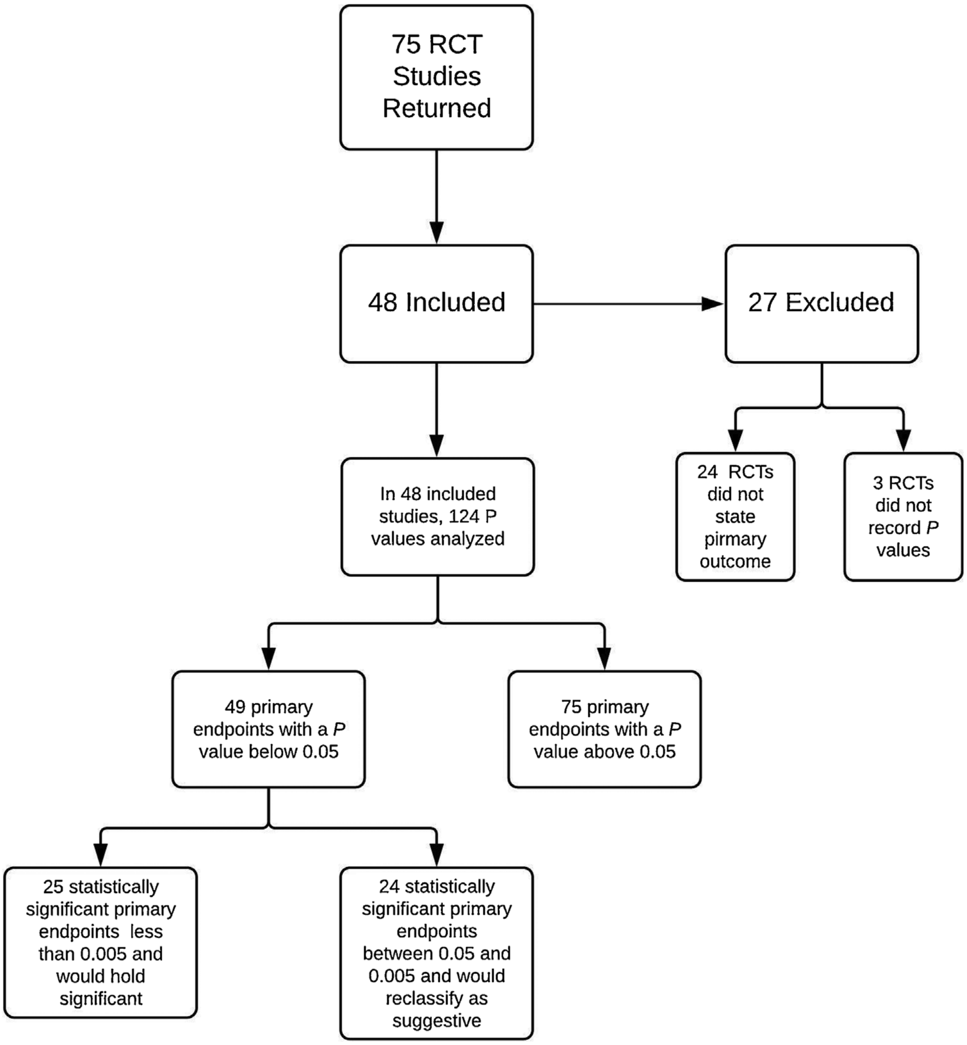
75 RCT Studies Returned
48 Included
27 Excluded
In 48 included studies, 124 P values analyzed
24 RCTs did not state pirmary outcome
3 RCTs did not record *P* values
49 primary endpoints with a *P* value below 0.05
75 primary endpoints with a *P* value above 0.05
25 statistically significant primary endpoints less than 0.005 and would hold significant
24 statistically significant primary endpoints between 0.05 and 0.005 and would reclassify as suggestive

Fig. 1. Flow diagram of study inclusion.

analysis of data using STATA 13.1. This methodology was previously conducted and reported in the senior author's (MV) work looking at the effects of this threshold change in high impact factor general medical journals [22].

Results

Of 75 articles retrieved, 48 were included. The 27 excluded were mostly those that did not state a primary outcome (n = 24) or those that did not report *P* values (n = 3).

Characteristics of the included trials

Surgery was the primary intervention type for 60.4% (29/48) of the included trials. A large number of trials (25/48) did not mention funding source. The majority (44/49) were randomized trials. Few (27.1%) were multi-centered (13/48) and none were multinational. Most trials were conducted in Europe (20/48) and North America (16/48). A total of 66.7% (32/48) of trials included a power analysis (Table 1).

Primary results

We identified 124 primary endpoints from 48 trials: 39.5% (49/124) of endpoints had a *P* value less than .05, and 60.5% (75/124) had a *P* value greater than .05. Of the 124 primary endpoints, only 20.2% (25/124) of the endpoints were less than .005, and would maintain statistical significance under the proposed threshold. Only 3 of the 48 included trials had all primary endpoints that met the new threshold of .005. Of the 3 studies where all primary endpoints met the new threshold, no power analyses were mentioned. Of statistically significant primary endpoints (.05), overall, 51.0% (25/49) reported a P value of less than .005, while 49.0% (24/49) would be reclassified as suggestive.

Table 1
Characteristics of included clinical trials (n = 48) or endpoints (n = 124).

Characteristic	No. (%)
Journal (n = 48)	
Archives of Orthopaedic and Trauma Surgery	15 (31.3%)
Injury	12 (25.0%
Journal of Orthopaedic Trauma	21 (43.8%)
Intervention (n = 48)	
Drug	5 (10.4%)
Procedure	5 (10.4%)
Anesthesia/Analgesia (Nerve Blocks/Pain Management)	5 (10.4%)
Surgery	29 (60.4%)
Other	4 (8.3%)
[a]*Funding Source (n = 48)*	
Industry	3 (6.3%)
Public	3 (6.3%)
Private	11 (22.9%)
Hospital	3 (6.3%)
Other	1 (2.1%)
Not Mentioned	25 (52.1%)
None	5 (10.4%)
Number of trial centers (n = 48)	
Multicenter	13 (27.1%)
Single center	35 (72.9%)
Location (n = 48)	
Multinational	0 (0%)
Single country	48 (100%)
Type of endpoint (n = 48)	
Subjective	13 (27.1%)
Objective	35 (72.9%)
Sample size (median, [IQR])	76 [50-134]

[a] In some of the trials, multiple funding sources were identified.

Of the 124 total *P* values included, 31.0% (13/42) in *Archives of Orthopaedic and Trauma Surgery,* 24.0% (6/25) in *Injury*, and 10.5% (6/57) in *Journal of Orthopaedic Trauma* were less than .005.

Discussion

The primary objective of our study was to assess the effect that lowering the statistical threshold of significance would have on previous published RCT within orthopaedic trauma journals. Our results show that 86.5% of primary outcome *P* values in orthopaedic trauma RCTs would not maintain statistical significance at the *P* <.005 threshold. When analyzing the statistically significant primary endpoints (.05), 41.5% would remain statistically significant with a *P* value threshold of less than .005. Although these results should be interpreted cautiously, should a new threshold be adopted, orthopaedic traumatology research not within the new threshold may be interpreted as suggestive rather than absolute in regard to significance.

In a recent study published in the *Journal of the American Medical Association,* Wayant et al. reported that 70.7% (123/174) of statistically significant primary endpoints were less than .005 in 3 major general medical journals with high impact factors (*New England Journal of Medicine*, *Journal of the American Medical Association*, and *The Lancet*). In comparison, we found that 41.5% of statistically significant orthopaedic traumatology RCTs were below a *P* value of .005. Wayant et al. further concluded that 29.3% (51/174) of trial endpoints would be reclassified as suggestive [22]. Our study shows comparatively lower findings. Fewer orthopaedic traumatology RCT endpoints would maintain statistical significance under the proposed threshold when compared to RCTs in general medical journals. The reason for these differences are unknown but may relate to issues of sample size or methodological rigor.

A .005 threshold for significance may address a few shortcomings of *P* values, such as underpowered RCTs, spurious false positive results, and a phenomenon known as *P* hacking [25,26]. *P* hacking occurs when researchers analyze data multiple ways until a significant effect is found. With a lower *P* value threshold of significance, the ability to *P* hack data may be significantly decreased [27]. Another fault in measuring significance with a *P* value is publication bias. It has been found that 74% of published articles in surgical literature reported positive results; however, this number is inflated because of publication bias that will overestimate the clinical relevance of treatment [28]. The methodological quality in surgical RCTs also tends to lag behind the general literature [16,23]. For example, Karanicolas et al. found that only 33% of studies published in surgical journals were of high quality, contrasted with 75% published in general medicine journals [29]. Efforts have been taken toward improving the design and reporting of surgical RCTs, and lowering the threshold of significance may place a greater emphasis on attaining high quality, evidence-based data when conducting RCTs [30–33].

Our study design was subject to both strengths and limitations. The strengths of this study include the use of double screening and double data extraction, which is the golden standard methodology recommended by the Cochrane Collaboration when perform systematic reviews [34]. While this study is not a systematic review of clinical interventions, the methodological process described herein is amenable to the screening and data extraction procedures employed in systematic reviews. Limitations of this study include selecting 3 orthopaedic trauma journals over a 2-year period; thus, the results may not be generalizable to RCTs in other orthopaedic journals and should be viewed descriptively rather than inferentially. The study upon which our methodology is based also included 3 journals [22]. Also, while we conducted our search in PubMed, which catalogues all of the Medline database and indexes all journals included in the present study, it is possible that our search did not return all RCTs published in these journals.

In conclusion, lowering the *P* value to .005 may address some shortcomings of RCTs in orthopaedics. A lower *P* value threshold may be a promising temporizing measure to improve RCT methodology and general reproducibility of findings [35] until more permanent solutions are found, which could take considerable time to implement. Adopting a lower threshold of significance would heavily alter the interpretation of orthopaedic trauma RCTs. Caution is thus warranted regarding the effects of such interpretations on clinical decision making.

Source of funding

This work was not funded.

Declaration of Competing Interest

The authors report no conflicts of interest.

References

[1] Narayanan UG, Wright JG. Evidence-based medicine: a prescription to change the culture of pediatric orthopaedics. J Pediatr Orthop 2002;22(3):277–8.

[2] Guyatt G, Montori V, Devereaux PJ, Schünemann H, Bhandari M. Patients at the centre: in our practice, and in our use of language. Bmj Evid Med 2004;9(1):6–7.

[3] Wright JG, Swiontkowski MF, Heckman JD. Introducing levels of evidence to the journal. J Bone Joint Surg Am 2003;85-A(1):1–3.

[4] Bhandari M, Montori VM, Schemitsch EH. The undue influence of significant p-values on the perceived importance of study results. Acta Orthop 2005;76 (3):291–5.

[5] Goodman SN. Toward evidence-based medical statistics. 1: the P value fallacy. Ann Intern Med 1999;130(12):995–1004.

[6] Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol 2008;45(3):135–40.

[7] Gelman A. Commentary: P values and statistical practice. Epidemiology 2013;24(1):69.

[8] Cohen J. The earth is round ($p < .05$). Am Psychol 1994;49(12):997–1003.

[9] Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. Evolution of reporting P values in the biomedical literature, 1990–2015. JAMA 2016;315(11):1141–8.

[10] Luus HG, Müller FO, Meyer BH. Statistical significance versus clinical relevance. Part II. The use and interpretation of confidence intervals. S Afr Med J 1989;76(11):626–9.

[11] Nurminen M. Statistical significance—a misconstrued notion in medical research. Scand J Work Environ Health 1997;232–5.

[12] Sterne JAC, Cox DR, Smith GD. Sifting the evidence—what's wrong with significance tests? Another comment on the role of statistical methods. BMJ 2001;322(7280):226–31.

[13] Thiese MS, Ronna B, Ott U. P value interpretations and considerations. J Thorac Dis 2016;8(9):E928–31.

[14] Cohen J. Statistical power analysis for the behavioral sciences, 1988. Hillsdale, NJ: L. Lawrence Earlbaum Associates; 1988. p. 2.

[15] Vandenbergh B. Creative strategy in advertising. In: Jewler A Jerome, editor. Paperback, 230 pp. *J Advert*. 1981;10(2):46-47. Belmont, California: Wadsworth Publishing Company; 1981.

[16] Chess LE, Gagnier J. Risk of bias of randomized controlled trials published in orthopaedic journals. BMC Med Res Methodol 2013;13:76.

[17] Khan M, Nathan E, Mark G, Anthony H, Olufemi RA, Asheesh B, et al. The fragility of statistically significant findings from randomized trials in sports surgery: a systematic survey. Am J Sports Med 2017;45(9):2164–70.

[18] Altman DG. Poor-quality medical research: what can journals do? JAMA 2002;287(21):2765–7.

[19] Navon D, Cohen Y. Consider avoiding the .05 significance level. arXiv [statOT] 2016(June). http://arxiv.org/abs/1606.09017.

[20] Ioannidis JPA. The proposal to lower P value thresholds to .005. JAMA 2018;319 (14):1429–30.

[21] Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. Redefine statistical significance. Nat Hum Behav 2017;2(1):6–10.

[22] Wayant C, Scott J, Vassar M. Evaluation of lowering the P value threshold for statistical significance from .05 to .005 in previously published randomized clinical trials in major medical journals. JAMA 2018;320(17):1813–5.

[23] Gummesson C, Atroshi I, Ekdahl C. The quality of reporting and outcome measures in randomized clinical trials related to upper-extremity disorders. J Hand Surg Am 2004;29(4)727–34 discussion 735-737.

[24] Iwashyna TJ, Deane AM. Individualizing endpoints in randomized clinical trials to better inform individual patient care: the TARGET proposal. Crit Care 2016;20(1):218.

[25] Halpern SD, Karlawish JHT, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA 2002;288(3):358–62.

[26] Ioannidis JPA. Why most published research findings are false. PLoS Med 2005;2(8):e124.

[27] 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.

[28] Hasenboehler EA, Choudhry IK, Newman JT, Smith WR, Ziran BH, Stahel PF. Bias towards publishing positive results in orthopedic and general surgery: a patient safety issue? Patient Saf Surg 2007;1(1):4.

[29] Karanicolas PJ, Bhandari M, Taromi B, Akl EA, Bassler D, Alonso-Coello P, et al. Blinding of outcomes in trials of orthopaedic trauma: an opportunity to enhance the validity of clinical trials. J Bone Joint Surg Am 2008;90(5):1026–33.

[30] Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276(8):637–9.

[31] Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996;17(1):1–12.

[32] Ioannidis JPA, Evans SJW, Gøtzsche PC, O'Neill RT, Altman DG, Schulz K, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med 2004;141(10):781–8.

[33] Bhandari M, Richards RR, Sprague S, Schemitsch EH. The quality of reporting of randomized trials in the Journal of Bone and Joint Surgery from 1988 through 2000. J Bone Joint Surg Am 2002;84-A(3):388–96.

[34] Guides and handbooks. 2019. . (Accessed 9 February 2019) https://training.cochrane.org/handbooks.

[35] de Ruiter J. Redefine or justify? Comments on the alpha debate. Psychon Bull Rev 2018(September), doi:http://dx.doi.org/10.3758/s13423-018-1523-9.