abrapg-ft
Associação Brasileira de Pesquisa e Pós-Graduação em Fisioterapia

Brazilian Journal of Physical Therapy

https://www.journals.elsevier.com/brazilian-journal-of-physical-therapy

SYSTEMATIC REVIEW

Reporting of confidence intervals, achievement of intended sample size, and adjustment for multiple primary outcomes in randomised trials of physical therapy interventions: an analysis of 100 representatively sampled trials

David Fernández Hernando[a], Mark Elkins[b], Ana Paula Coelho Figueira Freire[c,*]

[a] *Hospital Fundación Jiménez Díaz, Madrid, Spain*
[b] *University of Sydney, Faculty of Medicine and Health, Sydney, Australia*
[c] *Central Washington University, Health Sciences, Ellensburg, WA, United States*

Received 30 April 2023; received in revised form 15 November 2023; accepted 8 May 2024
Available online 21 May 2024

KEYWORDS
Adjustment for multiplicity;
Clinical trials;
Confidence intervals;
Physical therapy;
Sample size

Abstract

Background: The physical therapy profession has made efforts to increase the use of confidence intervals due to the valuable information they provide for clinical decision-making. Confidence intervals indicate the precision of the results and describe the strength and direction of a treatment effect measure.

Objectives: To determine the prevalence of reporting of confidence intervals, achievement of intended sample size, and adjustment for multiple primary outcomes in randomised trials of physical therapy interventions.

Methods: We randomly selected 100 trials published in 2021 and indexed on the Physiotherapy Evidence Database. Two independent reviewers extracted the number of participants, any sample size calculation, and any adjustments for multiple primary outcomes. We extracted whether at least one between-group comparison was reported with a 95 % confidence interval and whether any confidence intervals were interpreted.

Results: The prevalence of use of confidence intervals was 47 % (95 % CI 38, 57). Only 6 % of trials (95 % CI: 3, 12) both reported and interpreted a confidence interval. Among the 100 trials, 59 (95 % CI: 49, 68) calculated and achieved the required sample size. Among the 100 trials, 19 % (95 % CI: 13, 28) had a problem with unadjusted multiplicity on the primary outcomes.

Conclusions: Around half of trials of physical therapy interventions published in 2021 reported confidence intervals around between-group differences. This represents an increase of 5 % from five years earlier. Very few trials interpreted the confidence intervals. Most trials reported a

[*] Corresponding author at: Central Washington University, Department of Health Sciences, 400 E University Way, Zip code: 98926, Ellensburg, WA 98926, United States.
E-mail: anapcff@hotmail.com (A.P. Freire).

https://doi.org/10.1016/j.bjpt.2024.101079

sample size calculation, and among these most achieved that sample size. There is still a need to increase the use of adjustment for multiple comparisons.

Introduction

Traditionally, when the effect of an intervention estimated from a clinical trial was published, it was reported with a *p*-value. Many statistical authorities have recommended reporting the effect of an intervention with a 95 % confidence interval instead of with a *p*-value. Such authorities include the Cochrane Collaboration,[1] the British Medical Journal,[2] the American Statistical Association,[3] the journal Statistics in Medicine,[4] and proponents of evidence-based practice in medicine and physical therapy.[5–7]

Given how widely *p*-values are used in the reporting of clinical trials,[8,9] it may surprise some readers that statistical authorities recommend an alternative approach. However, null hypothesis statistical tests, which generate *p*-values, have numerous problems.[8,10] First, the statistical significance of null hypothesis statistical tests has poor replicability. If a study generated a *p*-value between 0.005 and 0.05, then replicating the study with a new random sample from the same clinical population would have a 33 % chance of a non-significant result.[10] Second, *p*-values do not provide evidence about the precision of the estimate derived from the study. In contrast, confidence intervals distinguish between important differences that *p*-values do not. For example, when the *p*-value is non-significant, it may be that (a) the intervention is ineffective and the study has generated a precise estimate of that, or (b) the study was underpowered and the effect remains uncertain. Confidence intervals help to distinguish between those two scenarios. Furthermore, even when a *p*-value is significant, confidence intervals can distinguish how precisely the study has estimated the magnitude of the intervention's effect and can contribute to interpretation of whether the intervention is clinically worthwhile. *P*-values don't provide this information.[7,11]

The physical therapy profession has made efforts to increase the use of confidence intervals, including in publications in the 2000s,[12] 2010s,[11,13,14] and recently.[7] An analysis of a representative sample of physical therapy trials showed a modest but steady increase in the use of confidence intervals over a 30-year period.[9]

Although the trends showing an increase in the use of confidence intervals are heartening, there are several reasons why it is important to continue to monitor the use of confidence intervals in physical therapy trials. First, the percentage of physical therapy trials reporting confidence intervals at the end of the 30-year analysis period (2016) had only reached 42 %,[9] so further improvement is needed. Second, in some professions and journals, the proportion of studies reporting confidence intervals has plateaued.[15,16] Third, in 2022, there was an initiative by the International Society of Physiotherapy Journal Editors to encourage routine use of confidence intervals instead of *p*-values,[7] so establishing a baseline prevalence immediately prior to that initiative will help to later discern its effect.

To get the most benefit from a confidence interval, researchers must do more than simply report it. Researchers must also interpret the relevance of the range of values within the confidence interval and consider the implications arising from them. Many researchers calculate confidence intervals at the request of editors, but then ignore them and interpret their trial's result dichotomously as statistically significant or non-significant instead.[17] Interpretation is crucial but no study of physical therapy trials has yet examined whether confidence intervals are interpreted when they are reported.

It is common that clinical trials present underpowered samples due to several factors, including small sample size, unexpected low inflow of patients, or high attrition rates. Confidence intervals can be useful to describe the effect of sample size and indirectly reflect the statistical power of the clinical trial, which is influenced by the sample size. Although some issues have been raised regarding common approaches to sample size calculation in clinical trials,[10] if researchers are going to do a sample size calculation then it also seems worth observing how often the intended sample size is achieved in trials of physical therapy interventions.

Another issue related to confidence intervals and sample size calculation is whether one or multiple primary outcomes are nominated. Although adjustment for multiple primary outcomes is widely recommended and often used in clinical trials,[18] it is more commonly used with p-values than confidence intervals.[19] Therefore, it also seems worthwhile to observe whether physical therapy trials that nominate multiple primary outcomes adjust for multiple primary outcomes. Therefore, the primary aim of this study was to answer the following question:

1. How prevalent is the use of confidence intervals in the current reporting of between-group differences in randomised trials of physical therapy interventions?

Secondarily we aimed to answer:

2. Where confidence intervals are used in the reporting of between-group differences, are the confidence intervals interpreted by the authors?
3. Do physical therapy trials achieve their intended sample size?
4. Where physical therapy trials nominate multiple primary outcomes, do they adjust for multiple primary outcomes?

Methods

Study design

The Physiotherapy Evidence Database or 'PEDro' is a free, web-based database of evidence relevant to physical therapy. It is available at http://www.pedro.org.au/. It is a highly comprehensive source of randomized trials in physical therapy.[20,21]

We randomly selected 100 trials published in 2021 from those indexed on PEDro to form a representative sample for analysis. Random sampling was performed using Random function of Microsoft Excel software (Microsoft Office 2007, Microsoft Corporation, Redmond, Washington).

For the purpose of this study, randomised trials that failed to report a statistical comparison of the between-group difference were irrelevant. Therefore, the very small proportion of trials published in 2021 that failed to report a between-group statistical comparison were excluded. Also, because some of the data coded on PEDro (eg, subdiscipline) was used to characterise the trials, the small proportion of papers that were awaiting consensus coding were also excluded before the random selection of trials. We focused on the primary report of each research trial so pilot studies and secondary analyses were also excluded. Trials using Bayesian methods would report credibility intervals rather than confidence intervals, and therefore were excluded. There was no restriction by language of publication or area of physical therapy practice.

Data extraction

To characterize the trials, we downloaded the subdiscipline of physical therapy from PEDro. If a trial was coded under more than one subdiscipline, we used all codes so studies may be coded under more than one subdiscipline. We also downloaded the language of publication and the PEDro Scale quality criteria for the 100 trials. We then extracted answers to the following questions from the trials:

- How many participants were randomised? (number)
- Was an a-priori sample size calculation reported? (Yes/No)
  - If yes, what was the sample size indicated by the calculation? (number)
- How many sites were involved in recruitment? (number)
- What was the location of data collection? (continent)
- Was the trial funded? (Yes/No) We accepted only funding for the trial, not authors.
- Was at least one primary outcome identified? (Yes/No)
  - If yes, how many primary outcomes were identified? (number)
  - If >1, was there adjustment for multiple primary outcomes? (Yes/No)

Finally, we extracted whether the analysis of all, some or none of the between-group comparisons were reported using 95 % confidence intervals. Other levels of confidence interval were also considered (e.g., 99 % confidence interval). Confidence intervals for other types of analysis (e.g., baseline characteristics, within-group comparisons) were not considered. We also extracted whether the types of outcomes reported with confidence intervals were continuous, dichotomousor both; and recorded whether the confidence intervals were presented numerically, graphically or both.

We recorded whether at least one confidence interval was interpreted. Interpretation included any mention of the clinical implications of the confidence interval limit(s), any reference to the null value being inside or outside the confidence interval, and/or any reference to location of the confidence interval relative to the smallest worthwhile effect (or a synonym such as the minimum clinically important difference) or a clinically relevant threshold. The reporting of interpretation was recorded as Yes/No.

We also recorded whether the trial reported a primary outcome (i.e., the terms *primary, principal*, *main* or *key* were used when specifying an outcome). For all trials that had more than one primary outcome, we recorded whether there was any adjustment for multiple comparisons (including Bonferroni, sharpened Bonferroni, Dunn, etc.).

Two independent reviewers extracted all these data, with any disagreements resolved by discussion. The primary outcome in our analysis was the prevalence of reporting of the between-group differences with 95 % confidence intervals.

Data analysis

The sample size of 100 trials was chosen because it gives overall estimates of prevalence that have confidence limits smaller than ±10 %, which we consider to be sufficiently precise estimates to characterize the use of 95 % confidence intervals. For the random sample of 100 trials, data related to trial characteristics was reported using descriptive statistics: mean and standard deviation (SD) for normally distributed, continuous data; median and interquartile range (IQR) for non-normally distributed, continuous data; and number (%) for dichotomous data. The PEDro scale quality criteria were tallied to a total score for the descriptive statistics.

From the random sample of 100 trials published in 2021, we calculated the percentage of trials that used confidence intervals when reporting the between-group comparison for at least one outcome. Because this result is calculated to estimate the prevalence of reporting confidence intervals among all trials of physical therapy interventions published in 2021, the prevalence estimate was accompanied by a confidence interval.

We compared trials that reported or did not report confidence intervals with respect to their PEDro Scale scores and number of participants randomised. The data for PEDro Scale scores and for the number of participants were not normally distributed, so we performed Hodges-Lehmann estimation of the median difference (95 % CI).[22] MedCalc software was used for analysis.

We compared the percentage of trials that used confidence intervals in 2021 to the percentage in 2016 from previously published data.[9]

Results

The June 2022 update of PEDro contained 2832 trials published in 2021. About 2 % of trials were excluded because they were in-process or did not report a between-group comparison. None of the randomly sampled trials used Bayesian methods so none was excluded for this reason.

Characteristics of the included trials

The main characteristics of the 100 randomly selected trials are presented in Table 1. In this cohort, all trials were published in English. Most studies were classified in

Table 1 Summary of characteristics extracted from the published reports of the 50 trials randomly selected from 2016 (previously published data[9]) and the 100 trials randomly selected from 2021.

	2016 (*n* = 50)	2021 (*n* = 100)
English language n (%)	48 (96)	100 (100)
Subdiscipline, n (%)		
Cardiothoracic	4 (8)	13 (13)
Continence and women's health	7 (14)	11 (11)
Ergonomics and occupational health	2 (4)	0 (0)
Gerontology	1 (2)	12 (12)
Musculoskeletal	13 (26)	29 (29)
Neurology	3 (6)	13 (13)
Oncology	1 (2)	5 (5)
Orthopaedics	6 (12)	8 (8)
Paediatrics	2 (4)	11 (11)
Sports	6 (12)	8 (8)
Other	5 (10)	9 (9)
Total PEDro score (0 to 10), median [IQR]	6 [5; 7]	6 [5; 7]
Randomised participants, median [IQR]	81 [39; 123]	71 [45; 117]
Sample size calculation presented, n yes (%)	32 (64)	71 (71)
Sample size calculated, median [IQR]	89 [48; 173]	60 [39; 137]
Multicentre recruitment, n (%)		
No	25 (50)	49 (49)
Yes	19 (38)	23 (23)
Not specified	6 (12)	28 (28)
Sites involved if multicentre, median [IQR]	3 [2;8]	4 [2;12]
Continent		
Asia	15 (30)	44 (44)
Africa	0 (0)	3 (3)
Europe	17 (34)	30 (30)
North America	9 (18)	9 (9)
Oceania	7 (14)	7 (7)
South America	2 (4)	7 (7)
Funding, n (%)		
Yes	29 (58)	45 (45)
No	17 (34)	39 (39)
Unclear	4 (8)	16 (16)
Primary outcome identified, n (%)	33 (66)	62 (62)
Number of primary outcomes, median [IQR]	1 [1;2]	1 [1;2]
Adjustment for multiple primary outcomes, n (%)	**(*n* = 15)**	**(*n* = 63)**
Yes	3 (20)	23 (37)
No	12 (80)	40 (63)

IQR, interquartile range. If a trial was coded under >1 subdiscipline, we used all codes so studies may be under more than one subdiscipline.

musculoskeletal (29 %), cardiothoracic (13 %) and gerontology (12 %) subdisciplines.

How prevalent is the use of confidence intervals?

The prevalence of confidence intervals for at least one outcome among the 100 trials published in 2021 was 47 % (95 % CI: 38, 57). This represents a marginal improvement over 2016, when 42 % (95 % CI: 29, 56) of trials reported some or all outcomes with confidence intervals (Fig. 1A). Among the 47 trials that presented confidence intervals, 36 trials (77 %) reported them for continuous outcomes (Fig. 1B). Among the 47 trials that presented confidence intervals, 45 trials (96 %) reported them numerically, none reported them graphically only, and only 2 (4 %) reported confidence intervals both numerically and graphically.

Trials that reported confidence intervals had a marginally higher PEDro Scale score than trials that did not report confidence intervals (Hodges-Lehmann median difference: 1; 95 % CI: 0, 1), as shown in Fig. 2A. Trials that reported confidence intervals had higher sample sizes than trials that did not report confidence intervals (Hodges-Lehmann median difference: 34; 95 % CI: 12, 66). (Fig. 2B).

Where confidence intervals are used in the reporting of between-group differences, are the confidence intervals interpreted by the authors?

Among the 47 trials that presented at least one outcome with confidence interval around the between-group difference, only 6 (13 %) interpreted the confidence interval. This means that among the 100 trials, only 6 % (95 % CI: 3, 12) both reported and interpreted a confidence interval for at least one outcome.

Do physical therapy trials achieve their intended sample size?

Among the 100 trials, 32 (32 %, 95 % CI: 24, 42) did not report sample size calculation. Among the remaining 68 trials, the mean number randomised exactly met the calculated required sample size in 16 trials, exceeded it in 43 trials, and fell short of it in 9 trials. In relative terms across the 68 trials, the number randomised was 1.18 (SD 0.33) times larger than the calculated required sample size. Among the 100 trials, 59 (59 %, 95 % CI: 49, 68) both calculated and achieved their required sample size.

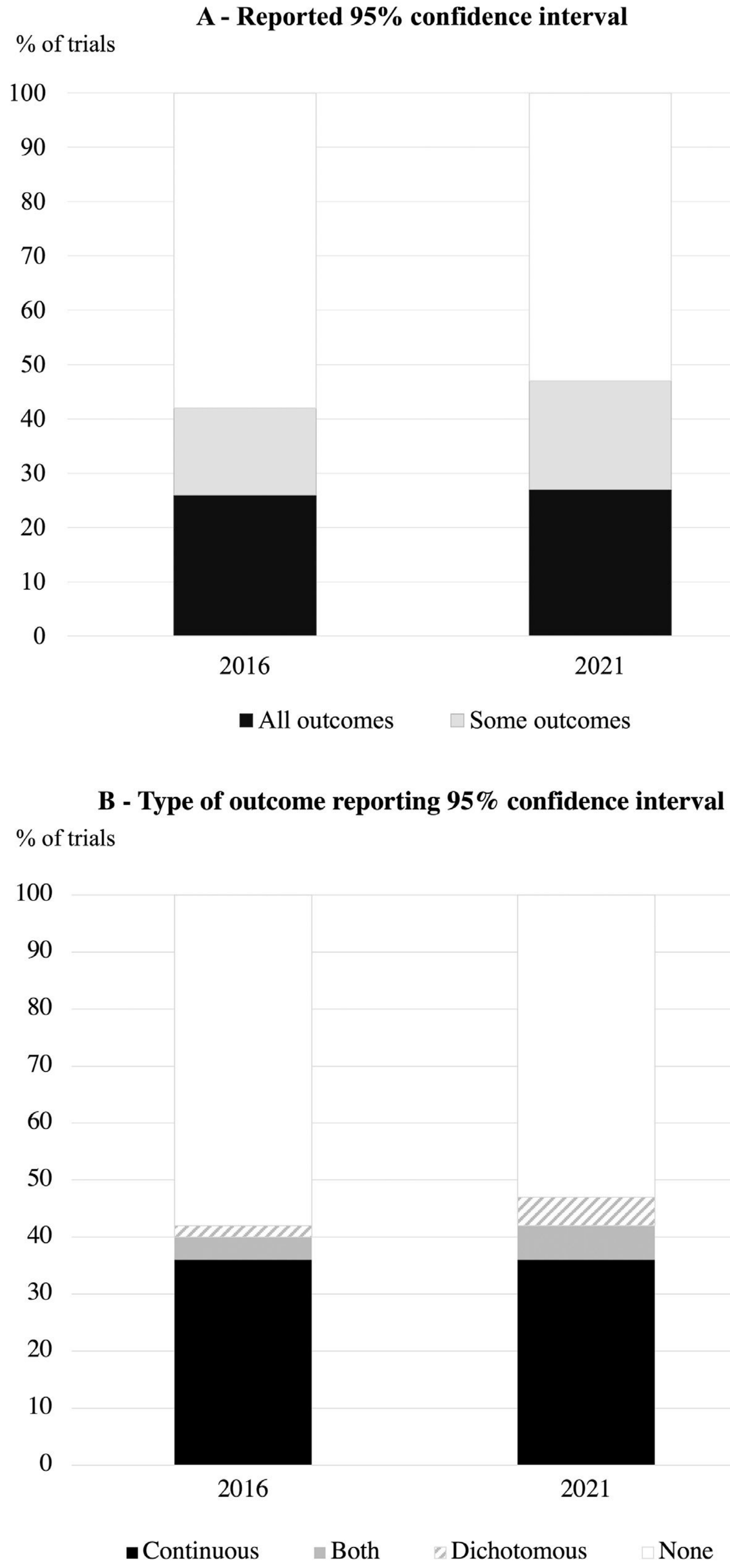

Fig. 1 Percentage of trials from 2016 to 2021 that (A) reported 95 % confidence intervals for at least one outcome and (B) reported 95 % confidence intervals for continuous and/or dichotomous outcomes.

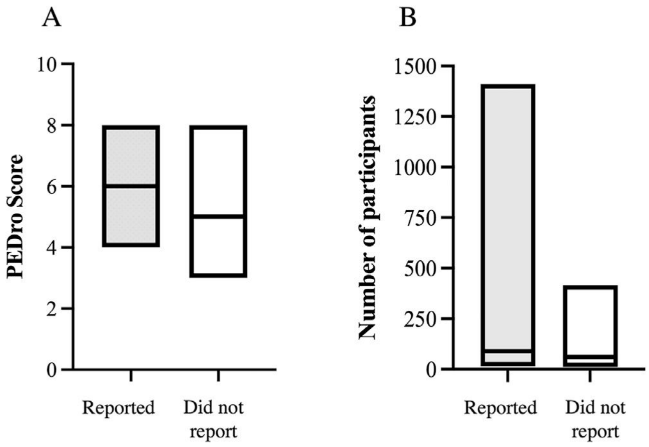

Fig. 2 Comparisons between the trials that did and did not report 95 % confidence intervals for two characteristics: (A) total PEDro score and (B) number of participants. Data are presented as median and IQR.

Where physical therapy trials nominate multiple primary outcomes, do they adjust for multiple primary outcomes?

Among the 100 trials, 37 did not nominate any primary outcome, 33 nominated a single primary outcome, 16 nominated two primary outcomes, and 14 nominated between three and eight primary outcomes. Among the 30 trials that nominated multiple primary outcomes, 11 (37 %) made an adjustment for multiplicity. This means that among the full cohort of 100 trials, only 19 % (95 % CI: 13, 28) had a problem with unadjusted multiplicity on the primary outcomes. Among the 11 trials that adjusted for multiplicity, only six also reported confidence intervals. Two of these studies adjusted p-values and four adjusted both p-values and confidence intervals.

Discussion

Overall, less than half of our sample of 100 trials reported confidence intervals. This represented an absolute increase of 5 % since 2016.[9] The improvement observed over this 5-year period might be an indication of advances in awareness regarding the important role of confidence intervals in providing meaningful information about the precision of estimates of clinical interventions. These trends are a positive sign in reducing the entrenched habit of using *p*-values as the method of statistical inference.

Previous studies observed the prevalence of confidence intervals reporting ranging from 86 % of trials in epidemiology, 54 % in public health, 27 % in dermatology, and 9 % in biomedical research.[17,23,24] The current reporting in physical therapy trials (47 %) was in the middle of this range. However, we did not restrict the inclusion of studies according to the publishing journals' impact factor, as performed in previous studies. Instead, we included any journals indexed on PEDro, which comprehensively indexes physical therapy trials,[20,21] thereby forming a representative sample for analysis. This might explain at least part of the differences in results across disciplines, as studies published in high impact journals are more likely to present confidence intervals.[9,15]

High impact journals tend to adhere to reporting guidelines, such as the Consolidated Standards of Reporting Trials (CONSORT), which require the presentation of estimated effect size and its precision.[25] Several journals already endorse the use of CONSORT as a requirement for submission and there is a trend to increase the adoption of this policy across journals.[26] This could be one of the factors related to the increase in reporting of confidence intervals observed between 2016 and 2021.

Confidence intervals are important to indicate the direction and size of the treatment effect, so improvements observed in the reporting of confidence intervals have important implications. Confidence intervals might help reduce "spin" (incorrect or misrepresentation of results) common in physical therapy research and help clinicians make more informed clinical decisions.[27] To continuously improve the reporting of confidence intervals, editors, reviewers and researchers should be encouraged to require that clinical trialists report confidence intervals as a standard component of their results.[7]

Reporting of confidence intervals is critical, but it's only half the job. Confidence intervals must also be interpreted appropriately. To our knowledge, this is the first study that investigated the format of representation (numerical or graphical) and whether these estimates were interpreted by researchers. A sound interpretation by the original authors can help clinicians to understand confidence intervals correctly. Among the 100 trials, only 6 % (95 % CI: 3, 12) of studies both reported and interpreted a confidence interval for at least one outcome. Our results are consistent with previous literature where the interpretation of results in terms of clinical relevance is sparsely used across trials.[17,28]

Some clinicians may not have a strong background in statistics or may have a misunderstanding of confidence intervals. These factors can lead to challenges in interpreting the clinical implications of confidence intervals.[29] Therefore, clinical trialists should discuss the clinical implications of the confidence intervals when they publish their results. This will help clinicians to appreciate the clinical relevance of the study findings. Providing an interpretation should involve discussion of the clinical implications of the range of values within the confidence interval. This might include reference to the null value being inside or outside the confidence interval, and reference to location of the confidence interval relative to the smallest worthwhile effect or a clinically relevant threshold.

Another aspect that could facilitate interpretation of confidence intervals by clinicians could be the format of presentation. Among the 47 trials that included confidence intervals, 96 % presented a numerical representation. Previous studies demonstrated that format of presentation of patient-reported outcomes in clinical trials may be associated with how accurately they are interpreted by clinicians.[30−32] Visual presentation strategies may be further explored, to improve efficacy in data communication around confidence intervals and proposed as an alternate method of presenting findings. Graphic presentations can provide an intuitive understanding regarding an outcome's estimates and help reduce the challenge for non-statisticians to understand the clinical implications of confidence intervals.

The sample size of a clinical trial directly affects confidence intervals. Lower sample sizes reduce the likelihood that the study generates estimates that are representative of the true effect in the wider patient population from which the study cohort was sampled. To acknowledge this larger degree of uncertainty in the findings, smaller samples generate wider confidence intervals. When a study's estimate is accompanied by a wide confidence interval, clinicians should acknowledge that uncertainty about what the true average effect of the intervention is. They need to be prepared to explain to patients that the evidence hasn't yet precisely estimated the average effect of the intervention on that outcome. They should be prepared to explain that the average effect might be favorable at the upper limit of the confidence interval, unfavorable at the lower limit of the confidence interval, or anywhere in between.

Our results showed some interesting findings regarding sample size. First, about one-third of the trials did not report a sample size calculation. Although some issues have been raised regarding common approaches to sample size calculation in clinical trials,[10] it is disappointing that this proportion of clinical trialists did not consider sample size in any formal way when planning their study. Second, among 68 trials that did report a sample size calculation, 59 achieved or exceeded the required sample size. This suggests that calculating a target sample size might inspire researchers to ensure it is achieved. Overall, though, there is still plenty of potential to improve the calculation and achievement of the required sample size, given that only 59 % (95 % CI 49, 68) achieved both of these elements. This problem has been observed elsewhere. For example, Gianola et al[33] identified that only 40 % of randomised controlled trials that analysed rehabilitation interventions for low back pain were adequately powered.

The last aim of our study was to determine the prevalence of adjustment for multiplicity among physical therapy trials that reported multiple primary outcomes. Among the 30 trials that nominated multiple primary outcomes, 11 (37 %) made an adjustment for multiplicity. Therefore, among the full cohort of 100 trials, only 19 % (95 % CI: 13, 28) had a problem with unadjusted multiplicity. Although it affected a relatively small proportion of the trials, adjustments for multiplicity are required to reduce the possibility of type I error.[18] Failure to adjust for multiple comparisons in randomised trials can result in unreliable findings, which can have negative consequences for patient care. Although adjustments for multiple comparisons are important, they also reduce the statistical power of the analysis for a given sample size and therefore widen the confidence intervals.[18] Adjustment of the sample size to account for this is recommended.

The strengths in our design included the assessment of multiple trial characteristics with reliable data extraction (two independent reviewers). Additionally, we used random selection to generate a large representative sample of trials, improving the generalisability of the findings. Our study presents important implications because the findings indicate the need for greater attention to the reporting of confidence intervals in randomised controlled trials. One limitation is that, despite applying no language restrictions during random sampling, only English-language trials were selected. This is likely because only 3 % of the eligible trials from which the 100 trials were selected were published in languages other than English. Another limitation is that we did not assess whether the studies that adjusted for multiple comparisons in their analysis also adjusted their sample size accordingly. This could be the focus of future research. Additionally, further research could address whether confidence intervals are calculated correctly and interpreted appropriately.

Conclusions

Around half of trials of physical therapy interventions published in 2021 reported a confidence interval around at least one between-group difference. This represents an increase of 5 % from 5 years earlier. Very few trials interpreted the confidence interval and presented the confidence interval graphically. Most trials of physical therapy interventions report a sample size calculation, and achieve the required sample size. There is still a need to increase the nomination of a primary outcome and, if multiple primary outcomes are nominated, the use of adjustment for multiple comparisons.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Higgins JPT, Thomas J, Chandler J, et al. Cochrane handbook for systematic reviews of interventions version 6.3 (updated February 2022). *Cochrane*2022.
2. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)*. 1986;292(6522):746–750. https://doi.org/10.1136/bmj.292.6522.746.
3. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "$p < 0.05$". *Am Stat*. 2019;73(sup1):1–19. https://doi.org/10.1080/00031305.2019.1583913.
4. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med*. 1998;17(8):857–872. https://doi.org/10.1002/(sici)1097-0258(19980430)17:8<857::aid−sim777>3.0.co;2-e.
5. Sharon Straus PG, Scott Richardson W, Brian Haynes R. *Evidence-Based Medicine: How to Practice and Teach EBM*. 5th ed. 2018.
6. Herbert R, Jamtvedt G, Hagen KB, Elkins MR. *Practical Evidence-Based Physiotherapy*. 2022.
7. Elkins MR, Pinto RZ, Verhagen A, Grygorowicz M, Söderlund A, Guemann M, et al. Statistical inference through estimation: recommendations from the International Society of Physiotherapy Journal Editors. *J Physiother*. 2022;68(1):1–4. https://doi.org/10.1016/j.jphys.2021.12.001.
8. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol Methods*. 2000;5(2):241–301. https://doi.org/10.1037/1082-989x.5.2.241.
9. Freire A, 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. https://doi.org/10.1016/j.bjpt.2018.10.004.

10. Boos DD, Stefanski LA. P-Value Precision and Reproducibility. *Am Stat*. 2011;65(4):213–221. https://doi.org/10.1198/tas.2011.10129.
11. Herbert R. Research Note: Significance testing and hypothesis testing: meaningless, misleading and mostly unnecessary. *J Physiother*. 2019;65(3):178–181. https://doi.org/10.1016/j.jphys.2019.05.001.
12. Herbert RD. How to estimate treatment effects from reports of clinical trials. I: continuous outcomes. *Aust J Physiother*. 2000;46 (3):229–235. https://doi.org/10.1016/s0004-9514(14)60334-2.
13. Kamper SJ. Showing confidence (intervals). *Braz J Phys Ther*. 2019;23(4):277–278. https://doi.org/10.1016/j.bjpt.2019.01.003.
14. Kamper SJ. Confidence intervals: linking evidence to practice. *J Orthop Sports Phys Ther*. 2019;49(10):763–764. https://doi.org/10.2519/jospt.2019.0706.
15. Stang A, Deckert M, Poole C, Rothman KJ. Statistical inference in abstracts of major medical and epidemiology journals 1975-2014: a systematic review. *Eur J Epidemiol*. 2017;32(1):21–29. https://doi.org/10.1007/s10654-016-0211-1.
16. Messam LLM, Weng HY, Rosenberger NWY, Tan ZH, Payet SDM, Santbakshsing M. The reporting of p values, confidence intervals and statistical significance in Preventive Veterinary Medicine (1997-2017). *PeerJ*. 2021;9:e12453. https://doi.org/10.7717/peerj.12453.
17. Fidler F, Thomason N, Cumming G, Finch S, Leeman J. Editors can lead researchers to confidence intervals, but can't make them think: statistical reform lessons from medicine. *Psychol Sci*. 2004;15(2):119–126. https://doi.org/10.1111/j.0963-7214.2004.01502008.x.
18. Vickerstaff V, Omar RZ, Ambler G. Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes. *BMC Med Res Methodol*. 2019;19(1):129. https://doi.org/10.1186/s12874-019-0754-4.
19. Benjamini Y, Yekutieli D, Edwards D, Shaffer JP, Tamhane AC, Westfall PH, et al. False discovery rate: adjusted multiple confidence intervals for selected parameters [with Comments, Rejoinder]. *J Am Stat Assoc*. 2005;100(469):71–93.
20. Moseley AM, Sherrington C, Elkins MR, Herbert RD, Maher CG. Indexing of randomised controlled trials of physiotherapy interventions: a comparison of AMED, CENTRAL, CINAHL, EMBASE, hooked on evidence, PEDro, PsycINFO and PubMed. *Physiotherapy*. 2009;95(3):151–156. https://doi.org/10.1016/j.physio.2009.01.006.
21. Michaleff ZA, Costa LO, Moseley AM, Maher CG, Elkins MR, Herbert RD, et al. CENTRAL, PEDro, PubMed, and EMBASE are the most comprehensive databases indexing randomized controlled trials of physical therapy interventions. *Phys Ther*. 2011;91 (2):190–197. https://doi.org/10.2522/ptj.20100116.
22. Hodges JL, Lehmann EL. Estimates of location based on rank tests. *Ann Math Statistics*. 1963;34(2):598–611.
23. Hopkins ZH, Moreno C, Secrest AM. Lack of confidence interval reporting in dermatology: a call to action. *Br J Dermatol*. 2019;180(4):910–915. https://doi.org/10.1111/bjd.17126.
24. Amaral EOS, Line SRP. Current use of effect size or confidence interval analyses in clinical and biomedical research. *Scientometrics*. 2021;126(11):9133–9145. https://doi.org/10.1007/s11192-021-04150-3.
25. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *Bmj*. 2010;340:c332. https://doi.org/10.1136/bmj.c332.
26. Shamseer L, Hopewell S, Altman DG, Moher D, Schulz KF. Update on the endorsement of CONSORT by high impact factor journals: a survey of journal “Instructions to Authors" in 2014. *Trials*. 2016;17(1):301. https://doi.org/10.1186/s13063-016-1408-z.
27. Khanpara H, Prakash V. Effect of spin in the abstract of a randomised controlled trial on physiotherapists' perception of treatment benefit: a randomised controlled trial. *BMJ Evid Based Med*. 2022;27(2):97–103. https://doi.org/10.1136/bmjebm-2021-111714.
28. Hoffmann TC, Thomas ST, Shin PN, Glasziou PP. Cross-sectional analysis of the reporting of continuous outcome measures and clinical significance of results in randomized trials of non-pharmacological interventions. *Trials*. 2014;15:362. https://doi.org/10.1186/1745-6215-15-362.
29. Paci M, Faedda G, Ugolini A, Pellicciari L. Barriers to evidence-based practice implementation in physiotherapy: a systematic review and meta-analysis. *Int J Qual Health Care*. 2021;33(2). https://doi.org/10.1093/intqhc/mzab093.
30. Brundage MD, Smith KC, Little EA, Bantug ET, Snyder CF. Communicating patient-reported outcome scores using graphic formats: results from a mixed-methods evaluation. *Qual Life Res*. 2015;24(10):2457–2472. https://doi.org/10.1007/s11136-015-0974-y.
31. Wanderer JP, Nelson SE, Ehrenfeld JM, Monahan S, Park S. Clinical data visualization: the current state and future needs. *J Med Syst*. 2016;40(12):275. https://doi.org/10.1007/s10916-016-0643-x.
32. Chen JC, Cooper RJ, McMullen ME, Schriger DL. Graph quality in top medical journals. *Ann Emergency Med*. 2017;69(4). https://doi.org/10.1016/j.annemergmed.2016.08.463. 453-61.e5.
33. Gianola S, Castellini G, Corbetta D, Moja L. Rehabilitation interventions in randomized controlled trials for low back pain: proof of statistical significance often is not relevant. *Health Qual Life Outcomes*. 2019;17(1):127. https://doi.org/10.1186/s12955-019-1196-8.