BMJ Open

Metaphor-meta-research in physiotherapy trials: reporting of statistical significance and clinical relevance in 2000 and 2018

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-054875
Article Type:	Original research
Date Submitted by the Author:	29-Jun-2021
Complete List of Authors:	Verhagen, Arianne; University of Technology Sydney Stubbs, Peter ; University of Technology Sydney, Graduate School of Health, Discipline of Physiotherapy Mehta, Poonam; University of Technology Sydney Kennedy, David; University of Technology Sydney Nasser, Anthony M; University of Technology Sydney Quel de Oliveira, Camila; University of Technology Sydney Pate, Joshua W; University of Technology Sydney Skinner, Ian W; University of Technology Sydney; Charles Sturt University McCambridge, Alana B; University of Technology Sydney
Keywords:	EPIDEMIOLOGY, EDUCATION & TRAINING (see Medical Education & Training), PRIMARY CARE

Metaphor-meta-research in physiotherapy trials: reporting of statistical significance and clinical relevance in 2000 and 2018

Authors:

Arianne P Verhagen[1], Peter W Stubbs[1], Poonam Mehta[1], David Kennedy[1], Anthony M Nasser[1], Camila Quel de Oliveira[1], Joshua W Pate[1], Ian W Skinner[1,2], Alana B McCambridge[1]

Affiliation:

[1] Graduate School of Health, Discipline of Physiotherapy, University of Technology Sydney, Sydney, Australia

[2] School of Community Health, Charles Sturt University, Port Macquarie, Australia

Funding:

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Protocol: https://doi.org/10.26195/ebqp-8a65

Corresponding Author:

Prof Arianne P Verhagen, Head of Discipline of Physiotherapy, Graduate School of Health, University of Technology Sydney Australia

E-mail: Arianne.Verhagen@uts.edu.au

Word count:

Abstract: 251; text: 3764

Number of references: 38; tables: 2; figures: 2

ABSTRACT

Design: meta-research

Objective: To compare the prevalence of reporting *p*-values, effect estimates and clinical relevance in physiotherapy randomized controlled trials (RCTs) published in the years 2000 and 2018.

Methods: We performed a meta-research study of physiotherapy RCTs obtained from six major physiotherapy peer-reviewed journals that were published in the years 2000 and 2018. We extracted data on the study characteristics and whether articles reported on statistical significance, effect estimates and confidence intervals for baseline, between-group, and within-group differences, and clinical relevance. Data were presented using descriptive statistics and inferences were made based on proportions. A 20% difference between 2000 and 2018 was regarded as a meaningful difference.

Results. We found 140 RCTs: 39 were published in 2000 and 101 in 2018. Overall, there was a high prevalence (>90%) of reporting *p*-values for the main (between-group) analysis, with no difference between years. Statistical significance testing was frequently used for evaluating baseline differences, increasing from 28% in 2000 to 61.4% in 2018. The prevalence of reporting effect estimates, confidence intervals and the mention of clinical relevance increased from 2000 to 2018 by 26.6%, 34% and 32.8% respectively. Despite an increase in use in 2018, over 40% of RCTs failed to report effect estimates, confidence intervals, and clinical relevance of results.

Conclusion. The prevalence of using *p*-values remains high in physiotherapy research. Although the proportion of reporting effect estimates, confidence intervals, and clinical relevance is higher in 2018 compared to 2000, many publications still fail to report and interpret study findings in this way.

Key words: Randomized clinical trials, Physiotherapy, reporting statistics, reporting clinical relevance

Strengths and Limitations

- This meta-research study will provide clear insight in the prevalence of (incorrect) use of p-values, and the prevalence of the use of effect estimates and clinical relevancy of outcomes
- We selected publications from six long-standing influential physiotherapy journals, assuming we select the best studies
- We defined a 20% difference as a meaningful difference
- We investigated reporting of *p*-values and effect estimates regardless of whether it was a primary or secondary outcome.

Introduction

As physiotherapy research informs clinical practice, it is important for clinicians to be confident in the quality of physiotherapy research. Meta-research is a relatively new scientific discipline that explores how research is performed, reported, reproduced, evaluated, and incentivised [1,2]. As all scientific research is prone to bias, it is important that each profession critically evaluates its own research methods, standards of reporting, and validity of the outcomes.

Continuing discussions about the use (and misuse) of the *p*-value prompted the American Statistical Association (ASA) to recommend in 2016 that authors avoid statements on statistical significance and interpretation of outcomes using a p-value as an arbitrary threshold [3,4,5]. Traditionally, the *p*-value has been used in randomised clinical trials (RCTs) in conjunction with the null hypothesis testing to answer study questions related to the effectiveness of interventions by dichotomising results as significant or not significant [6]. Although valuable if interpreted correctly, null hypothesis testing has its limitations; it does not measure the probability of the truth of the null hypothesis, it does not measure the size or magnitude of an effect, and its replicability is poor [3,7-10]. The recommendation of the ASA is endorsed by many academic journals, nevertheless, authors continue to conclude whether an intervention is effective and should be used clinically by a dichotomous interpretation based on *p*-values.

Well conducted and large RCTs are considered high quality evidence and reporting of RCTs should be guided by the CONSORT-statement (Consolidated Standards of Reporting Trials) [11]. There are several recommendations in the CONSORT-statement regarding the reporting and appropriate use of *p*-values. For example, authors should not report results solely as *p*-values and are encouraged to (also) use effect estimates and 95% confidence intervals (95% CIs) [11]. The advantage of effect estimates is their ability to demonstrate the strength and the direction of the effect, and the 95% CIs provide a range of values between which the estimated true effect estimate lies [10,12,13]. Nevertheless, a dichotomized interpretation of the confidence interval (CI) should be discouraged; it allows for discussing the accuracy, precision and/or relevance of the effect estimate. Clinical relevance is another parameter used to interpret the magnitude of the effect, and to deem if a finding is clinically meaningful.

According to the CONSORT-statement, authors should also compare baseline participant characteristics [11]. However, it discourages statistical significance testing of baseline covariates between randomized groups, as by using a proper randomization procedure all differences are based on chance. In addition, conclusions of a RCT should primarily be based on a between-group analysis by comparing post-intervention/follow-up outcomes between the groups or the between-group changes from baseline. Studies can additionally, with consideration, compare outcomes before and after the intervention using a ‘within-group’ analysis.

Previous meta-research within physiotherapy has investigated the use of randomization, blinding or intention-to-treat analysis [14-16] and one study evaluated the reporting of 95% CIs only [17]. To our knowledge, no study has examined the use of *p*-values, effect estimates or measures of clinical relevance in the physiotherapy literature before and after the CONSORT-statement was published in 2010. When selecting treatments, physiotherapists must be aware that statistical significance does not equate to clinical relevance [18]. Presenting effect estimates and variability of the effect (using 95% CIs) will also allow clinicians to consider how much a patient is likely to benefit from a given intervention compared to another (or no) intervention.

Therefore, the aim of this meta-research study was to investigate if the use of *p*-values, effect estimates, and clinical relevance differs between 2000 and 2018 in physiotherapy RCTs published in high quality influential journals (top 25%). Our secondary aim was to evaluate whether there is an association between the risk of bias of the studies and the incorrect use of *p*-values (i.e. baseline significance testing), and how clinical relevance was determined. This is because we assume that authors of studies with a lower risk of bias follow the reporting guidelines better.

Methods

Design

Meta-research study on the use of *p*-values, effect estimates (and 95% CI), and reporting and definition of clinical relevance in physiotherapy RCTs published in the years 2000 and 2018. The current study is part of a suite of research studies using the same sample of selected RCTs and was registered internally within the University of Technology Sydney, Discipline of Physiotherapy [19].

Ethics Approval

Not applicable as this involves a review of studies

Search strategy

We searched the databases Embase, Medline, and PubMed in May 2019. The search strategy was developed to identify RCTs with at least one physiotherapy intervention arm published in six high-ranked physiotherapy journals, all supporting the CONSORT-statement, restricted to publication years 2000 or 2018. Journals included were: (Ausn) Journal of Physiotherapy (J Physiother), Archives of Physical Medicine and Rehabilitation (Arch Phys Med Rehabil), Clinical Rehabilitation (Clin

Rehabil), Journal of Orthopedic and Sports Physical Therapy (J Orthop Sports Phys Ther), Physical Therapy (Phys Ther) and Spine. These journals were chosen based on SCImago Journal Rank (all Q1 = top 25%) across both years, suggesting a substantial influence within the physiotherapy profession. The search strategy was reviewed by a librarian. All articles retrieved in the search were imported into Covidence and duplicates were removed.

Study selection

Two independent assessors first screened each article by title and abstract, and then by the full texts. If required, a third assessor resolved conflicts. Articles were eligible if they were an RCT that used at least one physiotherapy intervention. The World Confederation of Physiotherapy Policy statement was used to determine whether the intervention was within the international scope of physiotherapy [20]. Studies were excluded if they were conference proceedings, editorials, reviews, published protocols, cost effectiveness analyses or secondary analyses of RCTs only, not performed on humans, or the full text could not be obtained.

Data Extraction

Data extraction. The following information was extracted from each included study: descriptive information (such as subdiscipline of physiotherapy practice, study population, sample size at randomisation and analysis); use of *p*-values, effect estimates and 95% CIs reported for baseline, between- and within-group analysis; whether clinical relevance was mentioned; and how clinical relevance was defined. Data was extracted from each article by two independent assessors with conflicts resolved by a third assessor.

Assessment of risk of bias. For all included studies, the risk of bias rating was performed using the PEDro scale obtained from the PEDro-database (Physiotherapy Evidence Database) or independently assessed by two assessors, when the score was not available. Conflicts in scoring were resolved by a third assessor. PEDro scale is considered to have good interrater reliability and convergent validity [21,22]. Ratings vary between 0 (very low quality (or high risk of bias)) to 10 (perfect quality (low risk of bias)). A score < 4 is considered 'poor', 4 to 5 'fair', 6 to 8 'good' and 9 to 10 'excellent' quality [21].

Statistical Analysis

First, we calculated frequencies and proportions for reporting of *p*-values, effect estimates, 95% CIs and clinical relevance. A *priori,* we defined that a difference of ≥20% between 2000 and 2018 was regarded as a meaningful difference [23]. For our secondary aim we calculated the correlation

(Pearson/Spearman correlation coefficient) between the PEDro score and a) the use of statistical significance testing at baseline and b) the mention of clinical relevance. We performed the analysis for the secondary aim in the trials of 2018 only as this dataset is the most recent representation of the literature. Correlation coefficients <0.20 were interpreted as no correlation, between 0.2 to 0.4 as low, 0.4 to 0.6 as moderate, 0.6 to 0.8 as high and above 0.8 as an almost perfect correlation [24,25]. Statistical analyses were performed using SPSS IBM 20.

Patient and Public involvement

No patients involved

Results

Search results

The search returned 1211 references, and after screening, 140 articles were included in the analysis (Figure 1). Of the 140 studies, 39 were published in 2000 and 101 in 2018 (Table 1). The number of published RCTs with at least one physiotherapy intervention was higher in 2018 compared to 2000 in Clin Rehabil, J Physiother, J Orthop Sports Phys Ther and Arch Phys Med Rehabil, while the number of published RCTs were similar in Spine and Phys Ther (Table 2). The RCTs were mainly performed in Europe/United Kingdom (n=51), USA/Canada (n=34), Australia/New Zealand (n=17) and Brazil (n=13).

Characteristics of included studies

Patient populations. Most studies were performed in musculoskeletal (50.7%) and neurological populations (30.7%) (Table 2). Other subdisciplines of physiotherapy were woman's health, oncology, and gerontology. The most common patient population in musculoskeletal studies included patients with low back pain (n=19) or neck pain (n=10). The most common patient populations in neurological studies were in stroke (n=22) and Parkinson's disease (n=7). Two journals (Spine and J Orthop Sports Phys Ther) published RCTs on musculoskeletal conditions only in both years, while the J Physiother did not publish any RCTs on musculoskeletal conditions in 2018.

Interventions. Of the 140 studies, most evaluated two interventions (n=115), while some evaluated three (n=21), or four or more interventions (n=4). Exercises or rehabilitation interventions (n=76; 54.2%) were the most common intervention evaluated followed by electrotherapy interventions (n=15, 10.7%). Most of the control interventions were exercise (n=32), followed by usual care (n=29), no treatment (n=26) or sham (n=16).

Sample size. The sample size in the studies ranged from 10 to 457 participants. The mean (standard deviation (SD)) sample size in all studies was 73.8 (62.2) at randomisation and 67.2 (58.6) in the analysis (Table 1). Between 2000 and 2018 the mean sample size across all journals was comparable, with a mean of 73-75 participants, but the difference between journals was large (Table 1).
In 2000 Spine published studies with an overall larger sample size (mean >125 participants) compared to the other journals (mean <65 participants). The sample size in the J Physiother and Phys They differed from 32 and 34 respectively in 2000, to over 100 participants, on average in 2018 (Table 2).

Risk of bias. Of the 140 articles, 15 (11%) had no PEDro-score and were rated by the researchers. Overall, the mean PEDro score was 6.6 (range from 3-10). Most studies (n=99; 70.7%) were of 'good' to 'excellent' quality (low risk of bias), n=31 (22.1%) was of 'fair' quality and 2 (1.4%) were of 'poor' quality (high risk of bias). The PEDro score differed slightly between 2000 and 2018, with a mean PEDro score of 5.8 in 2000 and 6.9 in 2018 (Table 1). The mean PEDro score in Spine did not differ between the years, while the PEDro score was higher in 2018, compared to 2000, in all other journals; with all included RCTs in the J Physiother in 2018 scoring 8/10 (Table 2).

Reporting prevalence

P-values. Most studies (n=128; 91.4%) used *p*-values to compare outcomes between groups (Table 1); one study (published in 2018) reported within-group differences only, nine studies reported only effect estimates and one study (published in 2000) did not report *p*-values or effect estimates.
The prevalence of *p*-values to determine between-group differences did not differ between 2000 and 2018 (92.3% and 91.1% respectively, Table 1). Of all studies that presented between-group *p*-values (n=130), 68 (52.3%) reported that the *p*-value was statistically significant, meaning <0.05, with a small difference between 2000 and 2018 (45.9% and 55.4% respectively). Of all studies reporting a non-significant difference regarding the primary outcome (n=62), 21 (33.3%) still reported positive findings in favour of the intervention, often based on the within-group differences or secondary outcomes. The number of studies that reported significance testing for baseline differences differed by 28.1%: 33.3% (95% CI: 19-50%) in 2000 and 61.4% (95% CI: 51-71%) in 2018.
The proportion of studies that reported (additional) within-group differences was 48.7% (95% CI: 32-65%) in 2000 and 55.4% (95% CI: 45-65%) in 2018 (Table 1). The J Physiother was the only journal where baseline statistical significance testing was not performed in 2018. The prevalence of *p*-values for between- and within-group differences decreased in J Physiother and J Orthop Sports Phys Ther by more than 20% (Table 2).

Effect estimates. Half of all studies (n=70, 50%) presented their results using an effect estimate (Table 1). The reporting of effect estimates for between-group analysis differed with 26.6% (30.8% (95% CI: 17-48%) in 2000 and 57.4% (95% CI: 47-67%) in 2018). The use of 95% CIs differed with 34% (20.5% (95% CI: 9-36%) in 2000 and 54.5% (95% CI: 44-64%) in 2018). Of the nine studies that reported only effect estimates (i.e., without *p*-values), seven were published in 2018. Overall, there was a meaningful difference (>20%) in the use of effect estimates (and 95% CIs) between 2000 and 2018, mainly due to the increases of >20% in Spine, J Physiother and Phys Ther journals.

Clinical relevance. Almost half of all studies (n=69; 49.3%) mentioned clinical relevance in their paper. In 25 studies, clinical relevance was related to the sample size calculation, but most of the studies mentioned clinical relevance (solely) in the discussion (Table 1). In 2018, only 23 studies (22.8%) defined clinically relevance and related it to the outcome. The overall mention of clinical relevance differed with 32.8% (25.6% (95% CI: 13-42%) in 2000 and 58.4% (95% CI: 48-68%) in 2018). Four journals showed a meaningful difference across years in mentioning clinical relevance (Table 2). The description of clinical relevance varied across studies, with 31 out of 69 (45%) studies clearly stating a minimal clinical important difference (MCID), mostly related to the sample size calculation, while others used the terms 'clinical change', 'minimal change', 'clinical meaningful change', 'clinically relevant difference', or 'significant clinical change' without specific reference to outcome data or cut-offs.

Risk of bias

The Pearson correlation coefficient between PEDro score and the use of statistical significance testing at baseline was -0.2 (Spearman: -0.23) in the studies in 2018 (figure 2). We found a low correlation between risk of bias and incorrect significance testing (baseline differences). This means that studies with a lower risk of bias are slightly less likely to present statistical significance testing at baseline. The Pearson correlation coefficient between the PEDro score and the mention of clinical relevance was 0.13 (Spearman: 0.14) in the studies in 2018. This means that there was no correlation between risk of bias and mention of clinical relevance.

Discussion

Main findings

Overall, we found that in the sample of physiotherapy journals investigated there was a high prevalence (>90%) of reporting *p*-values for the primary (between-group) analysis in both 2000 and 2018. Statistical significance testing for baseline differences differed between 28% in 2000 and 61.4% in 2018. Studies with lower risk of bias in 2018 tend to do slightly less statistical significance testing at baseline, indicating that the authors followed the reporting guidelines a bit better. Approximately half of all studies use statistical testing for within-group changes and there were no differences across years. The prevalence of reporting effect estimates, and the mention of clinical relevance differed >20% between 2000 and 2018, with it's reporting in almost 60% of all trials in 2018. However, many studies did not equate their study outcome to a known MCID.

Comparison with other studies

A previous study evaluating overall quality of methods in biomedical RCTs, including randomization, blinding and selective reporting, concluded that 59.3% of RCTs used inadequate methods (meaning scoring high risk of bias on one or more of the 6 Cochrane risk of bias items) and 35% of RCTs were poorly reported (meaning providing not enough information in the methods to decide on adequate or inadequate methods) [26]. Comparable findings have been found in physiotherapy RCTs in the PEDro database [21]. Whilst reporting of effect estimates in our selection of high-quality physiotherapy literature differs between 2000 and 2018, still most papers did not adhere to the reporting recommendations provided by the ASA and CONSORT-statements with regards to statistical significance testing and reliance on *p*-values to interpret results. Over a period of 18 years, presentation of effect estimates, and 95% CIs increased. Our results are consistent with another study that only evaluated the reporting of 95% CIs and found that these 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 [17]. However, in 2018, 42.6% of studies in our study still do not report the effect estimate, and solely present results using *p*-values. With an average increase of 2%, a one hundred percent compliance to the recommendations will only be achieved in 2049. Reporting of effect estimates (and CIs) are required if clinicians are to understand the magnitude and uncertainty of the treatment effect. Although the CONSORT-statement has been endorsed by these six major physiotherapy journals, in this study, only two journals (J Physiother, Phys Ther) successfully adhered to the reporting guidelines for effect estimates in 2018.

Although the reason for performing a RCT is to compare differences between randomised groups, about half of all studies also present the results of within-group analyses. Often participants in RCTs improve over time due to natural recovery or to the Hawthorne effect [27]. Therefore, it remains unclear why so many authors choose to test within-group differences in an RCT, and why journal editors permit authors to do so when it is conceivable that a reader may misinterpret the result.

The CONSORT-statement also recommends comparing baseline differences between groups, however statistical testing for baseline differences between randomized groups is not recommended [11,28]. The rationale is that when the randomization procedure is performed well, all differences at baseline are due to chance. Hypothesis testing at baseline means that we test the probability of a difference by chance, when we know these differences occur by chance and are therefore considered inappropriate and illogical [28,29]. We found that statistical significance testing for baseline differences had increased from 2000 to 2018, with over 60% of studies reporting *p*-values for baseline comparisons. Our results are higher than those in a previous study published in 2010 which found 38% of RCTs reported *p*-values for baseline differences in 114 RCTs published in leading medical journals [28]. A reason for this difference might be that the selection of the 114 RCTs came from four leading medical journals with higher impact factors than our six journals, and assuming their risk of bias was lower (though not assessed in that article) than in our sample.

Clinical relevance of outcomes is important when interpreting if the effects of an intervention are meaningful to patients [30]. Although the mention of clinical relevance increased over time, in 2018 only a small proportion of studies (n=23, 22.8%) related clinically relevance to their outcome, and most studies it was mentioned it in the discussion section only. Also, a wide variety of terminology was used, and the terms 'change' and 'difference' were used interchangeably in most studies. Recently, experts clarified the difference between these concepts more clearly [31]. They state that MCID are cross-sectional between-group differences, such as the difference between two intervention groups after treatment that are regarded clinically relevant, while minimal important changes (MIC) are longitudinal within-person changes in scores [31]. The lack of known clinically important values, particularly MCID for use in RCTs may be a barrier for researchers to report and interpret their findings in relation to clinical relevance. Future research that aims to determine MCIDs for core outcomes measures are warranted.

Performing underpowered studies is regarded as research waste [32,33]. The typical standardized effect estimate in physiotherapy trials is around 0.3 [34]. This is considered a small to medium effect estimate [35]. The sample size that on average should be sufficient to detect an effect estimate of 0.3 (in low back pain RCTs) is about 175 participants [36]. Almost all studies in our analysis had sample sizes that were too small to detect an effect estimate of 0.3. Nevertheless, about half the studies that presented between group *p*-values, reported statistical significance (using $p<0.05$). The mean sample size did not increase over time, although there was some variation between journals. This finding is a concern because sample sizes of physiotherapy RCTs remain small and therefore are likely underpowered.

Strengths & limitations

There are several limitations to our study. Firstly, the scope of physiotherapy practice is broad and may vary between countries. It is therefore possible that we may have missed some relevant publications or included publications that in other countries would not be defined as providing 'physiotherapy' intervention. Second, we selected publications from six long-standing influential physiotherapy journals. We assumed that these journals would publish the best RCTs, meaning that our findings might be more positive than if a sample was taken from the overall physiotherapy literature. Third, as the included RCTs from the six journals predominantly investigated musculoskeletal interventions, we cannot assume that our findings are representative of all physiotherapy research and subspecialties. Fourth, we arbitrarily defined a 20% difference as a meaningful difference. Unfortunately, we did not define what percentage of the literature should ideally report effect estimates or mention clinical relevance. In retrospect, that was pertinent to define. Lastly, we investigated reporting of *p*-values and effect estimates regardless of whether it was a primary or secondary outcome. However, we do not expect that our findings would differ majorly when only measured for the primary outcome.

Future Directions

Research is one of the pillars of evidence-based practice and plays a fundamental role in guiding treatment selection. 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 [37]. Unfortunately, the methodological quality of the RCTs in the PEDro database remains suboptimal [21]. Our findings confirm that the statistical reporting and use of clinical relevance in physiotherapy RCTs is also suboptimal. Researchers have an ethical obligation to accurately report findings to allow for evidence-based decision-making [7,38]. By 2018, authors should have been aware of reporting guidelines such as the CONSORT-statement and been obligated to adhere to publication guidelines [38]. The findings of our study show that there are some improvements in the physiotherapy literature, but there is still need for improvement concerning statistical reporting and reporting of clinical relevance. Overall, stronger incentives (or penalties) may be required to improve the quality and reporting of physiotherapy research.

Conclusion

The prevalence of the reporting of *p*-values remains high in physiotherapy research published in high ranked physiotherapy journals and the reporting of statistical significance testing for baseline differences was higher in 2018 compared to 2000. The prevalence of the reporting of effect estimates (and CI's) was >20% higher in 2018 compared to 2000 but was still reported in less than 60% of all publications. Our findings suggest that although reporting seems to have improved, there is still under-reporting of effect estimates. The prevalence of significance testing for baseline differences and within-group changes is also concerning, as it shows that authors do not completely understand the reason for randomisation in RCTs.

Acknowledgements:

Students of Master of Physiotherapy program of UTS assisted in searching, data extraction, and assessment of methodological quality: S. Rogan, D. Commerford, G. Milgate, K. Cummins, M. Beech, R. Briody, D. Hagtharp, L. Jovic, J. Lenn, and A Shah was the librarian that assisted with the search.

Author statement

Arianne P Verhagen: Conceptualization; Data curation; Formal analysis; Methodology; Supervision; Validation; Roles/Writing - original draft; Writing - review & editing. **Peter W Stubbs**: Data curation; Formal analysis; Validation; Roles/Writing - original draft; Writing - review & editing **Poonam Mehta**: Data curation; Supervision; Validation; Roles/Writing - original draft; Writing - review & editing. **David Kennedy**: Conceptualization; Roles/Writing - original draft; Writing - review & editing. **Anthony M Nasser**: Supervision; Roles/Writing - original draft; Writing - review & editing. **Camila Quel de Oliveira**: Roles/Writing - original draft; Writing - review & editing. **Joshua W Pate**: Roles/Writing - original draft; Writing - review & editing. **Ian W Skinner**: Data curation; Supervision; Roles/Writing - original draft; Writing - review & editing. **Alana B McCambridge**: Conceptualization; Data curation; Formal analysis; Methodology; Project administration; Resources; Software; Supervision; Validation; Roles/Writing - original draft; Writing - review & editing

Competing interests: None declared

Conflict of interest: AV was a member of the editorial board of the J Physiother (until 2020) and currently is an associate editor of the J Orthop Sports Phys Ther.

References

1. Ioannidis JPA, Fanelli D, Drake Dunne D, Goodman SN. Meta-research: Evaluation and Improvement of Research Methods and Practices. PLoS Biol 2015 13(10): e1002264. doi:10.1371/journal.pbio.1002264
2. Ioannidis JPA. Meta-research: why research on research matters. PLoS Biol 2018;16(3):e2005468. https//doi.org/10.1371/journal.pbio.2005468
3. ASA website: https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf. Last visited 21 September 2020
4. Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician. 2016;70(2):129-133
5. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p < 0.05". The American Statistician. 2019;73(sup1):1-19
6. 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
7. Verhagen AP, Ostelo RWJG, Rademaker A. Is the p value really so significant? Australian Journal of Physiotherapy 2004;50:261-2.
8. Sullivan GM, Feinn R. Using Effect Size-or Why the P Value Is Not Enough. J Grad Med Educ. 2012;4(3):279-282
9. Cohen J. The earth is round (p<. 05). In: *What if there were no significance tests?*: Routledge; 2016:69-82.
10. Herbert R. Research Note: significance testing and hypothesis testing: meaningless, misleading and mostly unnecessary. J Physiother 2019;65:178-181.
11. 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.
12. 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
13. 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
14. Armijo-Olivo S, Saltaji H, da Costa BR, Fuentes J, Ha C, Cummings GG. What is the influence of randomisation sequence generation and allocation concealment on treatment effects of physical therapy trials? A meta-epidemiological study. BMJ Open. 2015;5(9):e008562. doi: 10.1136/bmjopen-2015-008562.
15. Armijo-Olivo S, Fuentes J, da Costa BR, Saltaji H, Ha C, Cummings GG. Blinding in Physical Therapy Trials and Its Association with Treatment Effects: A Meta-epidemiological Study. Am J Phys Med Rehabil. 2017;96(1):34-44. doi: 10.1097/PHM.0000000000000521.
16. de Almeida MO, Saragiotto BT, Maher C, Costa LOP. Allocation Concealment and Intention-To-Treat Analysis Do Not Influence the Treatment Effects of Physical Therapy Interventions in Low Back Pain Trials: a Meta-epidemiologic Study. Arch Phys Med Rehabil. 2019;100(7):1359-1366. doi: 10.1016/j.apmr.2018.12.036.

17. 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

18. Thiese MS, Ronna B, Ott U. P value interpretations and considerations. Journal of thoracic disease. 2016;8(9):E928-e931

19. McCambridge AB, Nasser AM, Mehta P, Stubbs PW, Verhagen AP. METAPHoR: meta-research in physiotherapy trials – has reporting of physiotherapy interventions improved? JOSPT 2021, Accepted

20. Policy Statement: Description of Physical Therapy [press release]. World Confederation for Physical Therapy 2019

21. 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.

22. Cashin AG, McAuley JH. Clinimetrics: Physiotherapy Evidence Database (PEDro) Scale. J Physiother. 2020;66(1):59. doi: 10.1016/j.jphys.2019.08.005.

23. Moseley AM, Herbert RD, Maher CG, Sherrington C, Elkins MR. Reported quality of randomized controlled trials of physiotherapy interventions has improved over time. Journal of Clinical Epidemiology, 2011;64(6):594-601.

24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-74.

25. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Family medicine. 2005;37:360-3.

26. Catillon M. Trends and predictors of biomedical research quality, 1990-2015: a meta-research study. BMJ Open. 2019;9(9):e030342. doi: 10.1136/bmjopen-2019-030342.

27. Sedgwick P, Greenwood N. Understanding the Hawthorne effect. BMJ, 351 (2015), p. h4672, 10.1136/bmj.h4672

28. 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.

29. 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.

30. 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

31. Kamper SJ. Interpreting Outcomes 3-Clinical Meaningfulness: Linking Evidence to Practice. J Orthop Sports Phys Ther. 2019;49(9):677-678. doi: 10.2519/jospt.2019.0705.

32. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, Michie S, Moher D, Wager E. Reducing waste from incomplete or unusable reports of biomedical research. Lancet. 2014;383(9913):267-76. doi: 10.1016/S0140-6736(13)62228-X.

33. Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gülmezoglu AM, Howells DW, Ioannidis JP, Oliver S. How to increase value and reduce waste when research priorities are set. Lancet. 2014;383(9912):156-65. doi: 10.1016/S0140-6736(13)62229-1.

34. Lamb SE, Lall R, Hansen Z, et al. A multicentred randomised controlled trial of a primary care-based cognitive behavioural programme for low back pain. The Back Skills Training (BeST) trial. Health Technol Assess 2010;14:1–253.

35. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2 nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates Inc; 1988.

36. Froud R, Rajendran D, Patel S, Bright P, Bjørkli T, Eldridge S, Buchbinder R, Underwood M. The Power of Low Back Pain Trials: A Systematic Review of Power, Sample Size, and Reporting of Sample Size Calculations Over Time, in Trials Published Between 1980 and 2012. Spine (Phila Pa 1976). 2017;42(11):E680-E686. doi: 10.1097/BRS.0000000000001953.

37. 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.

38. 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

Table 1: Characteristics of included studies published in the years 2000 and 2018.

	2000, n=39	**2018, n=101**	**Total, n=140**
Journals, *n (%)*			
Arch Phys Med Rehabil	11 (28.2%)	30 (29.6%)	41 (29.3%)
(A)J Physiother	2 (5.1%)	7 (6.9%)	9 (6.4%)
Clin Rehabil	5 (12.8%)	45 (44.6%)	50 (35.7%)
J Orthop Sports Phys Ther	4 (10.2%)	6 (5.9%)	10 (7.1%)
Phys Ther	6 15.4%)	6 (5.9%)	12 (8.6%)
Spine	11 (28.2%)	7 (6.9%)	18 (12.9%)
Subdiscipline, *n (%)*			
Musculoskeletal	26 (66.7%)	45 (44.6%)	71 (50.7%)
Neurological	7 (17.9%)	36 (35.6%)	43 (30.7%)
Cardiorespiratory	2 (5.1%)	9 (8.9%)	11 (7.9%)
Other	4 (10.2%)	11 (11%)	15 (10.7%)
PEDro score **(0-10)**, mean (SD); (range)	5.8 (1.4); (3-8)	6.9 (1.3); (4-10)	6.6 (1.4); (3-10)
Sample size, mean (SD)	74.5 (88.3)	73.6 (49.1)	73.8 (62.2)
Use of p-value, *n (%)*			
Significance testing at baseline	13 (33.3%)	62 (61.4%)	75 (53.6%)
P-value for between-group analysis	36 (92.3%)	92 (91.1%)	128 (91.4%)
P-value for within-group analysis	19 (48.7%)	56 (55.4%)	75 (53.6%)
Effect estimates, *n (%)*			
Effect estimates for between-group analysis	12 (30.8%)	58 (57.4%)	70 (50%)
Effect estimates for within-group analysis	4 (10.6%)	29 (28.7%)	33 (23.6%)
Confidence intervals for between-group analysis	8 (20.5%)	55 (54.5%)	63 (45%)
Confidence intervals for within-group analysis	3 (7.7%)	28 (27.7%)	31 (22.1%)
Clinical relevance, *n (%)*			
Mentioned	10/39 (25.6%)	59/101 (58.4%)	69/140 (49.3%)
Used for sample size calculation	1/10	24/59	25/69
Specified a value for their outcome	3/10	23/59	26/69
Mentioned in discussion	9/10	49/59	58/69

(A)J Physiother = (Australian) Journal of Physiotherapy; Arch Phys Med Rehabil = Archives of Physical Medicine and Rehabilitation; Clin Rehabil = Clinical rehabilitation; J Orthop Sports Phys Ther = Journal of Orthopaedic and Sports Physical Therapy, Phys Ther = Physical Therapy

Table 2: Outcome data per journal

	Arch Phys Med Rehabil		***(A)J Physiother***		***Clin Rehabil***		***J Orthop Sports Phys Ther***		***Phys Ther***		***Spine***	
	2000	2018	2000	2018	2000	2018	2000	2018	2000	2018	2000	2018
N of studies	11	30	2	7	5	45	4	6	6	6	11	7
PEDro, mean (range)	5.6 (3-8)	6.7 (5-9)	6.5 (6-7)	8 (8-8)	5.6 (4-7)	7 (4-9)	5.5 (4-7)	6.8 (4-10)	5.3 (4-8)	6.7 (4-8)	6.3 (4-8)	6.3 (5-7)
Sample size, mean (range)	49.3 (10-135)	62.6 (19-180)	34 (28-40)	107.7 (46-198)	61.2 (27-98)	64.7 (19-181)	24.6 (10-52)	48.7 (24-103)	32.5 (18-44)	127.2 (52-208)	152.6 (21-457)	127.3 (23-304)
P-values												
Sign testing at baseline	3/11	18/30	1/2	0	2/5	33/45	1/4	2/6	1/6	3/6	5/11	6/7
Between-groups	10/11	29/30	2/2	4/7	5/5	44/45	4/4	4/6	6/6	6/6	9/11	7/7
Within-groups	3/11	18/30	0	1/7	3/5	26/45	3/4	3/6	4/6	3/6	4/11	4/7
Effect estimates												
Between-group	3/11	14/30	1/2	7/7	2/5	25/45	1/4	2/6	2/6	6/6	3/11	4/7
Within-group	1/11	5/30	0	2/7	1/5	17/45	1/4	1/6	1/6	3/6	0	1/7
Clinical relevance												
Mentioned	2/11	15/30	2/2	4/7	1/5	28/45	1/4	5/6	1/6	5/6	3/11	2/7
Related to outcome	0	5/15	1/2	2/4	0	10/28	0	2/5	1/6	3/5	1/3	1/2

(A)J Physiother = (Australian) Journal of Physiotherapy; Arch Phys Med Rehabil= Archives of Physical Medicine and Rehabilitation; Clin Rehabil = Clinical rehabilitation; J Orthop Sports Phys Ther = Journal of Orthopaedic and Sports Physical Therapy, Phys Ther = Physical Therapy; PEDro = Physiotherapy Evidence Database

Figure 1: Flow diagram of study selection

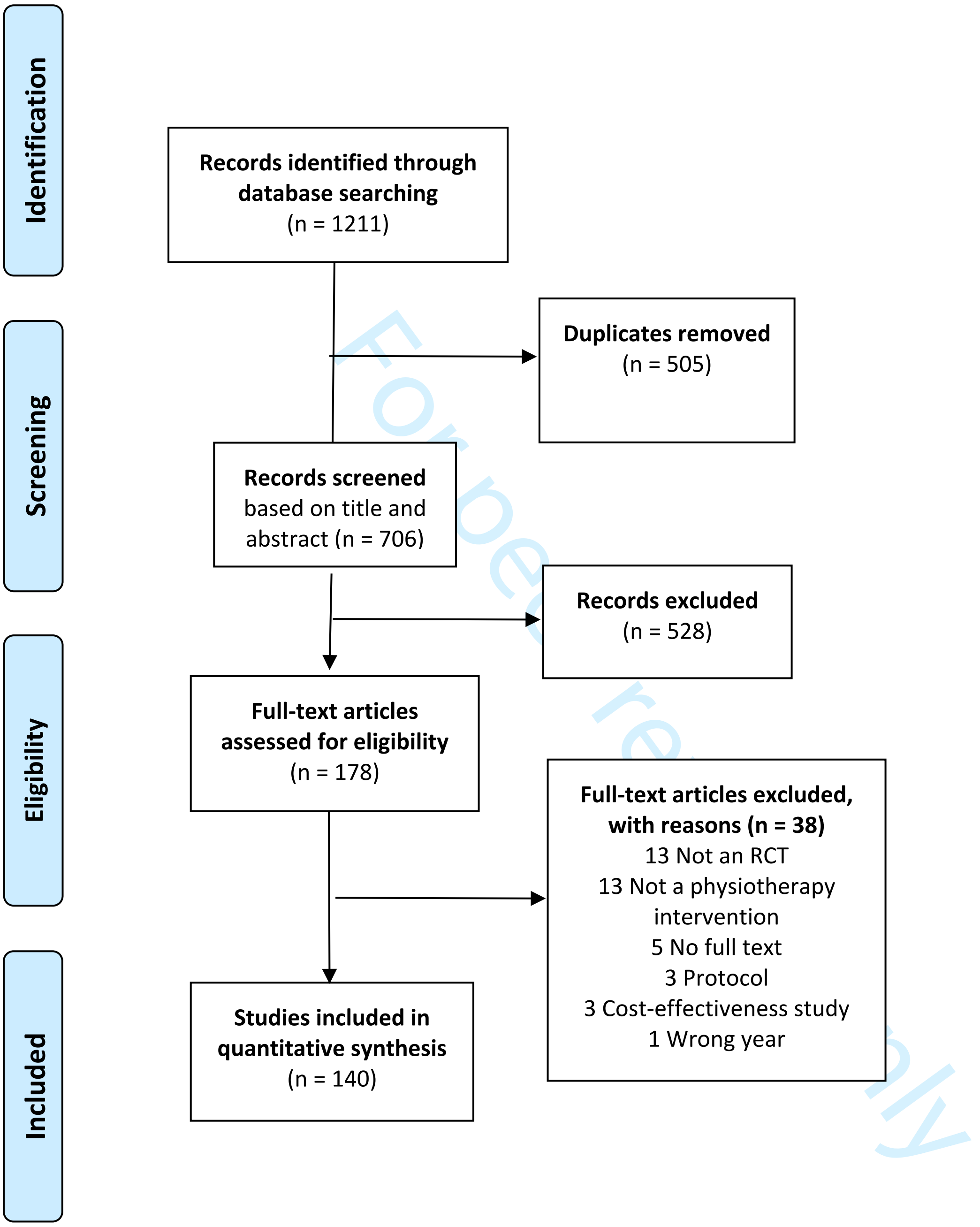
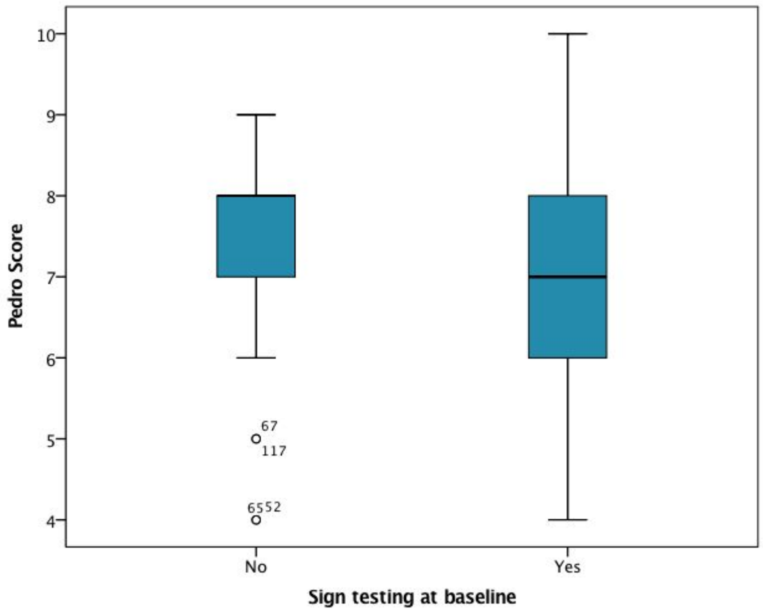

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 1211)
Duplicates removed (n = 505)
Records screened based on title and abstract (n = 706)
Records excluded (n = 528)
Full-text articles assessed for eligibility (n = 178)
Full-text articles excluded, with reasons (n = 38)
13 Not an RCT
13 Not a physiotherapy intervention
5 No full text
3 Protocol
3 Cost-effectiveness study
1 Wrong year
Studies included in quantitative synthesis (n = 140)

Figure 2: Boxplot on association between risk of bias (methodological quality (PEDro score)) and statistical significance testing for baseline variables.



Median, 25% quartile and range

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	1
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	5,6
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5,6
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	5,6
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5,6
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	5,6
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	5,6
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	6
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	6
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	6,7

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	6
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	6,7
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	7
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	7
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	8
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	8,9
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	8,9
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	8,9
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	8,9
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	9,10
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	11,12
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	10-12
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	1

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.

BMJ Open

Comparison between 2000 and 2018 on the reporting of statistical significance and clinical relevance in physiotherapy clinical trials in six major physiotherapy journals

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-054875.R1
Article Type:	Original research
Date Submitted by the Author:	19-Oct-2021
Complete List of Authors:	Verhagen, Arianne; University of Technology Sydney Stubbs, Peter ; University of Technology Sydney, Graduate School of Health, Discipline of Physiotherapy Mehta, Poonam; University of Technology Sydney Kennedy, David; University of Technology Sydney Nasser, Anthony M; University of Technology Sydney Quel de Oliveira, Camila; University of Technology Sydney Pate, Joshua W; University of Technology Sydney Skinner, Ian W; University of Technology Sydney; Charles Sturt University McCambridge, Alana B; University of Technology Sydney
<b>Primary Subject Heading</b>:	Epidemiology
Secondary Subject Heading:	Evidence based practice, Research methods
Keywords:	EPIDEMIOLOGY, EDUCATION & TRAINING (see Medical Education & Training), PRIMARY CARE, Clinical trials < THERAPEUTICS, Rehabilitation medicine < INTERNAL MEDICINE

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Comparison between 2000 and 2018 on the reporting of statistical significance and clinical relevance in physiotherapy clinical trials in six major physiotherapy journals

Authors:

Arianne P Verhagen[1], Peter W Stubbs[1], Poonam Mehta[1], David Kennedy[1], Anthony M Nasser[1], Camila Quel de Oliveira[1], Joshua W Pate[1], Ian W Skinner[1,2], Alana B McCambridge[1]

Affiliation:

[1] Graduate School of Health, Discipline of Physiotherapy, University of Technology Sydney, Sydney, Australia

[2] School of Community Health, Charles Sturt University, Port Macquarie, Australia

Funding:

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Protocol: https://doi.org/10.26195/ebqp-8a65

Corresponding Author:

Prof Arianne P Verhagen, Head of Discipline of Physiotherapy, Graduate School of Health, University of Technology Sydney Australia

E-mail: Arianne.Verhagen@uts.edu.au

Word count:

Abstract: 262; text: 3917

Number of references: 43; tables: 2; figures: 2

ABSTRACT

Design: meta-research

Objective: To compare the prevalence of reporting *p*-values, effect estimates and clinical relevance in physiotherapy randomized controlled trials (RCTs) published in the years 2000 and 2018.

Methods: We performed a meta-research study of physiotherapy RCTs obtained from six major physiotherapy peer-reviewed journals that were published in the years 2000 and 2018. We searched the databases Embase, Medline, and PubMed in May 2019, and extracted data on the study characteristics and whether articles reported on statistical significance, effect estimates and confidence intervals for baseline, between-group, and within-group differences, and clinical relevance. Data were presented using descriptive statistics and inferences were made based on proportions. A 20% difference between 2000 and 2018 was regarded as a meaningful difference.

Results. We found 140 RCTs: 39 were published in 2000 and 101 in 2018. Overall, there was a high prevalence (>90%) of reporting *p*-values for the main (between-group) analysis, with no difference between years. Statistical significance testing was frequently used for evaluating baseline differences, increasing from 28% in 2000 to 61.4% in 2018. The prevalence of reporting effect estimates, confidence intervals and the mention of clinical relevance increased from 2000 to 2018 by 26.6%, 34% and 32.8% respectively. Despite an increase in use in 2018, over 40% of RCTs failed to report effect estimates, confidence intervals, and clinical relevance of results.

Conclusion. The prevalence of using *p*-values remains high in physiotherapy research. Although the proportion of reporting effect estimates, confidence intervals, and clinical relevance is higher in 2018 compared to 2000, many publications still fail to report and interpret study findings in this way.

Key words: Randomized clinical trials, Physiotherapy, reporting statistics, reporting clinical relevance

Strengths and Limitations

- This meta-research study will provide clear insight in the prevalence of (incorrect) use of p-values, and the prevalence of the use of effect estimates and clinical relevancy of outcomes
- We selected publications from six long-standing influential physiotherapy journals, assuming we select the best studies
- We defined a 20% difference as a meaningful difference
- We investigated reporting of *p*-values and effect estimates regardless of whether it was a primary or secondary outcome.

Introduction

As high-quality physiotherapy research needs to be clear, transparent, reproducible, and well written to inform clinical practice, it is important for clinicians to be confident in the methodological quality of physiotherapy research. Meta-research is a relatively new scientific discipline that explores how research is performed, reported, reproduced, evaluated, and incentivised [1,2]. As all scientific research is prone to bias, it is important that each profession critically evaluates its own research methods, standards of reporting, and validity of the outcomes [3].

Continuing discussions about the use (and misuse) of the *p*-value prompted the American Statistical Association (ASA) to recommend in 2016 that authors avoid statements on statistical significance and interpretation of outcomes using a p-value as an arbitrary threshold [4,5,6]. Traditionally, the *p*-value has been used in randomised clinical trials (RCTs) in conjunction with the null hypothesis testing to answer study questions related to the effectiveness of interventions by dichotomising results as significant or not significant [7]. Although valuable if interpreted correctly, null hypothesis testing has its limitations; it does **not** measure the probability of the truth of the null hypothesis, it does **not** measure the size or magnitude of an effect, and its replicability is poor [4,8-11]. The recommendation of the ASA is endorsed by many academic journals, nevertheless, authors continue to conclude whether an intervention is effective and should be used clinically by a dichotomous interpretation based on *p*-values.

Well conducted and large RCTs are considered high quality evidence and reporting of RCTs should be guided by the CONSORT-statement (Consolidated Standards of Reporting Trials) [12]. There are several recommendations in the CONSORT-statement regarding the reporting and appropriate use of *p*-values. For example, authors should not report results solely as *p*-values and are encouraged to (also) use effect estimates and 95% confidence intervals (95% CIs) [12]. The advantage of effect estimates is their ability to demonstrate the strength and the direction of the effect, and the 95% CIs provide a range of values between which the estimated true effect estimate lies [11,13,14]. Nevertheless, a dichotomized interpretation of the confidence interval (CI) should be discouraged; it allows for discussing the accuracy, precision and/or relevance of the effect estimate. Clinical relevance is another parameter used to interpret the magnitude of the effect, and to deem if a finding is clinically meaningful. Clinical relevance (or a clinically meaningful/worthwhile change, a minimum important difference (MID) or a minimal clinical important difference (MICD)) is regarded the threshold value for which any change (or larger) in for instance pain or disability is considered meaningful to patients [15].

According to the CONSORT-statement, authors should also compare baseline participant characteristics [12]. However, it discourages statistical significance testing of baseline covariates between randomized groups, as by using a proper randomization procedure all differences are based on chance. In addition, conclusions of a RCT should primarily be based on a between-group analysis by comparing post-intervention (and follow-up) outcomes between the groups or the between-group changes from baseline. Studies can additionally, with consideration, compare outcomes before and after the intervention using a 'within-group' analysis.

Previous meta-research within physiotherapy has investigated the use of randomization, blinding or intention-to-treat analysis [16-18] and one study evaluated the reporting of 95% CIs only [19]. To our knowledge, no study has examined the use of *p*-values, effect estimates or measures of clinical relevance in the physiotherapy literature before and after the CONSORT-statement was published in 2010. When selecting treatments, physiotherapists must be aware that statistical significance does not equate to clinical relevance [20]. Presenting effect estimates and precision of the effect (using 95% CIs) will also allow clinicians to consider how much a patient is likely to benefit from a given intervention compared to another (or no) intervention.

Therefore, the aim of this meta-research study was to investigate if the use of *p*-values, effect estimates, and clinical relevance differs between 2000 and 2018 in physiotherapy RCTs published in high quality influential journals (top 25%). Our secondary aim was to evaluate whether there is an association between the methodological quality of the studies and the incorrect use of *p*-values (i.e. baseline significance testing), and how clinical relevance was determined. This is because we assume that authors of studies with a higher methodological quality follow the reporting guidelines better.

Methods

Design

Meta-research study on the use of *p*-values, effect estimates (and 95% CI), and reporting and definition of clinical relevance in physiotherapy RCTs published in the years 2000 and 2018. The current study is part of a suite of research studies using the same sample of selected RCTs and was registered internally within the University of Technology Sydney, Discipline of Physiotherapy [21].

Ethics Approval

Not applicable as this involves a review of studies

Search strategy

We searched the databases Embase, Medline, and PubMed on the 24th of May 2019 (see appendix). The search strategy was developed to identify RCTs with at least one physiotherapy intervention arm published in six high-ranked physiotherapy journals, all supporting the CONSORT-statement, restricted to publication years 2000 or 2018. Journals included were: (Aus) Journal of Physiotherapy (J Physiother), Archives of Physical Medicine and Rehabilitation (Arch Phys Med Rehabil), Clinical Rehabilitation (Clin Rehabil), Journal of Orthopedic and Sports Physical Therapy (J Orthop Sports Phys Ther), Physical Therapy (Phys Ther) and Spine. These journals were chosen based on SCImago Journal Rank (all Q1 = top 25%) across both years, suggesting a substantial influence within the physiotherapy profession. The search strategy was reviewed by a librarian. All articles retrieved in the search were imported into Covidence and duplicates were removed.

Study selection

Two independent assessors first screened each article by title and abstract, and then by the full texts. If required, a third assessor resolved conflicts. Articles were eligible if they were an RCT that used at least one physiotherapy intervention. The World Confederation of Physiotherapy (WCPT) Policy statement was used to determine whether the intervention was within the international scope of physiotherapy [22]. Studies were excluded if they were conference proceedings, editorials, reviews, published protocols, cost effectiveness analyses or secondary analyses of RCTs only, not performed on humans, or the full text could not be obtained.

Data Extraction

Data extraction. The following information was extracted from each included study: descriptive information (such as subdiscipline of physiotherapy practice, study population, sample size at randomisation and analysis); use of *p*-values, effect estimates and 95% CIs reported for baseline, between- and within-group analysis; whether clinical relevance was mentioned (as well as synonyms, such as clinically important difference/change, minimal clinical differences, clinical significance, clinically worthwhile difference etc); and how clinical relevance was defined. Data was extracted from each article by two independent assessors with conflicts resolved by a third assessor.

Assessment of methodological quality. For all included studies, the methodological quality assessment was performed using the PEDro scale obtained from the PEDro-database (Physiotherapy Evidence Database) or independently assessed by two assessors, when the score was not available. Conflicts in scoring were resolved by a third assessor. PEDro scale is considered to have good interrater reliability and convergent validity [23,24].

Statistical Analysis

First, we calculated frequencies and proportions for reporting of *p*-values, effect estimates, 95% CIs and clinical relevance. A *priori,* we defined that a difference of ≥20% between 2000 and 2018 was regarded as a meaningful difference [25]. For our secondary aim we calculated the correlation (Pearson/Spearman correlation coefficient) between the PEDro score and a) the use of statistical significance testing at baseline and b) the mention of clinical relevance. We performed the analysis for the secondary aim in the trials of 2018 only as this dataset is the most recent representation of the literature. Correlation coefficients <0.20 were interpreted as no correlation, between 0.2 to 0.4 as low, 0.4 to 0.6 as moderate, 0.6 to 0.8 as high and above 0.8 as an almost perfect correlation [26,27]. Statistical analyses were performed using SPSS IBM 20.

Patient and Public involvement

No patients involved

Results

Search results

The search returned 1211 references, and after screening, 140 articles were included in the analysis (Figure 1). Of the 140 studies, 39 were published in 2000 and 101 in 2018 (Table 1).

Please insert figure 1 here

The number of published RCTs with at least one physiotherapy intervention was higher in 2018 compared to 2000 in Clin Rehabil, J Physiother, J Orthop Sports Phys Ther and Arch Phys Med Rehabil, while the number of published RCTs were similar in Spine and Phys Ther (Table 2). The RCTs were mainly performed in Europe/United Kingdom (n=51), USA/Canada (n=34), Australia/New Zealand (n=17) and Brazil (n=13).

Please insert table 1 here

Characteristics of included studies

<u>*Patient populations*</u>. Most studies were performed in musculoskeletal (50.7%) and neurological populations (30.7%) (Table 2). Other subdisciplines of physiotherapy were woman's health, oncology, and gerontology. The most common patient population in musculoskeletal studies were patients with low back pain (n=19) or neck pain (n=10). The most common patient populations in neurological

studies were in stroke (n=22) and Parkinson's disease (n=7). Two journals (Spine and J Orthop Sports Phys Ther) published RCTs on musculoskeletal conditions only in both years, while the J Physiother did not publish any RCTs on musculoskeletal conditions in 2018.

Please insert table 2 here

Interventions. Of the 140 studies, most evaluated two interventions (n=115), while some evaluated three (n=21), or four or more interventions (n=4). Exercises or rehabilitation interventions (n=76; 54.2%) were the most common intervention evaluated followed by electrotherapy interventions (n=15, 10.7%). Most of the control interventions were exercise (n=32), followed by usual care (n=29), no treatment (n=26) or sham (n=16).

Sample size. The sample size in the studies ranged from 10 to 457 participants. The mean (standard deviation (SD)) sample size in all studies was 73.8 (62.2) at randomisation and 67.2 (58.6) in the analysis (Table 1). Between 2000 and 2018 the mean sample size across all journals was comparable, with a mean of 73-75 participants, but the difference between journals was large (Table 1).
In 2000 Spine published studies with an overall larger sample size (mean >125 participants) compared to the other journals (mean <65 participants). The sample size in the J Physiother and Phys They differed from 32 and 34 respectively in 2000, to over 100 participants, on average in 2018 (Table 2).

Methodological quality. Of the 140 articles, 15 (11%) had no PEDro-score and were rated by the researchers. Overall, the mean PEDro score was 6.6 (range from 3-10). The PEDro score differed slightly between 2000 and 2018, with a mean PEDro score of 5.8 in 2000 and 6.9 in 2018 (Table 1). The mean PEDro score in Spine did not differ between the years, while the PEDro score was higher in 2018, compared to 2000, in all other journals; with all included RCTs in the J Physiother in 2018 scoring 8/10 (Table 2).

Reporting prevalence

Most studies (n=128; 91.4%) used *p*-values to compare outcomes between groups (Table 1); one study (published in 2018) reported within-group differences only, nine studies reported only effect estimates and one study (published in 2000) did not report *p*-values or effect estimates. Complete reporting (presenting *p*-values, effect estimates and 95%CI on between group difference, and refraining from baseline sign testing), was observed in 5 studies (12.8%) in 2000 and 20 studies (19.8%) in 2018.

P-values.

The prevalence of *p*-values to determine between-group differences did not differ between 2000 and 2018 (92.3% and 91.1% respectively, Table 1). Of all studies that presented between-group *p*-values (n=130), 68 (52.3%) reported that the *p*-value was statistically significant, meaning <0.05, with a small difference between 2000 and 2018 (45.9% and 55.4% respectively). Of all studies reporting a non-significant difference regarding the primary outcome (n=62), 21 (33.3%) still reported positive findings in favour of the intervention, often based on the within-group differences or secondary outcomes. The number of studies that reported significance testing for baseline differences differed by 28.1%: 33.3% (95% CI: 19-50%) in 2000 and 61.4% (95% CI: 51-71%) in 2018.

The proportion of studies that reported (additional) within-group differences was 48.7% (95% CI: 32-65%) in 2000 and 55.4% (95% CI: 45-65%) in 2018 (Table 1). The J Physiother was the only journal where baseline statistical significance testing was not performed in 2018. The prevalence of *p*-values for between- and within-group differences decreased in J Physiother and J Orthop Sports Phys Ther by more than 20% (Table 2).

Effect estimates. Half of all studies (n=70, 50%) presented their results using an effect estimate (Table 1). The reporting of effect estimates for between-group analysis differed with 26.6% (30.8% (95% CI: 17-48%) in 2000 and 57.4% (95% CI: 47-67%) in 2018). The use of 95% CIs differed with 34% (20.5% (95% CI: 9-36%) in 2000 and 54.5% (95% CI: 44-64%) in 2018). Of the nine studies that reported only effect estimates (i.e., without *p*-values), seven were published in 2018. Overall, there was a meaningful difference (>20%) in the use of effect estimates (and 95% CIs) between 2000 and 2018, mainly due to the increases of >20% in Spine, J Physiother and Phys Ther journals.

Clinical relevance. Almost half of all studies (n=69; 49.3%) mentioned clinical relevance in their paper. In 25 studies, clinical relevance was related to the sample size calculation, but most of the studies mentioned clinical relevance (solely) in the discussion (Table 1). In 2018, only 23 studies (22.8%) defined clinically relevance and related it to the outcome. The overall mention of clinical relevance differed with 32.8% (25.6% (95% CI: 13-42%) in 2000 and 58.4% (95% CI: 48-68%) in 2018). Four journals showed a meaningful difference across years in mentioning clinical relevance (Table 2). The description of clinical relevance varied across studies, with 31 out of 69 (45%) studies clearly stating a minimal clinical important difference (MCID), mostly related to the sample size calculation, while others used the terms 'clinical change', 'minimal change', 'clinical meaningful change', 'clinically relevant difference', or 'significant clinical change' without specific reference to outcome data or cut-offs.

Methodological quality

The Pearson correlation coefficient between PEDro score and the use of statistical significance testing at baseline was -0.2 (Spearman: -0.23) in the studies in 2018 (see figure 2). We found a low correlation between methodological quality and incorrect significance testing (baseline differences). This means that studies with a higher methodological quality were slightly less likely to present statistical significance testing at baseline. The Pearson correlation coefficient between the PEDro score and the mention of clinical relevance was 0.13 (Spearman: 0.14) in the studies in 2018. This means that there was no correlation between methodological quality and mention of clinical relevance.

Please insert figure 2 here

Discussion

Main findings

Overall, we found that in the sample of physiotherapy journals investigated there was a high prevalence (>90%) of reporting *p*-values for the primary (between-group) analysis in both 2000 and 2018. Statistical significance testing for baseline differences differed between 28% in 2000 and 61.4% in 2018. Studies with higher methodological quality in 2018 tend to do slightly less statistical significance testing at baseline. Approximately half of all studies use statistical testing for within-group changes and there were no differences across years. The prevalence of reporting effect estimates, and the mention of clinical relevance differed >20% between 2000 and 2018, with it's reporting in almost 60% of all trials in 2018. However, many studies did not equate their study outcome to a known MCID. Although the CONSORT-statement has been endorsed by these six major physiotherapy journals, in this study, only two journals (J Physiother, Phys Ther) successfully adhered to the reporting guidelines for effect estimates in 2018.

Comparison with other studies

A previous study evaluating overall quality of methods in biomedical RCTs, including randomization, blinding and selective reporting, concluded that 59.3% of RCTs used inadequate methods (meaning scoring high risk of bias on one or more of the 6 Cochrane risk of bias items) and 35% of RCTs were poorly reported (meaning providing not enough information in the methods to decide on adequate or inadequate methods) [28]. Comparable findings have been found in physiotherapy RCTs in the PEDro database [23] and evaluation of manual therapy trials [29,30]. Whilst reporting of effect estimates in our selection of high-quality physiotherapy literature differs between 2000 and 2018, still most papers did not adhere to the reporting recommendations provided by the ASA and CONSORT-statements with regards to statistical significance testing and reliance on *p*-values to interpret results. Over a period of 18 years, presentation of effect estimates, and 95% CIs increased. Our results are consistent with another study that only evaluated the reporting of 95% CIs and found that these 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 [19]. However, in 2018, 42.6% of studies in our study still do not report the effect estimate, and solely present results using *p*-values. With an average increase of 2%, a one hundred percent compliance to the recommendations will only be achieved in 2049. Reporting of effect estimates (and CIs) are required if clinicians are to understand the magnitude and uncertainty of the treatment effect.

Although the reason for performing a RCT is to compare differences between randomised groups, about half of all studies also presented the results of within-group analyses. Often participants in RCTs improve over time due to e.g. natural recovery or to the Hawthorne effect [31]. Therefore, it remains unclear why so many authors choose to test within-group differences in an RCT, and why journal editors permit authors to do so when it is conceivable that a reader may misinterpret the result.

The CONSORT-statement also recommends comparing baseline differences between groups, however statistical testing for baseline differences between randomized groups is not recommended [12,32]. The rationale is that when the randomization procedure is performed well, all differences at baseline are due to chance. Hypothesis testing at baseline means that we test the probability of a difference by chance, when we know these differences occur by chance and are therefore considered inappropriate and illogical [32,33]. We found that statistical significance testing for baseline differences had increased from 2000 to 2018, with over 60% of studies reporting *p*-values for baseline comparisons. Our results are higher than those in a previous study published in 2010 which found 38% of RCTs reported *p*-values for baseline differences in 114 RCTs published in leading medical journals [32]. A reason for this difference might be that the selection of the 114 RCTs came from four leading medical journals with higher impact factors than our six journals, and assuming their risk of bias was lower (though not assessed in that article) than in our sample. The prevalence of significance testing for baseline differences and within-group changes is concerning, as it shows that authors do not completely understand the reason for randomisation in RCTs.

Clinical relevance of outcomes is important when interpreting if the effects of an intervention are meaningful to patients [34]. Although the mention of clinical relevance increased over time, in 2018 only a small proportion of studies (n=23, 22.8%) related clinically relevance to their outcome, and most studies it was mentioned it in the discussion section only. Also, a wide variety of terminology was used, and the terms ‘change’ and ‘difference’ were used interchangeably in most studies. Recently, experts clarified the difference between these concepts more clearly [35]. They state that MCID are cross-sectional between-group differences, such as the difference between two intervention groups after treatment that are regarded clinically relevant, while minimal important changes (MIC) are longitudinal within-person changes in scores [35]. The lack of known clinically important values, particularly MCID for use in RCTs may be a barrier for researchers to report and interpret their findings in relation to clinical relevance. Future research that aims to determine MCIDs for core outcomes measures are warranted.

Strengths & limitations

There are several limitations to our study. Firstly, the scope of physiotherapy practice is broad and may vary between countries. It is therefore possible that we may have missed some relevant publications or included publications that in other countries would not be defined as providing ‘physiotherapy’ intervention. As we have used the WCPT definitions as selection criteria we assume this will not potentially bias our results. Second, we selected publications from six long-standing influential physiotherapy journals. We assumed that these journals would publish the best RCTs, meaning that our findings might be more positive (meaning a higher percentage of improvement in 2018) than if a sample was taken from the overall physiotherapy literature. Third, as the included RCTs from the six journals predominantly investigated musculoskeletal interventions, we cannot assume that our findings are representative of all physiotherapy research and subspecialties. Fourth, we defined a 20% difference as a meaningful difference based on a previous study [25]. Unfortunately, we did not define what percentage of the literature should ideally report effect estimates or mention clinical relevance. In retrospect, that was pertinent to define. Fifth, as the number of published RCTs in 2018 was over twice as much as in 2000, this imbalance might have influenced our results, as results from a smaller number of studies are often a bit less precise. Lastly, we investigated reporting of *p*-values and effect estimates regardless of whether it was a primary or secondary outcome. However, we do not expect that our findings would differ majorly when only measured for the primary outcome.

Future Directions

Research is one of the pillars of evidence-based practice and plays a fundamental role in guiding treatment selection. 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 [36]. Unfortunately, the methodological quality of the RCTs in the PEDro database remains suboptimal [23]. Our findings confirm that the statistical reporting and use of clinical relevance in physiotherapy RCTs is also suboptimal. To further help authors, a consensus-based reporting checklist for primary outcomes in RCTs is currently under development: InsPECT statement, specifically focussing on reporting of outcomes in a transparent way [37].

Researchers have an ethical obligation to accurately report findings to allow for evidence-based decision-making [8,38]. By 2018, authors should have been aware of reporting guidelines such as the CONSORT-statement and been obligated to adhere to publication guidelines [38]. The findings of our study show that there are some improvements in the physiotherapy literature, but there is still need for improvement concerning statistical reporting and reporting of clinical relevance. Overall, stronger incentives (or penalties) may be required to improve the quality and reporting of physiotherapy research.

Performing underpowered studies is regarded as research waste [39,40]. The typical standardized effect estimate in physiotherapy trials is around 0.3 [41]. This is considered a small to medium effect estimate [42]. The sample size that on average should be sufficient to detect an effect estimate of 0.3 (in low back pain RCTs) is about 175 participants [43]. Almost all studies in our analysis had sample sizes that were too small to detect an effect estimate of 0.3. Nevertheless, about half the studies that presented between group *p*-values, reported statistical significance (using $p<0.05$). The mean sample size did not increase over time, although there was some variation between journals. This finding is a concern because sample sizes of physiotherapy RCTs remain small and therefore are likely underpowered. We strongly recommend future studies to be of sufficient power.

Conclusion

The prevalence of the reporting of *p*-values remains high in physiotherapy research published in high ranked physiotherapy journals and the reporting of statistical significance testing for baseline differences was higher in 2018 compared to 2000. The prevalence of the reporting of effect estimates (and CI's) was >20% higher in 2018 compared to 2000 but was still reported in less than 60% of all publications. Our findings suggest that although reporting seems to have improved, there is still under-reporting of effect estimates.

Acknowledgements:

Students of Master of Physiotherapy program of UTS assisted in searching, data extraction, and assessment of methodological quality: S. Rogan, D. Commerford, G. Milgate, K. Cummins, M. Beech, R. Briody, D. Hagtharp, L. Jovic, J. Lenn, and A Shah was the librarian that assisted with the search.

Author statement

Arianne P Verhagen: Conceptualization; Data curation; Formal analysis; Methodology; Supervision; Validation; Roles/Writing - original draft; Writing - review & editing. **Peter W Stubbs**: Data curation; Formal analysis; Validation; Roles/Writing - original draft; Writing - review & editing **Poonam Mehta**: Data curation; Supervision; Validation; Roles/Writing - original draft; Writing - review & editing. **David Kennedy**: Conceptualization; Roles/Writing - original draft; Writing - review & editing. **Anthony M Nasser**: Supervision; Roles/Writing - original draft; Writing - review & editing. **Camila Quel de Oliveira**: Roles/Writing - original draft; Writing - review & editing. **Joshua W Pate**: Roles/Writing - original draft; Writing - review & editing. **Ian W Skinner**: Data curation; Supervision; Roles/Writing - original draft; Writing - review & editing. **Alana B McCambridge**: Conceptualization; Data curation; Formal analysis; Methodology; Project administration; Resources; Software; Supervision; Validation; Roles/Writing - original draft; Writing - review & editing

Competing interests: None declared

Conflict of interest: AV was a member of the editorial board of the J Physiother (until 2020) and currently is an associate editor of the J Orthop Sports Phys Ther.

Figure 1: Study flowchart

Figure 2: Boxplot on association between methodological quality (PEDro score) and statistical significance testing for baseline variables.

References

1. Ioannidis JPA, Fanelli D, Drake Dunne D, Goodman SN. Meta-research: Evaluation and Improvement of Research Methods and Practices. PLoS Biol 2015 13(10): e1002264. doi:10.1371/journal.pbio.1002264
2. Ioannidis JPA. Meta-research: why research on research matters. PLoS Biol 2018;16(3):e2005468. https//doi.org/10.1371/journal.pbio.2005468
3. Kamper SJ. Interpreting outcomes 1—change and difference: linking evidence to practice. J Orthop Sports Phys Ther 2019;49:357–8
4. ASA website: https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf. Last visited 21 September 2020
5. Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician. 2016;70(2):129-133
6. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond “p < 0.05”. The American Statistician. 2019;73(sup1):1-19
7. 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
8. Verhagen AP, Ostelo RWJG, Rademaker A. Is the p value really so significant? Australian Journal of Physiotherapy 2004;50:261-2.
9. Sullivan GM, Feinn R. Using Effect Size-or Why the P Value Is Not Enough. J Grad Med Educ. 2012;4(3):279-282
10. Cohen J. The earth is round (p<. 05). In: *What if there were no significance tests?*: Routledge; 2016:69-82.
11. Herbert R. Research Note: significance testing and hypothesis testing: meaningless, misleading and mostly unnecessary. J Physiother 2019;65:178-181.
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. Kallogjeri D, Spitznagel EL Jr, Piccirillo JF. Importance of Defining and Interpreting a Clinically Meaningful Difference in Clinical Research. JAMA Otolaryngol Head Neck Surg. 2020 Feb 1;146(2):101-102. doi: 10.1001/jamaoto.2019.3744. PMID: 31804662.
16. Armijo-Olivo S, Saltaji H, da Costa BR, Fuentes J, Ha C, Cummings GG. What is the influence of randomisation sequence generation and allocation concealment on treatment effects of physical therapy trials? A meta-epidemiological study. BMJ Open. 2015;5(9):e008562. doi: 10.1136/bmjopen-2015-008562.
17. Armijo-Olivo S, Fuentes J, da Costa BR, Saltaji H, Ha C, Cummings GG. Blinding in Physical Therapy Trials and Its Association with Treatment Effects: A Meta-epidemiological Study. Am J Phys Med Rehabil. 2017;96(1):34-44. doi: 10.1097/PHM.0000000000000521.

18. de Almeida MO, Saragiotto BT, Maher C, Costa LOP. Allocation Concealment and Intention-To-Treat Analysis Do Not Influence the Treatment Effects of Physical Therapy Interventions in Low Back Pain Trials: a Meta-epidemiologic Study. Arch Phys Med Rehabil. 2019;100(7):1359-1366. doi: 10.1016/j.apmr.2018.12.036.

19. 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

20. Thiese MS, Ronna B, Ott U. P value interpretations and considerations. Journal of thoracic disease. 2016;8(9):E928-e931

21. McCambridge AB, Nasser AM, Mehta P, Stubbs PW, Verhagen AP. METAPHoR: meta-research in physiotherapy trials – has reporting of physiotherapy interventions improved? JOSPT 2021, Accepted

22. Policy Statement: Description of Physical Therapy [press release]. World Confederation for Physical Therapy 2019

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. Cashin AG, McAuley JH. Clinimetrics: Physiotherapy Evidence Database (PEDro) Scale. J Physiother. 2020;66(1):59. doi: 10.1016/j.jphys.2019.08.005.

25. Moseley AM, Herbert RD, Maher CG, Sherrington C, Elkins MR. Reported quality of randomized controlled trials of physiotherapy interventions has improved over time. Journal of Clinical Epidemiology, 2011;64(6):594-601.

26. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-74.

27. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Family medicine. 2005;37:360-3.

28. Catillon M. Trends and predictors of biomedical research quality, 1990-2015: a meta-research study. BMJ Open. 2019;9(9):e030342. doi: 10.1136/bmjopen-2019-030342.

29. Núñez-Cortés R, Alvarez G, Pérez-Bracchiglione J, et al. Reporting results in manual therapy clinical trials: A need for improvement. Int J Osteopath Med. 2021. doi:10.1016/j.ijosm.2021.06.002

30. Riley SP, Swanson B, Brismée J-M, Sawyer SF. A systematic review of orthopaedic manual therapy randomized clinical trials quality. J Man Manip Ther. 2016;24(5):241-252. doi:10.1080/10669817.2015.1119372

31. Sedgwick P, Greenwood N. Understanding the Hawthorne effect. BMJ, 351 (2015), p. h4672, 10.1136/bmj.h4672

32. 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.

33. 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.

34. 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

35. Kamper SJ. Interpreting Outcomes 3-Clinical Meaningfulness: Linking Evidence to Practice. J Orthop Sports Phys Ther. 2019;49(9):677-678. doi: 10.2519/jospt.2019.0705.

36. 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.

37. Butcher NJ, Monsour A, Mew EJ, Szatmari P, Pierro A, Kelly LE, et al. Improving outcome reporting in clinical trial reports and protocols: study protocol for the Instrument for reporting Planned Endpoints in Clinical Trials (InsPECT). Trials 2019;20:161.

38. 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

39. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, Michie S, Moher D, Wager E. Reducing waste from incomplete or unusable reports of biomedical research. Lancet. 2014;383(9913):267-76. doi: 10.1016/S0140-6736(13)62228-X.

40. Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gülmezoglu AM, Howells DW, Ioannidis JP, Oliver S. How to increase value and reduce waste when research priorities are set. Lancet. 2014;383(9912):156-65. doi: 10.1016/S0140-6736(13)62229-1.

41. Lamb SE, Lall R, Hansen Z, et al. A multicentred randomised controlled trial of a primary care-based cognitive behavioural programme for low back pain. The Back Skills Training (BeST) trial. Health Technol Assess 2010;14:1–253.

42. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2 nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates Inc; 1988.

43. Froud R, Rajendran D, Patel S, Bright P, Bjørkli T, Eldridge S, Buchbinder R, Underwood M. The Power of Low Back Pain Trials: A Systematic Review of Power, Sample Size, and Reporting of Sample Size Calculations Over Time, in Trials Published Between 1980 and 2012. Spine (Phila Pa 1976). 2017;42(11):E680-E686. doi: 10.1097/BRS.0000000000001953.

Table 1: Characteristics of included studies published in the years 2000 and 2018.

	2000, n=39	**2018, n=101**	**Total, n=140**
Journals, *n (%)*			
Arch Phys Med Rehabil	11 (28.2%)	30 (29.6%)	41 (29.3%)
(A)J Physiother	2 (5.1%)	7 (6.9%)	9 (6.4%)
Clin Rehabil	5 (12.8%)	45 (44.6%)	50 (35.7%)
J Orthop Sports Phys Ther	4 (10.2%)	6 (5.9%)	10 (7.1%)
Phys Ther	6 15.4%)	6 (5.9%)	12 (8.6%)
Spine	11 (28.2%)	7 (6.9%)	18 (12.9%)
Subdiscipline, *n (%)*			
Musculoskeletal	26 (66.7%)	45 (44.6%)	71 (50.7%)
Neurological	7 (17.9%)	36 (35.6%)	43 (30.7%)
Cardiorespiratory	2 (5.1%)	9 (8.9%)	11 (7.9%)
Other	4 (10.2%)	11 (11%)	15 (10.7%)
***PEDro score* (0-10)**, mean (SD); (range)	5.8 (1.4); (3-8)	6.9 (1.3); (4-10)	6.6 (1.4); (3-10)
Sample size, mean (SD)	74.5 (88.3)	73.6 (49.1)	73.8 (62.2)
Use of p-value, *n (%)*			
Significance testing at baseline	13 (33.3%)	62 (61.4%)	75 (53.6%)
P-value for between-group analysis	36 (92.3%)	92 (91.1%)	128 (91.4%)
P-value for within-group analysis	19 (48.7%)	56 (55.4%)	75 (53.6%)
Effect estimates, *n (%)*			
Effect estimates for between-group analysis	12 (30.8%)	58 (57.4%)	70 (50%)
Effect estimates for within-group analysis	4 (10.6%)	29 (28.7%)	33 (23.6%)
Confidence intervals for between-group analysis	8 (20.5%)	55 (54.5%)	63 (45%)
Confidence intervals for within-group analysis	3 (7.7%)	28 (27.7%)	31 (22.1%)
Clinical relevance, *n (%)*			
Mentioned	10/39 (25.6%)	59/101 (58.4%)	69/140 (49.3%)
Used for sample size calculation	1/10	24/59	25/69
Specified a value for their outcome	3/10	23/59	26/69
Mentioned in discussion	9/10	49/59	58/69

(A)J Physiother = (Australian) Journal of Physiotherapy; Arch Phys Med Rehabil = Archives of Physical Medicine and Rehabilitation; Clin Rehabil = Clinical rehabilitation; J Orthop Sports Phys Ther = Journal of Orthopaedic and Sports Physical Therapy, Phys Ther = Physical Therapy

Table 2: Outcome data per journal

	Arch Phys Med Rehabil		***(A)J Physiother***		***Clin Rehabil***		***J Orthop Sports Phys Ther***		***Phys Ther***		***Spine***	
	2000	2018	2000	2018	2000	2018	2000	2018	2000	2018	2000	2018
N of studies	11	30	2	7	5	45	4	6	6	6	11	7
PEDro, mean (range)	5.6 (3-8)	6.7 (5-9)	6.5 (6-7)	8 (8-8)	5.6 (4-7)	7 (4-9)	5.5 (4-7)	6.8 (4-10)	5.3 (4-8)	6.7 (4-8)	6.3 (4-8)	6.3 (5-7)
Sample size, mean (range)	49.3 (10-135)	62.6 (19-180)	34 (28-40)	107.7 (46-198)	61.2 (27-98)	64.7 (19-181)	24.6 (10-52)	48.7 (24-103)	32.5 (18-44)	127.2 (52-208)	152.6 (21-457)	127.3 (23-304)
P-values												
Sign testing at baseline	3/11	18/30	1/2	0	2/5	33/45	1/4	2/6	1/6	3/6	5/11	6/7
Between-groups	10/11	29/30	2/2	4/7	5/5	44/45	4/4	4/6	6/6	6/6	9/11	7/7
Within-groups	3/11	18/30	0	1/7	3/5	26/45	3/4	3/6	4/6	3/6	4/11	4/7
Effect estimates												
Between-group	3/11	14/30	1/2	7/7	2/5	25/45	1/4	2/6	2/6	6/6	3/11	4/7
Within-group	1/11	5/30	0	2/7	1/5	17/45	1/4	1/6	1/6	3/6	0	1/7
Clinical relevance												
Mentioned	2/11	15/30	2/2	4/7	1/5	28/45	1/4	5/6	1/6	5/6	3/11	2/7
Related to outcome	0	5/15	1/2	2/4	0	10/28	0	2/5	1/6	3/5	1/3	1/2

(A)J Physiother = (Australian) Journal of Physiotherapy; Arch Phys Med Rehabil= Archives of Physical Medicine and Rehabilitation; Clin Rehabil = Clinical rehabilitation; J Orthop Sports Phys Ther = Journal of Orthopaedic and Sports Physical Therapy, Phys Ther = Physical Therapy; PEDro = Physiotherapy Evidence Database

Figure 1: Flow diagram of study selection

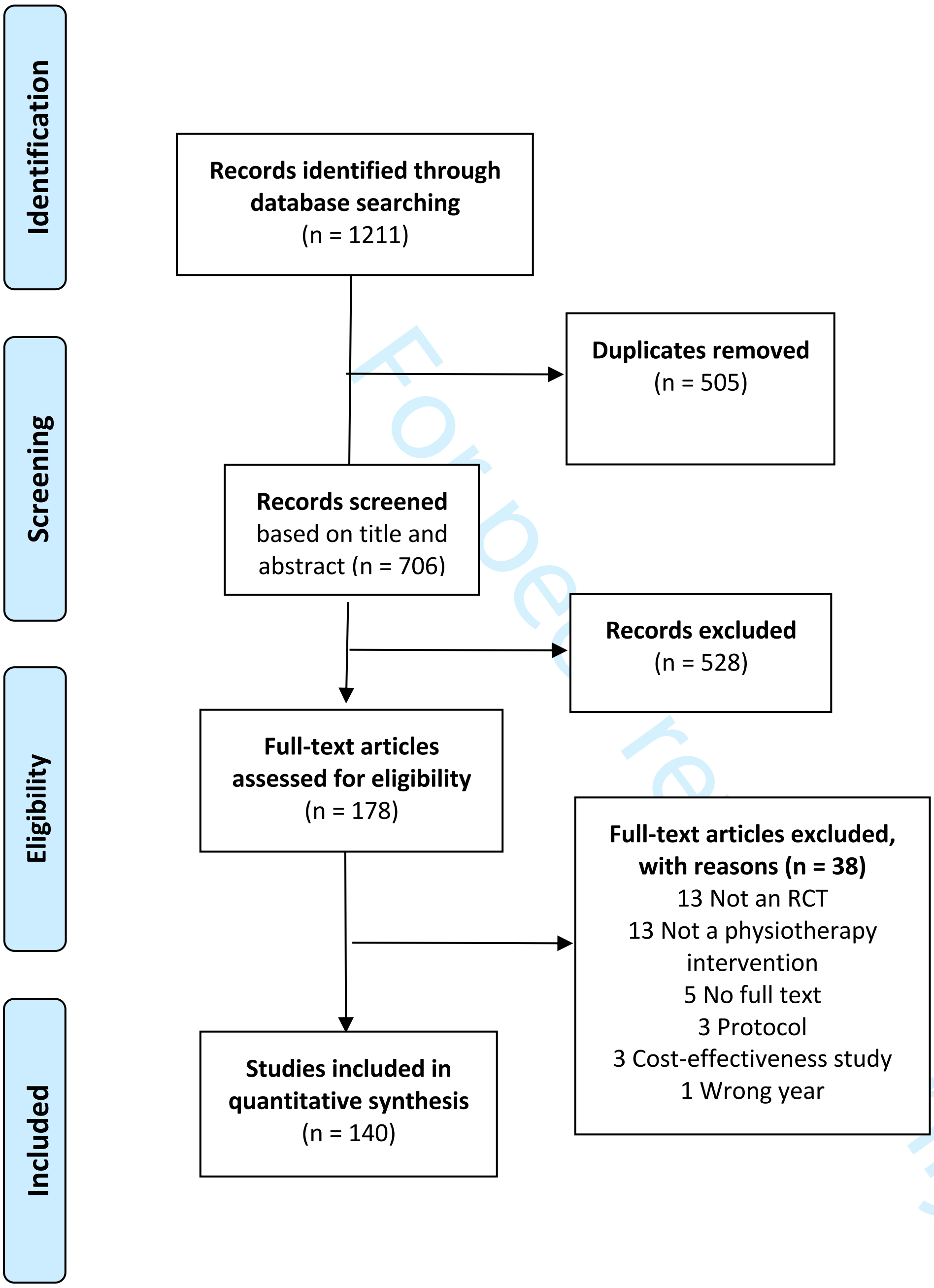
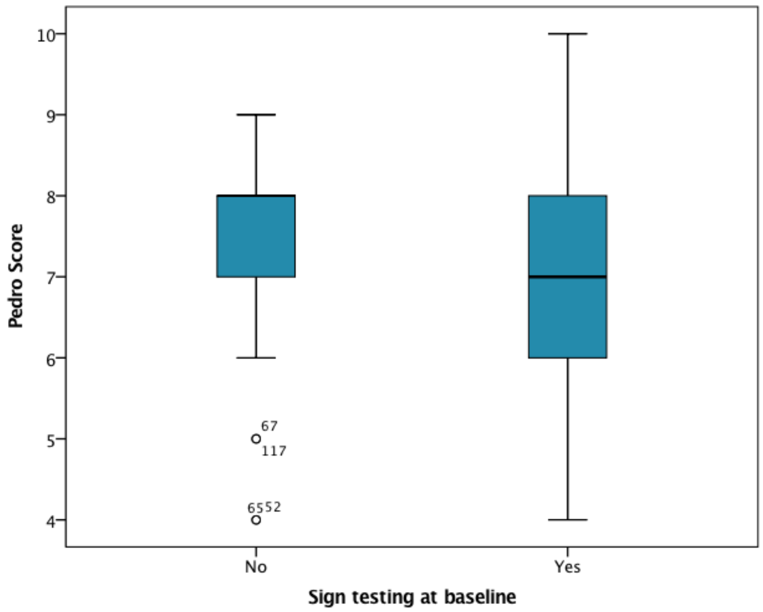

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 1211)
Duplicates removed (n = 505)
Records screened based on title and abstract (n = 706)
Records excluded (n = 528)
Full-text articles assessed for eligibility (n = 178)
Full-text articles excluded, with reasons (n = 38)
13 Not an RCT
13 Not a physiotherapy intervention
5 No full text
3 Protocol
3 Cost-effectiveness study
1 Wrong year
Studies included in quantitative synthesis (n = 140)

Figure 2: Boxplot on association between methodological quality (PEDro score) and statistical significance testing for baseline variables.



Median, 25% quartile and range

Supplemental material: Search strategy:

Basic search strategy, adapted for different databases if necessary.

(((“randomized controlled trial”[Publication Type]) OR (“controlled clinical trial”[Publication Type]) OR (randomized[Title/Abstract]) OR (placebo[Title/Abstract]) OR (clinical trials as topic[MeSH]) OR (randomly[Title/Abstract]) OR (trial[Title]) NOT ((animals[mh] NOT humans [mh])) AND ((Therapeutics[MeSH Terms]) OR (Therapeutics[Title/Abstract]) OR (“Musculoskeletal Manipulations”[MeSH Terms]) OR (“Musculoskeletal Manipulations”[Title/Abstract]) OR (“physical therapy modalities”[MeSH Terms]) OR (“physical therapy modalities”[Title/Abstract]) OR (“physical therapy specialty”[MeSH Terms]) OR (“physical therapy specialty”[Title/Abstract]) OR (rehabilitation[MeSH Terms]) OR (rehabilitation[Title/Abstract]) OR (“rehabilitation research”[MeSH Terms]) OR (“rehabilitation research”[Title/Abstract]) OR (“Manual therapy”[Title/Abstract]) OR (physiotherap*[Title/Abstract]) OR (“physical therap*”[Title/Abstract]) OR (exercis*[Title/Abstract]) OR (therap*[Title/Abstract]) OR (“physical activity”[Title/Abstract]) OR (education[Title/Abstract]) OR (electrotherap*[Title/Abstract]) OR (“Electrical stimulation therapy"[MeSH Terms]) OR (“Electrical stimulation therapy"[Title/Abstract]) OR (“motor control”[Title/Abstract]) OR (management[Title/Abstract]) OR (telehealth[Title/Abstract]) OR (telemedicine[MeSH Terms]) OR (“Respiratory therapy”[MeSH Terms]) OR (“Pain management”[MeSH Terms])) AND ((“1538-6724”[Journal]) OR (“0031-9023”[Journal]) OR (“1938-1344”[Journal]) OR (“0190-6011”[Journal]) OR (“1528-1159”[Journal]) OR (“0362- 2436”[Journal]) OR ("0004-9514"[Journal]) OR ("1836-9553"[Journal]) OR (“1532-821X”[Journal]) OR (“0003-9993”[Journal]) OR (“1477-0873”[Journal]) OR (“0269-2155”[Journal]) AND (("2000/01/01"[PDat]: "2000/12/31"[PDat]) OR ("2018/01/01"[PDat]: "2018/12/31"[PDat])))

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	1
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	5,6
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5,6
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	5,6
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5,6
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	5,6
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	5,6
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	6
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	6
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	6,7

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	6
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	6,7
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	7
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	7
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	8
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	8,9
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	8,9
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	8,9
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	8,9
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	9,10
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	11,12
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	10-12
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	1

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.

BMJ Open

Comparison between 2000 and 2018 on the reporting of statistical significance and clinical relevance in physiotherapy clinical trials in six major physiotherapy journals; a meta-research design

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-054875.R2
Article Type:	Original research
Date Submitted by the Author:	20-Nov-2021
Complete List of Authors:	Verhagen, Arianne; University of Technology Sydney, Discipline of Physiotherapy, Graduate School of Health Stubbs, Peter ; University of Technology Sydney, Graduate School of Health, Discipline of Physiotherapy Mehta, Poonam; University of Technology Sydney Kennedy, David; University of Technology Sydney Nasser, Anthony M; University of Technology Sydney Quel de Oliveira, Camila; University of Technology Sydney Pate, Joshua W; University of Technology Sydney Skinner, Ian W; University of Technology Sydney; Charles Sturt University McCambridge, Alana B; University of Technology Sydney
<b>Primary Subject Heading</b>:	Epidemiology
Secondary Subject Heading:	Evidence based practice, Research methods
Keywords:	EPIDEMIOLOGY, EDUCATION & TRAINING (see Medical Education & Training), PRIMARY CARE, Clinical trials < THERAPEUTICS, Rehabilitation medicine < INTERNAL MEDICINE

Comparison between 2000 and 2018 on the reporting of statistical significance and clinical relevance in physiotherapy clinical trials in six major physiotherapy journals; a meta-research design

Authors:

Arianne P Verhagen[1], Peter W Stubbs[1], Poonam Mehta[1], David Kennedy[1], Anthony M Nasser[1], Camila Quel de Oliveira[1], Joshua W Pate[1], Ian W Skinner[1,2], Alana B McCambridge[1]

Affiliation:

[1] Graduate School of Health, Discipline of Physiotherapy, University of Technology Sydney, Sydney, Australia

[2] School of Allied Health, Exercise and Sports Sciences, Charles Sturt University, Port Macquarie, Australia

Funding:

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Protocol: https://doi.org/10.26195/ebqp-8a65

Corresponding Author:

Prof Arianne P Verhagen, Head of Discipline of Physiotherapy, Graduate School of Health, University of Technology Sydney Australia

E-mail: Arianne.Verhagen@uts.edu.au

Word count:

Abstract: 262; text: 3917

Number of references: 43; tables: 2; figures: 2

ABSTRACT

Design: meta-research

Objective: To compare the prevalence of reporting *p*-values, effect estimates and clinical relevance in physiotherapy randomized controlled trials (RCTs) published in the years 2000 and 2018.

Methods: We performed a meta-research study of physiotherapy RCTs obtained from six major physiotherapy peer-reviewed journals that were published in the years 2000 and 2018. We searched the databases Embase, Medline, and PubMed in May 2019, and extracted data on the study characteristics and whether articles reported on statistical significance, effect estimates and confidence intervals for baseline, between-group, and within-group differences, and clinical relevance. Data were presented using descriptive statistics and inferences were made based on proportions. A 20% difference between 2000 and 2018 was regarded as a meaningful difference.

Results. We found 140 RCTs: 39 were published in 2000 and 101 in 2018. Overall, there was a high prevalence (>90%) of reporting *p*-values for the main (between-group) analysis, with no difference between years. Statistical significance testing was frequently used for evaluating baseline differences, increasing from 28% in 2000 to 61.4% in 2018. The prevalence of reporting effect estimates, confidence intervals and the mention of clinical relevance increased from 2000 to 2018 by 26.6%, 34% and 32.8% respectively. Despite an increase in use in 2018, over 40% of RCTs failed to report effect estimates, confidence intervals, and clinical relevance of results.

Conclusion. The prevalence of using *p*-values remains high in physiotherapy research. Although the proportion of reporting effect estimates, confidence intervals, and clinical relevance is higher in 2018 compared to 2000, many publications still fail to report and interpret study findings in this way.

Key words: Randomized clinical trials, Physiotherapy, reporting statistics, reporting clinical relevance

Strengths and Limitations

- This meta-research study will provide clear insight in the prevalence of (incorrect) use of p-values, and the prevalence of the use of effect estimates and clinical relevancy of outcomes
- We selected publications from six long-standing influential physiotherapy journals, assuming we select the best studies
- We defined a 20% difference as a meaningful difference
- We investigated reporting of *p*-values and effect estimates regardless of whether it was a primary or secondary outcome.

Introduction

As high-quality physiotherapy research needs to be clear, transparent, reproducible, and well written to inform clinical practice, it is important for clinicians to be confident in the methodological quality of physiotherapy research. Meta-research is a relatively new scientific discipline that explores how research is performed, reported, reproduced, evaluated, and incentivised [1,2]. As all scientific research is prone to bias, it is important that each profession critically evaluates its own research methods, standards of reporting, and validity of the outcomes [3].

Continuing discussions about the use (and misuse) of the *p*-value prompted the American Statistical Association (ASA) to recommend in 2016 that authors avoid statements on statistical significance and interpretation of outcomes using a p-value as an arbitrary threshold [4,5,6]. Traditionally, the *p*-value has been used in randomised clinical trials (RCTs) in conjunction with the null hypothesis testing to answer study questions related to the effectiveness of interventions by dichotomising results as significant or not significant [7]. Although valuable if interpreted correctly, null hypothesis testing has its limitations; it does **not** measure the probability of the truth of the null hypothesis, it does **not** measure the size or magnitude of an effect, and its replicability is poor [4,8-11]. The recommendation of the ASA is endorsed by many academic journals, nevertheless, authors continue to conclude whether an intervention is effective and should be used clinically by a dichotomous interpretation based on *p*-values.

Well conducted and large RCTs are considered high quality evidence and reporting of RCTs should be guided by the CONSORT-statement (Consolidated Standards of Reporting Trials) [12]. There are several recommendations in the CONSORT-statement regarding the reporting and appropriate use of *p*-values. For example, authors should not report results solely as *p*-values and are encouraged to (also) use effect estimates and 95% confidence intervals (95% CIs) [12]. The advantage of effect estimates is their ability to demonstrate the strength and the direction of the effect, and the 95% CIs provide a range of values between which the estimated true effect estimate lies [11,13,14]. Nevertheless, a dichotomized interpretation of the confidence interval (CI) should be discouraged; it allows for discussing the accuracy, precision and/or relevance of the effect estimate. Clinical relevance is another parameter used to interpret the magnitude of the effect, and to deem if a finding is clinically meaningful. Clinical relevance (or a clinically meaningful/worthwhile change, a minimum important difference (MID) or a minimal clinical important difference (MICD)) is regarded the threshold value for which any change (or larger) in for instance pain or disability is considered meaningful to patients [15].

According to the CONSORT-statement, authors should also compare baseline participant characteristics [12]. However, it discourages statistical significance testing of baseline covariates between randomized groups, as by using a proper randomization procedure all differences are based on chance. In addition, conclusions of a RCT should primarily be based on a between-group analysis by comparing post-intervention (and follow-up) outcomes between the groups or the between-group changes from baseline. Studies can additionally, with consideration, compare outcomes before and after the intervention using a 'within-group' analysis.

Previous meta-research within physiotherapy has investigated the use of randomization, blinding or intention-to-treat analysis [16-18] and one study evaluated the reporting of 95% CIs only [19]. To our knowledge, no study has examined the use of *p*-values, effect estimates or measures of clinical relevance in the physiotherapy literature before and after the CONSORT-statement was published in 2010. When selecting treatments, physiotherapists must be aware that statistical significance does not equate to clinical relevance [20]. Presenting effect estimates and precision of the effect (using 95% CIs) will also allow clinicians to consider how much a patient is likely to benefit from a given intervention compared to another (or no) intervention.

Therefore, the aim of this meta-research study was to investigate if the use of *p*-values, effect estimates, and clinical relevance differs between 2000 and 2018 in physiotherapy RCTs published in high quality influential journals (top 25%). Our secondary aim was to evaluate whether there is an association between the methodological quality of the studies and the incorrect use of *p*-values (i.e. baseline significance testing), and how clinical relevance was determined. This is because we assume that authors of studies with a higher methodological quality follow the reporting guidelines better.

Methods

Design

Meta-research study on the use of *p*-values, effect estimates (and 95% CI), and reporting and definition of clinical relevance in physiotherapy RCTs published in the years 2000 and 2018. The current study is part of a suite of research studies using the same sample of selected RCTs and was registered internally within the University of Technology Sydney, Discipline of Physiotherapy [21].

Ethics Approval

Not applicable as this involves a review of studies

Search strategy

We searched the databases Embase, Medline, and PubMed on the 24th of May 2019 (see appendix). The search strategy was developed to identify RCTs with at least one physiotherapy intervention arm published in six high-ranked physiotherapy journals, all supporting the CONSORT-statement, restricted to publication years 2000 or 2018. Journals included were: (Aus) Journal of Physiotherapy (J Physiother), Archives of Physical Medicine and Rehabilitation (Arch Phys Med Rehabil), Clinical Rehabilitation (Clin Rehabil), Journal of Orthopedic and Sports Physical Therapy (J Orthop Sports Phys Ther), Physical Therapy (Phys Ther) and Spine. These journals were chosen based on SCImago Journal Rank (all Q1 = top 25%) across both years, suggesting a substantial influence within the physiotherapy profession. The search strategy was reviewed by a librarian. All articles retrieved in the search were imported into Covidence and duplicates were removed.

Study selection

Two independent assessors first screened each article by title and abstract, and then by the full texts. If required, a third assessor resolved conflicts. Articles were eligible if they were an RCT that used at least one physiotherapy intervention. The World Confederation of Physiotherapy (WCPT) Policy statement was used to determine whether the intervention was within the international scope of physiotherapy [22]. Studies were excluded if they were conference proceedings, editorials, reviews, published protocols, cost effectiveness analyses or secondary analyses of RCTs only, not performed on humans, or the full text could not be obtained.

Data Extraction

Data extraction. The following information was extracted from each included study: descriptive information (such as subdiscipline of physiotherapy practice, study population, sample size at randomisation and analysis); use of *p*-values, effect estimates and 95% CIs reported for baseline, between- and within-group analysis; whether clinical relevance was mentioned (as well as synonyms, such as clinically important difference/change, minimal clinical differences, clinical significance, clinically worthwhile difference etc); and how clinical relevance was defined. Data was extracted from each article by two independent assessors with conflicts resolved by a third assessor.

Assessment of methodological quality. For all included studies, the methodological quality assessment was performed using the PEDro scale obtained from the PEDro-database (Physiotherapy Evidence Database) or independently assessed by two assessors, when the score was not available. Conflicts in scoring were resolved by a third assessor. PEDro scale is considered to have good interrater reliability and convergent validity [23,24].

Statistical Analysis

First, we calculated frequencies and proportions for reporting of *p*-values, effect estimates, 95% CIs and clinical relevance. A *priori,* we defined that a difference of ≥20% between 2000 and 2018 was regarded as a meaningful difference [25]. For our secondary aim we calculated the correlation (Pearson/Spearman correlation coefficient) between the PEDro score and a) the use of statistical significance testing at baseline and b) the mention of clinical relevance. We performed the analysis for the secondary aim in the trials of 2018 only as this dataset is the most recent representation of the literature. Correlation coefficients <0.20 were interpreted as no correlation, between 0.2 to 0.4 as low, 0.4 to 0.6 as moderate, 0.6 to 0.8 as high and above 0.8 as an almost perfect correlation [26,27]. Statistical analyses were performed using SPSS IBM 20.

Patient and Public involvement

No patients involved

Results

Search results

The search returned 1211 references, and after screening, 140 articles were included in the analysis (Figure 1). Of the 140 studies, 39 were published in 2000 and 101 in 2018 (Table 1).

Please insert figure 1 here

The number of published RCTs with at least one physiotherapy intervention was higher in 2018 compared to 2000 in Clin Rehabil, J Physiother, J Orthop Sports Phys Ther and Arch Phys Med Rehabil, while the number of published RCTs were similar in Spine and Phys Ther (Table 2). The RCTs were mainly performed in Europe/United Kingdom (n=51), USA/Canada (n=34), Australia/New Zealand (n=17) and Brazil (n=13).

Please insert table 1 here

Characteristics of included studies

Patient populations. Most studies were performed in musculoskeletal (50.7%) and neurological populations (30.7%) (Table 2). Other subdisciplines of physiotherapy were woman's health, oncology, and gerontology. The most common patient population in musculoskeletal studies were patients with low back pain (n=19) or neck pain (n=10). The most common patient populations in neurological studies were in stroke (n=22) and Parkinson's disease (n=7). Two journals (Spine and J Orthop Sports Phys Ther) published RCTs on musculoskeletal conditions only in both years, while the J Physiother did not publish any RCTs on musculoskeletal conditions in 2018.

Please insert table 2 here

Interventions. Of the 140 studies, most evaluated two interventions (n=115), while some evaluated three (n=21), or four or more interventions (n=4). Exercises or rehabilitation interventions (n=76; 54.2%) were the most common intervention evaluated followed by electrotherapy interventions (n=15, 10.7%). Most of the control interventions were exercise (n=32), followed by usual care (n=29), no treatment (n=26) or sham (n=16).

Sample size. The sample size in the studies ranged from 10 to 457 participants. The mean (standard deviation (SD)) sample size in all studies was 73.8 (62.2) at randomisation and 67.2 (58.6) in the analysis (Table 1). Between 2000 and 2018 the mean sample size across all journals was comparable, with a mean of 73-75 participants, but the difference between journals was large (Table 1).
In 2000 Spine published studies with an overall larger sample size (mean >125 participants) compared to the other journals (mean <65 participants). The sample size in the J Physiother and Phys They differed from 32 and 34 respectively in 2000, to over 100 participants, on average in 2018 (Table 2).

Methodological quality. Of the 140 articles, 15 (11%) had no PEDro-score and were rated by the researchers. Overall, the mean PEDro score was 6.6 (range from 3-10). The PEDro score differed slightly between 2000 and 2018, with a mean PEDro score of 5.8 in 2000 and 6.9 in 2018 (Table 1). The mean PEDro score in Spine did not differ between the years, while the PEDro score was higher in 2018, compared to 2000, in all other journals; with all included RCTs in the J Physiother in 2018 scoring 8/10 (Table 2).

Reporting prevalence

Most studies (n=128; 91.4%) used *p*-values to compare outcomes between groups (Table 1); one study (published in 2018) reported within-group differences only, nine studies reported only effect estimates and one study (published in 2000) did not report *p*-values or effect estimates. Complete reporting (presenting *p*-values, effect estimates and 95%CI on between group difference, and refraining from baseline sign testing), was observed in 5 studies (12.8%) in 2000 and 20 studies (19.8%) in 2018.

P-values.

The prevalence of *p*-values to determine between-group differences did not differ between 2000 and 2018 (92.3% and 91.1% respectively, Table 1). Of all studies that presented between-group *p*-values (n=130), 68 (52.3%) reported that the *p*-value was statistically significant, meaning <0.05, with a small difference between 2000 and 2018 (45.9% and 55.4% respectively). Of all studies reporting a non-significant difference regarding the primary outcome (n=62), 21 (33.3%) still reported positive findings in favour of the intervention, often based on the within-group differences or secondary outcomes. The number of studies that reported significance testing for baseline differences differed by 28.1%: 33.3% (95% CI: 19-50%) in 2000 and 61.4% (95% CI: 51-71%) in 2018.
The proportion of studies that reported (additional) within-group differences was 48.7% (95% CI: 32-65%) in 2000 and 55.4% (95% CI: 45-65%) in 2018 (Table 1). The J Physiother was the only journal where baseline statistical significance testing was not performed in 2018. The prevalence of *p*-values for between- and within-group differences decreased in J Physiother and J Orthop Sports Phys Ther by more than 20% (Table 2).

Effect estimates. Half of all studies (n=70, 50%) presented their results using an effect estimate (Table 1). The reporting of effect estimates for between-group analysis differed with 26.6% (30.8% (95% CI: 17-48%) in 2000 and 57.4% (95% CI: 47-67%) in 2018). The use of 95% CIs differed with 34% (20.5% (95% CI: 9-36%) in 2000 and 54.5% (95% CI: 44-64%) in 2018). Of the nine studies that reported only effect estimates (i.e., without *p*-values), seven were published in 2018. Overall, there was a meaningful difference (>20%) in the use of effect estimates (and 95% CIs) between 2000 and 2018, mainly due to the increases of >20% in Spine, J Physiother and Phys Ther journals.

Clinical relevance. Almost half of all studies (n=69; 49.3%) mentioned clinical relevance in their paper. In 25 studies, clinical relevance was related to the sample size calculation, but most of the studies mentioned clinical relevance (solely) in the discussion (Table 1). In 2018, only 23 studies (22.8%) defined clinically relevance and related it to the outcome. The overall mention of clinical relevance differed with 32.8% (25.6% (95% CI: 13-42%) in 2000 and 58.4% (95% CI: 48-68%) in 2018). Four journals showed a meaningful difference across years in mentioning clinical relevance (Table 2). The description of clinical relevance varied across studies, with 31 out of 69 (45%) studies clearly stating a minimal clinical important difference (MCID), mostly related to the sample size calculation, while others used the terms ‘clinical change’, ‘minimal change’, ‘clinical meaningful change’, ‘clinically relevant difference’, or ‘significant clinical change’ without specific reference to outcome data or cut-offs.

Methodological quality

The Pearson correlation coefficient between PEDro score and the use of statistical significance testing at baseline was -0.2 (Spearman: -0.23) in the studies in 2018 (see figure 2). We found a low correlation between methodological quality and incorrect significance testing (baseline differences). This means that studies with a higher methodological quality were slightly less likely to present statistical significance testing at baseline. The Pearson correlation coefficient between the PEDro score and the mention of clinical relevance was 0.13 (Spearman: 0.14) in the studies in 2018. This means that there was no correlation between methodological quality and mention of clinical relevance.

Please insert figure 2 here

Discussion

Main findings

Overall, we found that in the sample of physiotherapy journals investigated there was a high prevalence (>90%) of reporting *p*-values for the primary (between-group) analysis in both 2000 and 2018. Statistical significance testing for baseline differences differed between 28% in 2000 and 61.4% in 2018. Studies with higher methodological quality in 2018 tend to do slightly less statistical significance testing at baseline. Approximately half of all studies use statistical testing for within-group changes and there were no differences across years. The prevalence of reporting effect estimates, and the mention of clinical relevance differed >20% between 2000 and 2018, with it's reporting in almost 60% of all trials in 2018. However, many studies did not equate their study outcome to a known MCID. Although the CONSORT-statement has been endorsed by these six major physiotherapy journals, in this study, only two journals (J Physiother, Phys Ther) successfully adhered to the reporting guidelines for effect estimates in 2018.

Comparison with other studies

A previous study evaluating overall quality of methods in biomedical RCTs, including randomization, blinding and selective reporting, concluded that 59.3% of RCTs used inadequate methods (meaning scoring high risk of bias on one or more of the 6 Cochrane risk of bias items) and 35% of RCTs were poorly reported (meaning providing not enough information in the methods to decide on adequate or inadequate methods) [28]. Comparable findings have been found in physiotherapy RCTs in the PEDro database [23] and evaluation of manual therapy trials [29,30]. Whilst reporting of effect estimates in our selection of high-quality physiotherapy literature differs between 2000 and 2018, still most papers did not adhere to the reporting recommendations provided by the ASA and CONSORT-statements with regards to statistical significance testing and reliance on *p*-values to interpret results. Over a period of 18 years, presentation of effect estimates, and 95% CIs increased. Our results are consistent with another study that only evaluated the reporting of 95% CIs and found that these 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 [19]. However, in 2018, 42.6% of studies in our study still do not report the effect estimate, and solely present results using *p*-values. With an average increase of 2%, a one hundred percent compliance to the recommendations will only be achieved in 2049. Reporting of effect estimates (and CIs) are required if clinicians are to understand the magnitude and uncertainty of the treatment effect.

Although the reason for performing a RCT is to compare differences between randomised groups, about half of all studies also presented the results of within-group analyses. Often participants in RCTs improve over time due to e.g. natural recovery or to the Hawthorne effect [31]. Therefore, it remains unclear why so many authors choose to test within-group differences in an RCT, and why journal editors permit authors to do so when it is conceivable that a reader may misinterpret the result.

The CONSORT-statement also recommends comparing baseline differences between groups, however statistical testing for baseline differences between randomized groups is not recommended [12,32]. The rationale is that when the randomization procedure is performed well, all differences at baseline are due to chance. Hypothesis testing at baseline means that we test the probability of a difference by chance, when we know these differences occur by chance and are therefore considered inappropriate and illogical [32,33]. We found that statistical significance testing for baseline differences had increased from 2000 to 2018, with over 60% of studies reporting *p*-values for baseline comparisons. Our results are higher than those in a previous study published in 2010 which found 38% of RCTs reported *p*-values for baseline differences in 114 RCTs published in leading medical journals [32]. A reason for this difference might be that the selection of the 114 RCTs came from four leading medical journals with higher impact factors than our six journals, and assuming their risk of bias was lower (though not assessed in that article) than in our sample. Another reason might be that statistical testing of baseline data in clinical trials is common practice and authors might just replicate the analysis of other authors [33,34]. In addition, reviewers (and maybe even editors) may suggest authors to present statistical baseline testing for this reason.

The prevalence of significance testing for baseline differences and within-group changes is concerning, as it shows that authors do not completely understand the reason for randomisation in RCTs.

Clinical relevance of outcomes is important when interpreting if the effects of an intervention are meaningful to patients [35]. Although the mention of clinical relevance increased over time, in 2018 only a small proportion of studies (n=23, 22.8%) related clinically relevance to their outcome, and most studies it was mentioned it in the discussion section only. Also, a wide variety of terminology was used, and the terms ‘change’ and ‘difference’ were used interchangeably in most studies. Recently, experts clarified the difference between these concepts more clearly [36]. They state that MCID are cross-sectional between-group differences, such as the difference between two intervention groups after treatment that are regarded clinically relevant, while minimal important changes (MIC) are longitudinal within-person changes in scores [36]. The lack of known clinically important values, particularly MCID for use in RCTs may be a barrier for researchers to report and interpret their findings in relation to clinical relevance. Future research that aims to determine MCIDs for core outcomes measures are warranted.

Strengths & limitations

There are several limitations to our study. Firstly, the scope of physiotherapy practice is broad and may vary between countries. It is therefore possible that we may have missed some relevant publications or included publications that in other countries would not be defined as providing ‘physiotherapy’ intervention. As we have used the WCPT definitions as selection criteria we assume this will not potentially bias our results. Second, we selected publications from six long-standing influential physiotherapy journals. We assumed that these journals would publish the best RCTs, meaning that our findings might be more positive (meaning a higher percentage of improvement in 2018) than if a sample was taken from the overall physiotherapy literature. Third, as the included RCTs from the six journals predominantly investigated musculoskeletal interventions, we cannot assume that our findings are representative of all physiotherapy research and subspecialties. Fourth, we defined a 20% difference as a meaningful difference based on a previous study [25]. Unfortunately, we did not define what percentage of the literature should ideally report effect estimates or mention clinical relevance. In retrospect, that was pertinent to define. Fifth, as the number of published RCTs in 2018 was over twice as much as in 2000, this imbalance might have influenced our results, as results from a smaller number of studies are often a bit less precise. Lastly, we investigated reporting of *p*-values and effect estimates regardless of whether it was a primary or secondary outcome. However, we do not expect that our findings would differ majorly when only measured for the primary outcome.

Future Directions

Research is one of the pillars of evidence-based practice and plays a fundamental role in guiding treatment selection. 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 [36]. Unfortunately, the methodological quality of the RCTs in the PEDro database remains suboptimal [23]. Our findings confirm that the statistical reporting and use of clinical relevance in physiotherapy RCTs is also suboptimal. To further help authors, a consensus-based reporting checklist for primary outcomes in RCTs is currently under development: InsPECT statement, specifically focussing on reporting of outcomes in a transparent way [37].

Researchers have an ethical obligation to accurately report findings to allow for evidence-based decision-making [8,38]. By 2018, authors should have been aware of reporting guidelines such as the CONSORT-statement and been obligated to adhere to publication guidelines [38]. The findings of our study show that there are some improvements in the physiotherapy literature, but there is still need for improvement concerning statistical reporting and reporting of clinical relevance. Overall, stronger incentives (or penalties) may be required to improve the quality and reporting of physiotherapy research.

Performing underpowered studies is regarded as research waste [39,40]. The typical standardized effect estimate in physiotherapy trials is around 0.3 [41]. This is considered a small to medium effect estimate [42]. The sample size that on average should be sufficient to detect an effect estimate of 0.3 (in low back pain RCTs) is about 175 participants [43]. Almost all studies in our analysis had sample sizes that were too small to detect an effect estimate of 0.3. Nevertheless, about half the studies that presented between group *p*-values, reported statistical significance (using *p*<0.05). The mean sample size did not increase over time, although there was some variation between journals. This finding is a concern because sample sizes of physiotherapy RCTs remain small and therefore are likely underpowered [44]. We strongly recommend future studies to be of sufficient power.

Conclusion

The prevalence of the reporting of *p*-values remains high in physiotherapy research published in high ranked physiotherapy journals and the reporting of statistical significance testing for baseline differences was higher in 2018 compared to 2000. The prevalence of the reporting of effect estimates (and CI's) was >20% higher in 2018 compared to 2000 but was still reported in less than 60% of all publications. Our findings suggest that although reporting seems to have improved, there is still under-reporting of effect estimates.

Acknowledgements:

Students of Master of Physiotherapy program of UTS assisted in searching, data extraction, and assessment of methodological quality: S. Rogan, D. Commerford, G. Milgate, K. Cummins, M. Beech, R. Briody, D. Hagtharp, L. Jovic, J. Lenn, and A Shah was the librarian that assisted with the search.

Author statement

Arianne P Verhagen: Conceptualization; Data curation; Formal analysis; Methodology; Supervision; Validation; Roles/Writing - original draft; Writing - review & editing. **Peter W Stubbs**: Data curation; Formal analysis; Validation; Roles/Writing - original draft; Writing - review & editing **Poonam Mehta**: Data curation; Supervision; Validation; Roles/Writing - original draft; Writing - review & editing. **David Kennedy**: Conceptualization; Roles/Writing - original draft; Writing - review & editing. **Anthony M Nasser**: Supervision; Roles/Writing - original draft; Writing - review & editing. **Camila Quel de Oliveira**: Roles/Writing - original draft; Writing - review & editing. **Joshua W Pate**: Roles/Writing - original draft; Writing - review & editing. **Ian W Skinner**: Data curation; Supervision; Roles/Writing - original draft; Writing - review & editing. **Alana B McCambridge**: Conceptualization; Data curation; Formal analysis; Methodology; Project administration; Resources; Software; Supervision; Validation; Roles/Writing - original draft; Writing - review & editing

Competing interests: None declared

Data availability statement: "No additional data available".

Conflict of interest: AV was a member of the editorial board of the J Physiother (until 2020) and currently is an associate editor of the J Orthop Sports Phys Ther.

Figure 1: Study flowchart

Figure 2: Boxplot on association between methodological quality (PEDro score) and statistical significance testing for baseline variables.

References

1. Ioannidis JPA, Fanelli D, Drake Dunne D, Goodman SN. Meta-research: Evaluation and Improvement of Research Methods and Practices. PLoS Biol 2015 13(10): e1002264. doi:10.1371/journal.pbio.1002264
2. Ioannidis JPA. Meta-research: why research on research matters. PLoS Biol 2018;16(3):e2005468. https//doi.org/10.1371/journal.pbio.2005468
3. Kamper SJ. Interpreting outcomes 1—change and difference: linking evidence to practice. J Orthop Sports Phys Ther 2019;49:357–8
4. ASA website: https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf. Last visited 21 September 2020
5. Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician. 2016;70(2):129-133
6. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p < 0.05". The American Statistician. 2019;73(sup1):1-19
7. 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
8. Verhagen AP, Ostelo RWJG, Rademaker A. Is the p value really so significant? Australian Journal of Physiotherapy 2004;50:261-2.
9. Sullivan GM, Feinn R. Using Effect Size-or Why the P Value Is Not Enough. J Grad Med Educ. 2012;4(3):279-282
10. Cohen J. The earth is round (p<. 05). In: *What if there were no significance tests?*: Routledge; 2016:69-82.
11. Herbert R. Research Note: significance testing and hypothesis testing: meaningless, misleading and mostly unnecessary. J Physiother 2019;65:178-181.
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. Kallogjeri D, Spitznagel EL Jr, Piccirillo JF. Importance of Defining and Interpreting a Clinically Meaningful Difference in Clinical Research. JAMA Otolaryngol Head Neck Surg. 2020 Feb 1;146(2):101-102. doi: 10.1001/jamaoto.2019.3744.
16. Armijo-Olivo S, Saltaji H, da Costa BR, Fuentes J, Ha C, Cummings GG. What is the influence of randomisation sequence generation and allocation concealment on treatment effects of physical therapy trials? A meta-epidemiological study. BMJ Open. 2015;5(9):e008562. doi: 10.1136/bmjopen-2015-008562.
17. Armijo-Olivo S, Fuentes J, da Costa BR, Saltaji H, Ha C, Cummings GG. Blinding in Physical Therapy Trials and Its Association with Treatment Effects: A Meta-epidemiological Study. Am J Phys Med Rehabil. 2017;96(1):34-44. doi: 10.1097/PHM.0000000000000521.

18. de Almeida MO, Saragiotto BT, Maher C, Costa LOP. Allocation Concealment and Intention-To-Treat Analysis Do Not Influence the Treatment Effects of Physical Therapy Interventions in Low Back Pain Trials: a Meta-epidemiologic Study. Arch Phys Med Rehabil. 2019;100(7):1359-1366. doi: 10.1016/j.apmr.2018.12.036.

19. 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

20. Thiese MS, Ronna B, Ott U. P value interpretations and considerations. Journal of thoracic disease. 2016;8(9):E928-e931

21. McCambridge AB, Nasser AM, Mehta P, Stubbs PW, Verhagen AP. Has Reporting on Physical Therapy Interventions Improved in 2 Decades? An Analysis of 140 Trials Reporting on 225 Interventions. J Orthop Sports Phys Ther 2021;51(10):503-9.

22. Policy Statement: Description of Physical Therapy [press release]. World Confederation for Physical Therapy 2019

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. Cashin AG, McAuley JH. Clinimetrics: Physiotherapy Evidence Database (PEDro) Scale. J Physiother. 2020;66(1):59. doi: 10.1016/j.jphys.2019.08.005.

25. Moseley AM, Herbert RD, Maher CG, Sherrington C, Elkins MR. Reported quality of randomized controlled trials of physiotherapy interventions has improved over time. Journal of Clinical Epidemiology, 2011;64(6):594-601.

26. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-74.

27. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Family medicine. 2005;37:360-3.

28. Catillon M. Trends and predictors of biomedical research quality, 1990-2015: a meta-research study. BMJ Open. 2019;9(9):e030342. doi: 10.1136/bmjopen-2019-030342.

29. Núñez-Cortés R, Alvarez G, Pérez-Bracchiglione J, et al. Reporting results in manual therapy clinical trials: A need for improvement. Int J Osteopath Med. 2021. doi:10.1016/j.ijosm.2021.06.002

30. Riley SP, Swanson B, Brismée J-M, Sawyer SF. A systematic review of orthopaedic manual therapy randomized clinical trials quality. J Man Manip Ther. 2016;24(5):241-252. doi:10.1080/10669817.2015.1119372

31. Sedgwick P, Greenwood N. Understanding the Hawthorne effect. BMJ, 351 (2015), p. h4672, 10.1136/bmj.h4672

32. 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.

33. 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.

34. de Boer, M.R., Waterlander, W.E., Kuijper, L.D. *et al* Testing for baseline differences in randomized controlled trials: an unhealthy research behavior that is hard to eradicate. *Int J Behav Nutr Phys Act* 2015;12,4. doi.org/10.1186/s12966-015-0162-z

35. 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

36. Kamper SJ. Interpreting Outcomes 3-Clinical Meaningfulness: Linking Evidence to Practice. J Orthop Sports Phys Ther. 2019;49(9):677-678. doi: 10.2519/jospt.2019.0705.

37. 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.

38. Butcher NJ, Monsour A, Mew EJ, Szatmari P, Pierro A, Kelly LE, et al. Improving outcome reporting in clinical trial reports and protocols: study protocol for the Instrument for reporting Planned Endpoints in Clinical Trials (InsPECT). Trials 2019;20:161.

39. 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

40. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, Michie S, Moher D, Wager E. Reducing waste from incomplete or unusable reports of biomedical research. Lancet. 2014;383(9913):267-76. doi: 10.1016/S0140-6736(13)62228-X.

41. Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gülmezoglu AM, Howells DW, Ioannidis JP, Oliver S. How to increase value and reduce waste when research priorities are set. Lancet. 2014;383(9912):156-65. doi: 10.1016/S0140-6736(13)62229-1.

42. Lamb SE, Lall R, Hansen Z, et al. A multicentred randomised controlled trial of a primary care-based cognitive behavioural programme for low back pain. The Back Skills Training (BeST) trial. Health Technol Assess 2010;14:1–253.

43. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2 nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates Inc; 1988.

44. Froud R, Rajendran D, Patel S, Bright P, Bjørkli T, Eldridge S, Buchbinder R, Underwood M. The Power of Low Back Pain Trials: A Systematic Review of Power, Sample Size, and Reporting of Sample Size Calculations Over Time, in Trials Published Between 1980 and 2012. Spine (Phila Pa 1976). 2017;42(11):E680-E686. doi: 10.1097/BRS.0000000000001953.

Table 1: Characteristics of included studies published in the years 2000 and 2018.

	2000, n=39	**2018, n=101**	**Total, n=140**
Journals, *n (%)*			
Arch Phys Med Rehabil	11 (28.2%)	30 (29.6%)	41 (29.3%)
(A)J Physiother	2 (5.1%)	7 (6.9%)	9 (6.4%)
Clin Rehabil	5 (12.8%)	45 (44.6%)	50 (35.7%)
J Orthop Sports Phys Ther	4 (10.2%)	6 (5.9%)	10 (7.1%)
Phys Ther	6 15.4%)	6 (5.9%)	12 (8.6%)
Spine	11 (28.2%)	7 (6.9%)	18 (12.9%)
Subdiscipline, *n (%)*			
Musculoskeletal	26 (66.7%)	45 (44.6%)	71 (50.7%)
Neurological	7 (17.9%)	36 (35.6%)	43 (30.7%)
Cardiorespiratory	2 (5.1%)	9 (8.9%)	11 (7.9%)
Other	4 (10.2%)	11 (11%)	15 (10.7%)
***PEDro score* (0-10)**, mean (SD); (range)	5.8 (1.4); (3-8)	6.9 (1.3); (4-10)	6.6 (1.4); (3-10)
Sample size, mean (SD)	74.5 (88.3)	73.6 (49.1)	73.8 (62.2)
Use of p-value, *n (%)*			
Significance testing at baseline	13 (33.3%)	62 (61.4%)	75 (53.6%)
P-value for between-group analysis	36 (92.3%)	92 (91.1%)	128 (91.4%)
P-value for within-group analysis	19 (48.7%)	56 (55.4%)	75 (53.6%)
Effect estimates, *n (%)*			
Effect estimates for between-group analysis	12 (30.8%)	58 (57.4%)	70 (50%)
Effect estimates for within-group analysis	4 (10.6%)	29 (28.7%)	33 (23.6%)
Confidence intervals for between-group analysis	8 (20.5%)	55 (54.5%)	63 (45%)
Confidence intervals for within-group analysis	3 (7.7%)	28 (27.7%)	31 (22.1%)
Clinical relevance, *n (%)*			
Mentioned	10/39 (25.6%)	59/101 (58.4%)	69/140 (49.3%)
Used for sample size calculation	1/10	24/59	25/69
Specified a value for their outcome	3/10	23/59	26/69
Mentioned in discussion	9/10	49/59	58/69

(A)J Physiother = (Australian) Journal of Physiotherapy; Arch Phys Med Rehabil = Archives of Physical Medicine and Rehabilitation; Clin Rehabil = Clinical rehabilitation; J Orthop Sports Phys Ther = Journal of Orthopaedic and Sports Physical Therapy, Phys Ther = Physical Therapy

Table 2: Outcome data per journal

	Arch Phys Med Rehabil		***(A)J Physiother***		***Clin Rehabil***		***J Orthop Sports Phys Ther***		***Phys Ther***		***Spine***	
	2000	2018	2000	2018	2000	2018	2000	2018	2000	2018	2000	2018
N of studies	11	30	2	7	5	45	4	6	6	6	11	7
PEDro, mean (range)	5.6 (3-8)	6.7 (5-9)	6.5 (6-7)	8 (8-8)	5.6 (4-7)	7 (4-9)	5.5 (4-7)	6.8 (4-10)	5.3 (4-8)	6.7 (4-8)	6.3 (4-8)	6.3 (5-7)
Sample size, mean (range)	49.3 (10-135)	62.6 (19-180)	34 (28-40)	107.7 (46-198)	61.2 (27-98)	64.7 (19-181)	24.6 (10-52)	48.7 (24-103)	32.5 (18-44)	127.2 (52-208)	152.6 (21-457)	127.3 (23-304)
P-values												
Sign testing at baseline	3/11	18/30	1/2	0	2/5	33/45	1/4	2/6	1/6	3/6	5/11	6/7
Between-groups	10/11	29/30	2/2	4/7	5/5	44/45	4/4	4/6	6/6	6/6	9/11	7/7
Within-groups	3/11	18/30	0	1/7	3/5	26/45	3/4	3/6	4/6	3/6	4/11	4/7
Effect estimates												
Between-group	3/11	14/30	1/2	7/7	2/5	25/45	1/4	2/6	2/6	6/6	3/11	4/7
Within-group	1/11	5/30	0	2/7	1/5	17/45	1/4	1/6	1/6	3/6	0	1/7
Clinical relevance												
Mentioned	2/11	15/30	2/2	4/7	1/5	28/45	1/4	5/6	1/6	5/6	3/11	2/7
Related to outcome	0	5/15	1/2	2/4	0	10/28	0	2/5	1/6	3/5	1/3	1/2

(A)J Physiother = (Australian) Journal of Physiotherapy; Arch Phys Med Rehabil= Archives of Physical Medicine and Rehabilitation; Clin Rehabil = Clinical rehabilitation; J Orthop Sports Phys Ther = Journal of Orthopaedic and Sports Physical Therapy, Phys Ther = Physical Therapy; PEDro = Physiotherapy Evidence Database

Figure 1: Flow diagram of study selection

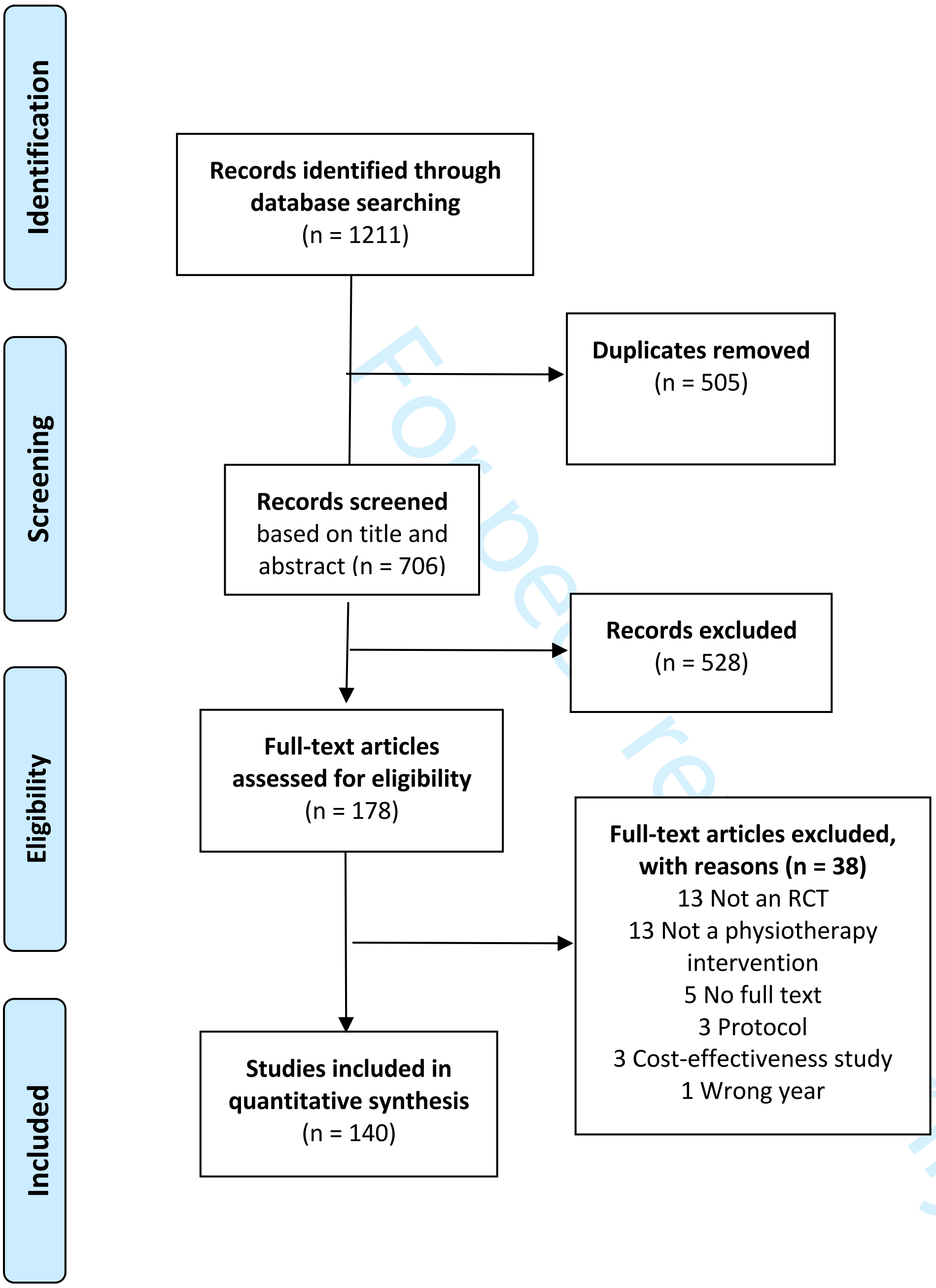
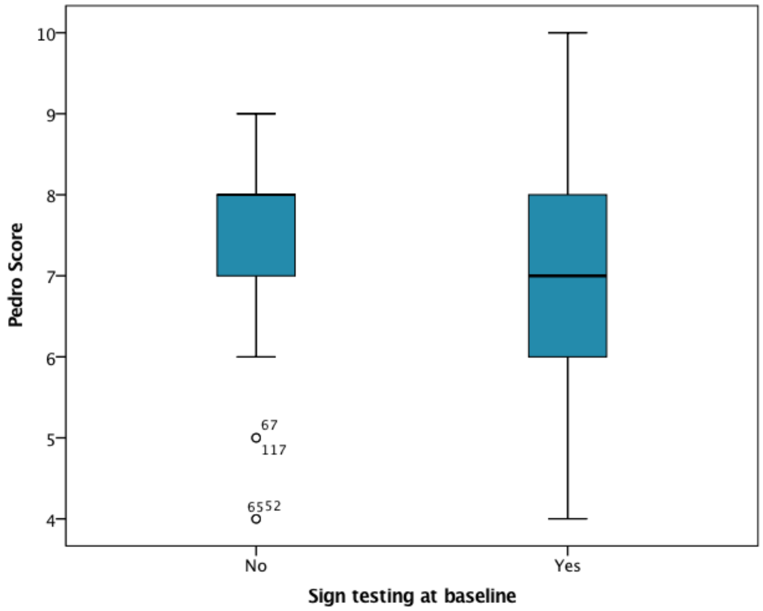

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 1211)
Duplicates removed (n = 505)
Records screened based on title and abstract (n = 706)
Records excluded (n = 528)
Full-text articles assessed for eligibility (n = 178)
Full-text articles excluded, with reasons (n = 38)
13 Not an RCT
13 Not a physiotherapy intervention
5 No full text
3 Protocol
3 Cost-effectiveness study
1 Wrong year
Studies included in quantitative synthesis (n = 140)

Figure 2: Boxplot on association between methodological quality (PEDro score) and statistical significance testing for baseline variables.



Median, 25% quartile and range

Supplemental material: Search strategy:

Basic search strategy, adapted for different databases if necessary.

(((“randomized controlled trial”[Publication Type]) OR (“controlled clinical trial”[Publication Type]) OR (randomized[Title/Abstract]) OR (placebo[Title/Abstract]) OR (clinical trials as topic[MeSH]) OR (randomly[Title/Abstract]) OR (trial[Title]) NOT ((animals[mh] NOT humans [mh])) AND ((Therapeutics[MeSH Terms]) OR (Therapeutics[Title/Abstract]) OR (“Musculoskeletal Manipulations”[MeSH Terms]) OR (“Musculoskeletal Manipulations”[Title/Abstract]) OR (“physical therapy modalities”[MeSH Terms]) OR (“physical therapy modalities”[Title/Abstract]) OR (“physical therapy specialty”[MeSH Terms]) OR (“physical therapy specialty”[Title/Abstract]) OR (rehabilitation[MeSH Terms]) OR (rehabilitation[Title/Abstract]) OR (“rehabilitation research”[MeSH Terms]) OR (“rehabilitation research”[Title/Abstract]) OR (“Manual therapy”[Title/Abstract]) OR (physiotherap*[Title/Abstract]) OR (“physical therap*”[Title/Abstract]) OR (exercis*[Title/Abstract]) OR (therap*[Title/Abstract]) OR (“physical activity”[Title/Abstract]) OR (education[Title/Abstract]) OR (electrotherap*[Title/Abstract]) OR (“Electrical stimulation therapy"[MeSH Terms]) OR (“Electrical stimulation therapy"[Title/Abstract]) OR (“motor control”[Title/Abstract]) OR (management[Title/Abstract]) OR (telehealth[Title/Abstract]) OR (telemedicine[MeSH Terms]) OR (“Respiratory therapy”[MeSH Terms]) OR (“Pain management”[MeSH Terms])) AND ((“1538-6724”[Journal]) OR (“0031-9023”[Journal]) OR (“1938-1344”[Journal]) OR (“0190-6011”[Journal]) OR (“1528-1159”[Journal]) OR (“0362- 2436”[Journal]) OR ("0004-9514"[Journal]) OR ("1836-9553"[Journal]) OR (“1532-821X”[Journal]) OR (“0003-9993”[Journal]) OR (“1477-0873”[Journal]) OR (“0269-2155”[Journal]) AND (("2000/01/01"[PDat]: "2000/12/31"[PDat]) OR ("2018/01/01"[PDat]: "2018/12/31"[PDat])))

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	1
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	5,6
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5,6
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	5,6
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5,6
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	5,6
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	5,6
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	6
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	6
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., $I^{2)}$ for each meta-analysis.	6,7

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	6
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	6,7
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	7
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	7
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	8
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	8,9
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	8,9
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	8,9
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	8,9
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	9,10
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	11,12
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	10-12
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	1

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.