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

Arianne Verhagen ,[1] Peter William 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]

To cite: Verhagen A, Stubbs PW, Mehta P, *et al*. 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. *BMJ Open* 2022;**12**:e054875. doi:10.1136/bmjopen-2021-054875

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-054875).

Received 29 June 2021
Accepted 10 December 2021

[1]Discipline of Physiotherapy, Graduate School of Health, University of Technology Sydney, Sydney, New South Wales, Australia
[2]School of Allied Health, Department Exercise and Sports Sciences, Charles Sturt University, Port Macquarie, New South Wales, Australia

Correspondence to
Professor Arianne Verhagen; arianne.verhagen@uts.edu.au

ABSTRACT

Design Meta-research.

Objective To compare the prevalence of reporting p values, effect estimates and clinical relevance in physiotherapy randomised 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, CIs 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, CIs and clinical relevance of results.

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

Strengths and limitations of this study

- ► 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.[4–6] Traditionally, the p value has been used in randomised controlled 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% 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 dichotomised interpretation of the 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 or a minimal clinical important difference (MCID)) 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 randomised groups, as by using a proper randomisation procedure all differences are based on chance. In addition, conclusions of an 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 randomisation, 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 with 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 (ie, 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]

Search strategy

We searched the databases Embase, Medline and PubMed on the 24 May 2019 (see online supplemental 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-group 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 inter-rater 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 0.2–0.4 as low, 0.4–0.6 as moderate, 0.6–.8 as high and above 0.8 as an almost perfect correlation.[26 27] Statistical analyses were performed using SPSS IBM V.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 with 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/UK (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

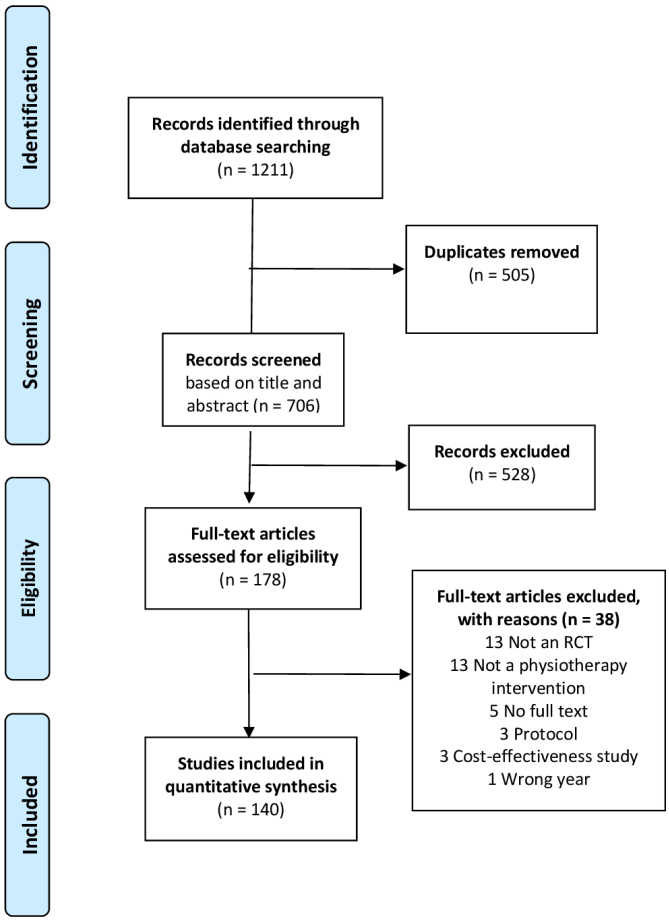

Figure 1 Study flow chart. RCT, randomised controlled trial.

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.

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 (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

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.

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 with 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 to 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 with 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%

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; PEDro, Physiotherapy Evidence Database; Phys Ther, Physical Therapy.

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% to 50%) in 2000 and 61.4% (95% CI 51% to 71%) in 2018.

The proportion of studies that reported (additional) within-group differences was 48.7% (95% CI 32% to 65%) in 2000 and 55.4% (95% CI 45% to 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-group 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% to 48%) in 2000 and 57.4% (95% CI 47% to 67%) in 2018). The use of 95% CIs differed with 34% (20.5% (95% CI 9% to 36%) in 2000 and 54.5% (95% CI 44% to 64%) in 2018). Of the nine studies that reported only effect estimates (ie, 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% to 42%) in 2000 and 58.4% (95% CI 48% to 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 an 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.

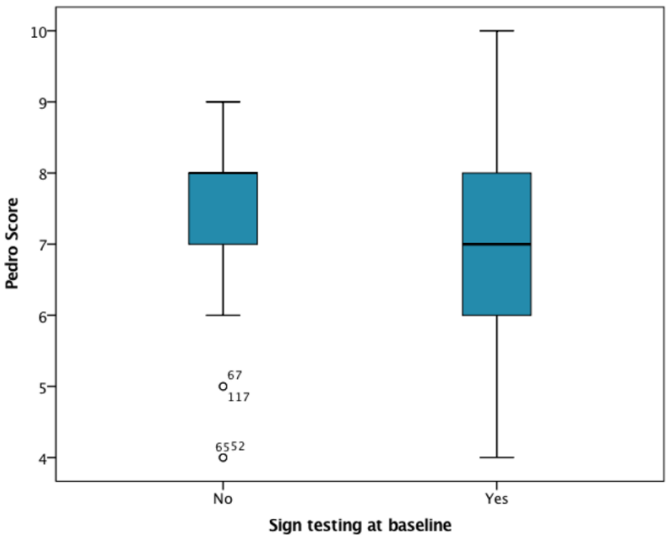

Figure 2 Boxplot on association between methodological quality (PEDro score) and statistical significance testing for baseline variables. PEDro, Physiotherapy Evidence Database.

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.

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 its 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 randomisation, 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 six 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] While 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 per cent 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, for example, 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 randomised groups is not recommended.[12 32] The rationale is that when the randomisation 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 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 and limitations

There are several limitations to our study. First, 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 (Instrument for reporting Planned Endpoints in Clinical Trials) 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 standardised 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 with 2000. The prevalence of the reporting of effect estimates (and CIs) was >20% higher in 2018 compared with 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.

Twitter Peter William Stubbs @PeterStubbsPT

Acknowledgements Students of Master of Physiotherapy programme 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.

Contributors AV: conceptualisation; data curation; formal analysis; methodology; supervision; validation; roles/writing-original draft; writing-review, editing and guarantor. PWS: data curation; formal analysis; validation; roles/writing-original draft; writing-review and editing. PM: data curation; supervision; validation; roles/writing-original draft; writing-review and editing. DK: conceptualisation; roles/writing-original draft; writing-review and editing. AMN: supervision; roles/writing-original draft; writing-review and editing. CQdO: roles/writing-original draft; writing-review and editing. JWP: roles/writing-original draft; writing-review and editing. IWS: data curation; supervision; roles/writing-original draft; writing-review and editing. ABM: conceptualisation; data curation; formal analysis; methodology; project administration; resources; software; supervision; validation; roles/writing-original draft; writing-review and editing.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests 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*.

Patient consent for publication Not applicable.

Ethics approval Not applicable as this involves a review of studies.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Data will be available on request with the first and/or last author. It is data on published randomised controlled trials.

ORCID iD

Arianne Verhagen http://orcid.org/0000-0002-6195-0128

REFERENCES

1 Ioannidis JPA, Fanelli D, Dunne DD, *et al*. Meta-research: evaluation and improvement of research methods and practices. *PLoS Biol* 2015;13:e1002264.
2 Ioannidis JPA. Meta-research: why research on research matters. *PLoS Biol* 2018;16:e2005468.
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, 2020. Available: https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf
5 Wasserstein RL, Lazar NA. The ASA Statement on *p* -Values: Context, Process, and Purpose. *Am Stat* 2016;70:129–33.
6 Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p < 0.05". *The American Statistician* 2019;73:1–19.

7 Greenland S, Senn SJ, Rothman KJ, *et al*. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* 2016;31:337–50.
8 Verhagen AP, Ostelo RWJG, Rademaker A. Is the p value really so significant?*. *Aust J Physiother* 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;4:279–82.
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;65:178–81.
12 Moher D, Hopewell S, Schulz KF, *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;44:560–4.
14 McLeod SA. What are confidence intervals in statistics? simply psychology, 2020. Available: https://www.simplypsychology.org/confidence-interval.html
15 Kallogjeri D, Spitznagel EL, Piccirillo JF. Importance of defining and interpreting a clinically meaningful difference in clinical research. *JAMA Otolaryngol Head Neck Surg* 2020;146:101–2.
16 Armijo-Olivo S, Saltaji H, da Costa BR, *et al*. 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:e008562.
17 Armijo-Olivo S, Fuentes J, da Costa BR, *et al*. Blinding in physical therapy trials and its association with treatment effects: a Meta-epidemiological study. *Am J Phys Med Rehabil* 2017;96:34–44.
18 de Almeida MO, Saragiotto BT, Maher C, *et al*. 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;100:1359–66.
19 Freire APCF, Elkins MR, Ramos EMC, *et al*. 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;23:302–10.
20 Thiese MS, Ronna B, Ott U. P value interpretations and considerations. *J Thorac Dis* 2016;8:E928–31.
21 McCambridge AB, Nasser AM, Mehta P, *et al*. 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;51:503–9.
22 Policy Statement. *Description of Physical Therapy [press release]. World Confederation for Physical Therapy*, 2019.
23 Gonzalez GZ, Moseley AM, Maher CG, *et al*. Methodologic quality and statistical reporting of physical therapy randomized controlled trials relevant to musculoskeletal conditions. *Arch Phys Med Rehabil* 2018;99:129–36.
24 Cashin AG, McAuley JH. Clinimetrics: physiotherapy evidence database (PEDro) scale. *J Physiother* 2020;66:59.
25 Moseley AM, Herbert RD, Maher CG, *et al*. Reported quality of randomized controlled trials of physiotherapy interventions has improved over time. *J Clin Epidemiol* 2011;64:594–601.
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. *Fam Med* 2005;37:360–3.
28 Catillon M. Trends and predictors of biomedical research quality, 1990-2015: a meta-research study. *BMJ Open* 2019;9:e030342.
29 Núñez-Cortés R, Alvarez G, Pérez-Bracchiglione J. Reporting results in manual therapy clinical trials: a need for improvement. *Int J Osteopath Med* 2021.
30 Riley SP, Swanson B, Brismée J-M, *et al*. A systematic review of orthopaedic manual therapy randomized clinical trials quality. *J Man Manip Ther* 2016;24:241–52.
31 Sedgwick P, Greenwood N. Understanding the Hawthorne effect. *BMJ* 2015;351:h4672.
32 Austin PC, Manca A, Zwarenstein M, *et al*. 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;63:142–53.
33 Harvey LA. Statistical testing for baseline differences between randomised groups is not meaningful. *Spinal Cord* 2018;56:919.
34 de Boer MR, Waterlander WE, Kuijper LDJ, *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.
35 Ferreira ML, Herbert RD, Ferreira PH, *et al*. A critical review of methods used to determine the smallest worthwhile effect of interventions for low back pain. *J Clin Epidemiol* 2012;65:253–61.
36 Kamper SJ. Interpreting outcomes 3-Clinical Meaningfulness: linking evidence to practice. *J Orthop Sports Phys Ther* 2019;49:677–8.
37 Moseley AM, Elkins MR, Van der Wees PJ. Using research to guide practice: the physiotherapy evidence database (PEDro). *Braz J Phys Ther* 2019:30914–1.
38 Butcher NJ, Monsour A, Mew EJ, *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 J-B, Hommel G, Röhrig B, *et al*. Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications. *Dtsch Arztebl Int* 2009;106:335–9.
40 Glasziou P, Altman DG, Bossuyt P, *et al*. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* 2014;383:267–76.
41 Chalmers I, Bracken MB, Djulbegovic B, *et al*. How to increase value and reduce waste when research priorities are set. *Lancet* 2014;383:156–65.
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, *et al*. 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* 2017;42:E680–6.