

REVIEW

Sample size calculations are poorly conducted and reported in many randomized trials of hip and knee osteoarthritis: results of a systematic review

Bethan Copsey^{a,*}, Jacqueline Y. Thompson^a, Karan Vadher^a, Usama Ali^a, Susan J. Dutton^a, Raymond Fitzpatrick^b, Sarah E. Lamb^a, Jonathan A. Cook^a

^aNuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS), University of Oxford, Botnar Research Centre, Windmill Road, Headington, Oxford OX3 7LD, UK

^bNuffield Department of Population Health, University of Oxford, Richard Doll Building, Old Road Campus, Oxford OX3 7LF, UK

Accepted 17 August 2018; Published online 23 August 2018

Abstract

Objectives: To review the methodology and reporting of sample size calculations in a contemporary sample of trials in osteoarthritis.

Study Design and Setting: Randomized trials in hip and/or knee osteoarthritis published in 2016 were identified by searching MEDLINE, Cochrane library, CINAHL, EMBASE, PsycINFO, PEDro, and AMED until March 31, 2017. Data were extracted on study characteristics, methods used to calculate the sample size, and the reporting and justification of components used in the sample size calculation. We attempted to replicate the sample size calculation using the reported information.

Results: This review included 116 trials. Seventy-eight (67%, $n = 78/116$) reported a power calculation. Less than a quarter reported all core components of the sample size calculation (21%, $n = 16/78$). The sample size calculation was only reproducible in 53% of the trials that reported a power calculation ($n = 41/78$). The replicated calculation produced a sample size over 10% larger than the reported value in 12% of trials ($n = 9/78$). Insufficient information was reported to allow the sample size calculation to be replicated in a quarter of trials (27%, $n = 21/78$).

Conclusion: Sample size calculations in trials of hip and knee osteoarthritis are not adequately reported, and the calculation frequently cannot be reproduced. © 2018 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Keywords: Sample size; Osteoarthritis; Reporting; Systematic review; Clinical trial; Research methods

Funding: This project is funded by a doctoral studentship from the EPSRC (Engineering and Physical Sciences Research Council) and Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences via the Medical Sciences Division of the University of Oxford. The funders had no input into the study design, the collection, analysis and interpretation of data, the writing of the report, or in the decision to submit the article for publication. Sarah E Lamb, Ray Fitzpatrick, Usama Ali, and Bethan Copsey receive funding from the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care Oxford at Oxford Health NHS Foundation Trust. Jonathan Cook is funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health.

Conflict of interest: None.

* Corresponding author. Tel.: +44 0 1865 737923.

E-mail address: bethan.copsey@csm.ox.ac.uk (B. Copsey).

1. Introduction

Sample size calculation is a key part of designing a clinical trial and is important for ethical, practical, and financial reasons. An overly large sample size can increase trial costs, delay dissemination of study findings, and result in more participants receiving a treatment when there is already sufficient evidence to show it is inferior to an alternative [1]. An overly small sample size can lead to underpowered trials that are more likely to “miss” a clinically important treatment effect, should it exist [2,3].

Altman et al. emphasized the importance of reporting the justification for the target sample size, especially when the trial does not recruit as many participants as planned [4]. When the sample size calculation is adequately reported, the reader can understand what the study was designed to achieve. The difference between the treatments that the trial was designed to statistically detect (the target difference), with associated assumptions, should be

What is new?**Key findings**

- Sample size calculations in hip and knee osteoarthritis trials are often poorly reported, and where reported, the calculation often cannot be replicated.
- The standard deviation assumed in the sample size calculation was a poor estimate of the observed standard deviation in a substantial proportion of trials.

What this adds to what was known?

- This is the first review of sample size calculations in trials of hip and knee osteoarthritis and shows that the problems of poor reporting and lack of reproducibility of sample size calculations exist in this clinical area.

What is the implication and what should change now?

- Trialists and reviewers should ensure that sample size calculations are reported clearly and completely to facilitate the interpretation of trial results and prevent the conduct of underpowered trials.
- Trialists should perform a sensitivity analysis at the design stage to explore how a difference in the estimate of the standard deviation could affect the power of the study.

specified [5]. If well justified, the target difference can inform the interpretation of the trial findings, clarifying the presence (or absence) of a meaningful difference. Appropriate calculation of the sample size and reporting of the calculation help to avoid research waste, preventing the conduct of trials that are likely to produce inconclusive and potentially misleading results.

Previous systematic reviews have found that power calculations are often not performed, inadequately reported, or based on inaccurate assumptions [5–7]. A study may be underpowered if the parameters used to calculate its sample size are based on inaccurate assumptions [8–10]. Reviews of trials in a handful of specific conditions, such as back pain and rheumatology, have highlighted poor reporting of sample size calculations [11–13]. Focusing on a specific clinical area reduces the heterogeneity in the assumptions made in the sample size calculation. For instance, oncology trials are more likely to be powered on survival, which is not usually applicable to low-mortality conditions such as osteoarthritis [14].

We explored whether recently published osteoarthritis trials also poorly reported their sample size calculations.

To our knowledge, the sample size calculations of hip and knee osteoarthritis trials have not previously been reviewed. Few reviews of any clinical area have attempted to replicate the sample size calculation of published trials [5,7,15,16]. Even fewer have compared the standard deviation assumed in the sample size calculation with the observed values in the trial results [7,9].

1.1. Objectives

Primary objective of this study was to summarize current practice in calculating the sample size for trials of hip and knee osteoarthritis, including the sample size, target difference, and justification for the chosen inputs.

Secondary objectives were to assess the reporting and reproducibility of these sample size calculations.

2. Materials and methods

The study methods were described in a published protocol and are summarized below [17].

2.1. Identification of studies

Seven databases were searched to identify relevant articles published in 2016: MEDLINE, Cochrane library (CENTRAL), CINAHL, EMBASE, PsycINFO, PEDro, and AMED (MEDLINE search strategy in Appendix B). The final search was performed on March 31, 2017, to allow for a 3-month lag between publication and database indexing.

2.2. Selection of studies

Abstracts and full texts were each screened independently by two of four reviewers (B.C., U.A., K.V., and J.Y.T.). Disagreements were resolved by discussion with a third reviewer (J.A.C.).

2.2.1. Inclusion criteria

Studies were eligible for inclusion if the article was the primary report of a randomized controlled trial of two treatment arms in a hip and/or knee osteoarthritis population. Included articles were published online or in a journal issue in 2016.

2.2.2. Exclusion criteria

The following article and study types were excluded:

- Conference abstracts
- Study protocols
- Non-English language articles
- Quasirandomized and nonrandomized studies
- Pilot and feasibility studies
- Factorial designs
- Cross-over trials
- Trials with three or more arms

- Studies that did not evaluate treatments (e.g., comparing different methods of providing information to improve patient knowledge)
- Studies examining osteoarthritis prevention
- Studies combining osteoarthritis and nonosteoarthritis populations (e.g., participants with osteoarthritis or rheumatoid arthritis, or trials of total knee arthroplasty where it was not explicitly stated that all participants had osteoarthritis)
- Secondary analyses of trials (e.g., long-term follow-up or subgroup analyses)

2.3. Data extraction

Data extraction on study characteristics included the study design, population, eligibility criteria, intervention and comparison treatments, and primary outcome. Data extraction on the sample size calculation included the target sample size, calculation method, values used, and justification (e.g., effect size, target difference, standard deviation, loss to follow-up, use of a one-tailed or two-tailed test, significance level, and power). Data extraction on the study results included the number of participants randomized, number lost to follow-up, and standard deviation of the primary outcome.

A second reviewer independently extracted the data from a sample of 20% of the included studies. Additional details of the sample size calculation were extracted from the study protocol if cited in the main article.

2.4. Sample size replication

Core values for the sample size calculation were defined as the power, significance level, whether a one-tailed or two-tailed test was used, level of attrition, and

- for continuous outcomes in superiority trials: the target difference as a standardized effect size or mean difference and standard deviation.
- for continuous outcomes in noninferiority trials: the noninferiority margin, mean difference, and standard deviation.
- for binary outcomes in superiority trials: any two of the anticipated between-group risk difference, effect in the intervention group, and effect in the control group.

We attempted to replicate the sample size calculations using the reported values. Unless otherwise stated, we assumed that 80% power and 5% two-tailed significance level with a superiority hypothesis were used, anticipating no attrition. For noninferiority trials, where not reported, we assumed the anticipated mean difference was 0.

To compare the replicated and reported target sample sizes, we calculated the ratio:

$$\frac{\text{replicated value} - \text{reported value}}{\text{reported value}}.$$

We present the number of studies with a replicated value over 10% or 30% above or below the reported sample size (ratio above 1.1 and 1.3 or below 0.9 and 0.7). The calculations were considered reproducible if the replicated value was within 10% of the reported value, to account for potential differences in software and rounding errors.

2.5. Data synthesis

For categorical and binary outcomes, data were summarized using the number and proportion of studies. For continuous outcomes, data were summarized using the median and interquartile range (IQR).

For continuous outcomes, the standard deviation assumed in the sample size calculation was compared with the corresponding value in the study results at the final follow-up time point. Again, we present the number of studies with a ratio above 1.1 and 1.3 or below 0.9 and 0.7.

2.6. Subgroup analyses

Subgroup analyses were performed to assess whether trial characteristics were associated with the number of participants randomized, whether the sample size calculation was reported and fully specified, and the reproducibility of the sample size calculation (reported value within 10% of the replicated value). Subgroup analyses examined differences by intervention (surgical or nonsurgical), number of trial centers (single center or multicenter), funding source (full/partial industry funding or no industry funding), and comparator treatment (placebo/waitlist or active control).

For continuous outcomes, subgroups were compared using the median difference and 95% confidence interval, estimated using Hodges–Lehmann [18,19]. For binary outcomes, subgroups were compared using absolute risk differences with 95% confidence intervals. A significance level of 0.05 was used. As the subgroup analyses were exploratory, no adjustments were made for multiple testing.

3. Results

3.1. Flow of studies and study characteristics

From database searches, 116 of 2955 articles were eligible for this review (Fig. 1). Most were parallel-group, superiority, single-center trials of knee osteoarthritis with nonindustry funding (Table 1). The number of randomized participants ranged from 20 to 633 (median 73). Only 13 trials (13/116, 11%) referred or provided access to the trial protocol. Although the studies used different definitions of osteoarthritis, several studies used the American College of Rheumatology criteria (41%, $n = 47/116$) [20,21]. The studies often restricted eligibility by age (76%, $n = 88/116$), prior surgery (66%, $n = 77/116$), and osteoarthritis

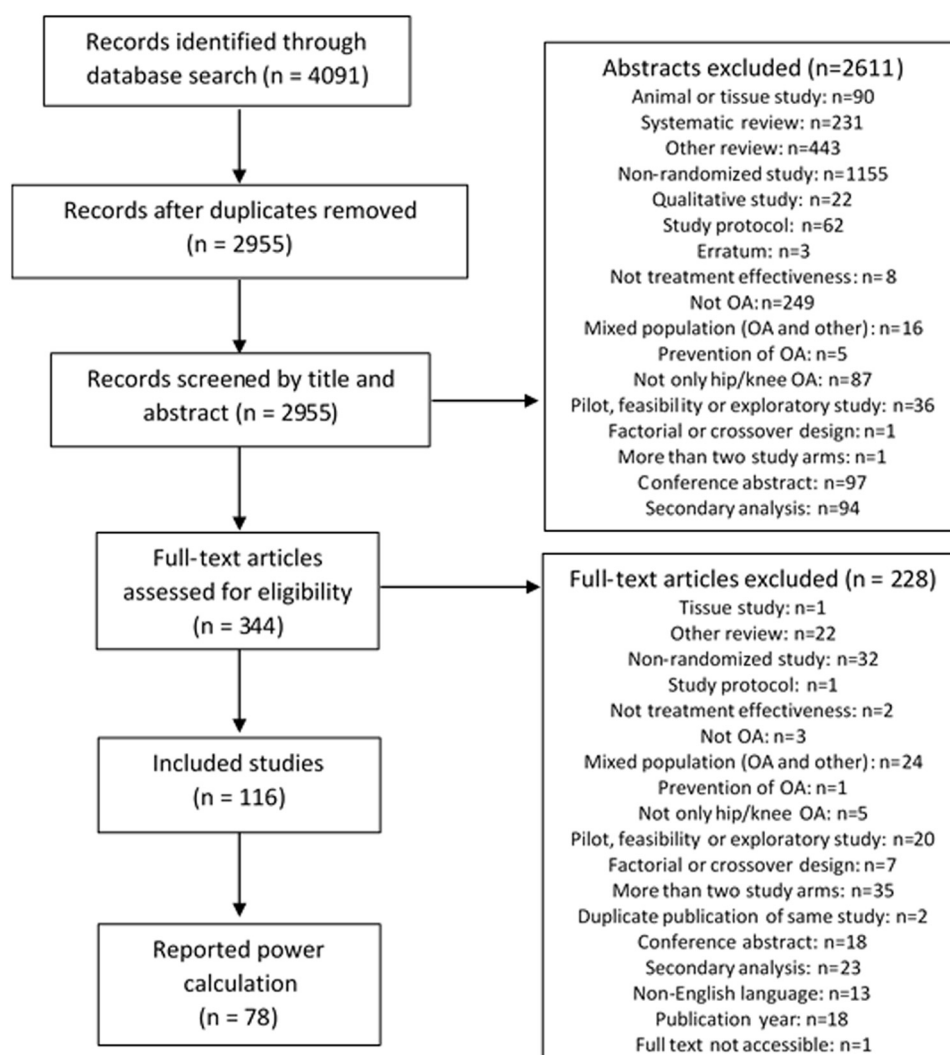


Fig. 1. Flow of studies in this review. OA, osteoarthritis.

severity, measured by Kellgren–Lawrence gradings (46%, $n = 53/116$) and pain symptoms (41%, $n = 48/116$) [22].

Of the 116 included trials, 78 (67%) reported a power calculation. Trials reporting a power calculation were more likely to have a larger sample size, cite the trial protocol, and report the trial funding source (Table 1). They were otherwise generally similar to trials not reporting a power calculation.

Among the 38 trials that did not report a power calculation, one reported a post hoc power calculation and six reported that a power calculation was conducted but did not provide details. Of the remaining 31 trials, four reported that the sample size was based on the predefined recruitment period and 27 did not justify the number of participants. In the 31 trials that did not report conducting a power calculation, 65% (20/31) mentioned the small sample size or lack of power calculation as a limitation of the trial and another 10% (3/31) stated that future trials with larger sample sizes were necessary.

The results that follow are based on the 78 trials that reported details of a power calculation.

3.2. Sample size calculation methodology

All of the included trials used a conventional (Neyman–Pearson or statistical hypothesis testing) power calculation approach [23,24]. None used alternative techniques, such as Bayesian approaches or simulations [25–27]. Two trials (3%, $n = 2/78$) reported a sample size calculation that was inappropriate for the study design; one used a sample size calculation for a paired sample when using an unpaired design and the other used a survey-based approach [28,29].

Most of the trials had a continuous primary outcome (97%, $n = 76/78$) (Table 2). One trial used a binary primary outcome, and none used a time-to-event primary outcome. The type of primary outcome in the remaining trial was unclear as the target difference was reported as a percentage. The trials were usually powered on one primary outcome (91%, $n = 71/78$).

Table 1. Study characteristics

Trial characteristics	Reported power calculation		Total
	Yes	No	
N (number of trials)	78	38	116
Study design			
Parallel-group randomized trial	77 (99%)	38 (100%)	115 (99%)
Cluster randomized trial	1 (1%)	0 (0%)	1 (1%)
Population			
Knee OA	69 (88%)	32 (84%)	101 (87%)
Hip OA	7 (9%)	4 (11%)	11 (9%)
Hip or knee OA	2 (3%)	2 (5%)	4 (3%)
Intervention			
Drug	21 (27%)	12 (32%)	33 (28%)
Surgery	17 (22%)	4 (11%)	21 (18%)
Exercise	19 (24%)	3 (8%)	22 (19%)
Other	21 (27%)	19 (50%)	40 (34%)
Comparator			
Active treatment	48 (62%)	27 (71%)	75 (65%)
Usual care	12 (15%)	5 (13%)	17 (15%)
Placebo or no treatment	18 (23%)	6 (16%)	24 (21%)
Study hypothesis			
Superiority	51 (65%)	18 (47%)	69 (59%)
Noninferiority	8 (10%)	1 (3%)	9 (8%)
Multiple	1 (1%)	0 (0%)	1 (1%)
Unclear	18 (23%)	19 (50%)	37 (32%)
Study centers			
Single centre	55 (71%)	28 (74%)	83 (72%)
Multicenter	16 (21%)	3 (8%)	19 (16%)
Unclear	7 (9%)	7 (18%)	14 (12%)
Funding source			
Industry	13 (17%)	2 (5%)	15 (13%)
Nonindustry	31 (40%)	9 (24%)	40 (34%)
Combination	4 (5%)	3 (8%)	7 (6%)
No funding	7 (9%)	3 (8%)	10 (9%)
Not reported	23 (29%)	21 (55%)	44 (38%)
Number randomized			
Median (IQR)	86.5 (55–150)	59 (40–76)	73 (50–120)
Range	26–633	20–140	20–633

Abbreviation: OA, osteoarthritis.

One trial planned to re-estimate the sample size if attrition was higher than expected, but this was found to be unnecessary. Three trials conducted unplanned sample size re-estimations (4%, $n = 3/78$) due to poor recruitment, low attrition, or post hoc analysis. Nine trials conducted sensitivity analyses on their sample size calculations (12%, $n = 9/78$), usually to assess the power for secondary outcomes ($n = 4$).

3.3. Reporting sample size calculations

Only 21% ($n = 16/78$) of the studies reported all of the core components of their sample size calculation. Almost all of the trials reported the power and significance level

(96%, $n = 75/78$). However, other components were not well reported, including the level of attrition (73%, $n = 57/78$) and whether a one-tailed or two-tailed test was used (41%, $n = 32/78$).

Almost all superiority trials powered on a continuous outcome reported the mean difference (90%, $n = 61/68$), and most reported the standard deviation (66%, $n = 45/68$). Most trials reported the standardized effect size or enough information to calculate it (79%, $n = 54/68$).

The single included cluster-randomized trial reported the intraclass correlation coefficient assumed in the power calculation to adjust for clustering. The single included trial with a binary outcome reported the risk difference between

Table 2. Reporting of components

Component of sample size calculation			n (%)
Type of primary outcome			
Continuous			76 (97%)
Binary			1 (1%)
Unclear			1 (1%)
Number of primary outcomes			
1			71 (91%)
2			3 (4%)
3			1 (1%)
Primary outcome not specified			3 (4%)
Alpha			
One-tailed		0.025	6 (8%)
		0.05	2 (3%)
Two-tailed		0.05	21 (27%)
		0.10	1 (1%)
		Not reported	1 (1%)
One-tailed and two-tailed		0.05	1 (1%)
Number of tails not reported or unclear		0.05	45 (58%)
		Not reported	1 (1%)
Power			
0.8			55 (71%)
0.85			3 (4%)
0.90			11 (14%)
0.95			4 (5%)
Other			1 (1%)
Multiple			2 (3%)
Not reported			2 (3%)
Attrition			
< 5%			2 (3%)
5%–9%			2 (3%)
10%–14%			15 (19%)
15%–19%			12 (15%)
20%–24%			20 (26%)
≥25%			6 (8%)
Not reported or unclear			21 (27%)

the two arms but not the anticipated effect in the intervention or control group.

Thirteen trials referred to a study protocol. Most of them (77%, $n = 10/13$) reported the sample size calculation consistently between the study protocol and results publication. The three trials with discrepancies reported (i) different target sample sizes, (ii) different levels of attrition, and (iii) only a revised sample size calculation in the results publication.

Most studies used a 5% significance level (88%, $n = 69/78$) or 2.5% for one-tailed tests (8%, $n = 6/78$). Statistical power was at least 80% for all studies; most used 80% power (71%, $n = 55/78$). The studies most commonly used 80% power and 5% significance level (65%, $n = 51/78$). Where reported, the level of attrition assumed in the sample size calculation was commonly 10%–24% (60%, $n = 47/78$) but ranged from 2% to 44%.

Where reported, the standardized effect size that the trial was powered to detect was moderate to large for most trials (median 0.75, IQR 0.50 to 0.86). Only one trial reported a standardized effect size of 0.2 or less for the target difference. However, it is likely that this target difference was incorrectly reported as the reproduced sample size for this trial was much greater than the reported sample size [30].

3.4. Justification of components

Table 3 shows the proportion of trials with continuous outcomes that justified the mean difference and standard deviation. The mean difference was most commonly based on a treatment difference from a published trial (29%, $n = 22/76$) or a published minimum clinically important difference (17%, $n = 13/76$). Where reported, justifications for the standard deviation used were almost exclusively based on previously published trials (33%, $n = 25/76$). Very few trials justified the anticipated level of attrition (3%, $n = 2/76$).

3.5. Reproducibility

Only half of the reported sample size calculations were reproducible (53%, $n = 41/78$) (Table 4). The replicated calculations produced a sample size over 10% larger than the reported value in 12% of trials ($n = 9/78$) (Fig. 2). One-quarter of the trials did not report enough information for us to replicate the sample size calculation (28%, $n = 22/78$). The sample size could be replicated in most of the trials that reported all of the core components, so that no assumptions were needed (88%, $n = 14/16$).

The absolute difference between the reproduced and reported sample size in terms of the number of participants was small for most studies (median difference 1 participant, IQR 0 to 5). Five studies showed a difference between the reproduced and reported sample size that was greater than 50 participants (9%, $n = 5/56$). Four of these five trials underestimated the sample size (replicated value over 30% larger than the reported value) and one overestimated it (replicated value at least 30% smaller than the reported value).

Table 3. Reporting of components for continuous outcomes ($n = 71$)

Component for continuous outcomes	Reported	Justification provided
Superiority trials: $n = 68$		
Standardized effect size	54 (79%) ^a	n/a
Mean difference	61 (90%)	38 (56%)
Standard deviation	45 (66%)	30 (44%)
Noninferiority trials: $n = 8$		
Noninferiority margin	8 (100%)	7 (88%)
Mean difference	3 (38%)	2 (25%)
Standard deviation	6 (75%)	5 (63%)

^a 12 (18%) of trials reported the target standardized effect size explicitly. For the remaining trials, the standardized effect size was calculated from the reported mean difference and standard deviation.

Table 4. Reproducibility of sample size calculation ($n = 78$)

Comparison of replicated value and reported value of target sample size	N (%)
Replicated value >30% larger than reported value	6 (8%)
Replicated value 10%–30% larger than reported value	3 (4%)
Replicated value within 10% of reported value	41 (53%) ^a
Replicated value 10%–30% smaller than reported value	1 (1%)
Replicated value >30% smaller than reported value	5 (6%)
Insufficient information	22 (28%)

^a This included four trials where additional assumptions were required on the interpretation of the reported information to replicate the sample size; for two trials, the reported target difference had to be translated into a different scale, and for two trials, the reported value and replicated value before accounting for attrition were compared because the anticipated attrition rate was unusually high.

3.6. Accuracy of components

Comparing the standard deviation of the primary outcome to the follow-up value, the standard deviation used in the sample size calculation was accurate (within 10%) in only one-third of trials (31%, $n = 9/29$) (Table 5). The follow-up standard deviation was over 30% larger than the value assumed in the sample size calculation in six trials (21%, $n = 6/29$), leading to a reduction in power.

3.7. Subgroup analysis

Exploratory subgroup analysis did not detect any significant differences in reporting based on funding source, study intervention, or comparator type (Appendix C). Multicenter trials recruited significantly larger sample sizes.

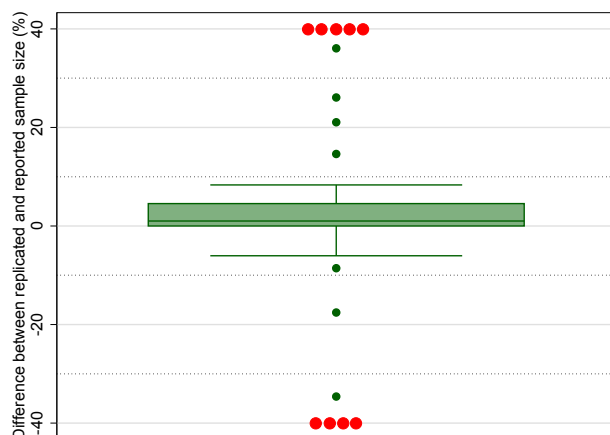


Fig. 2. Comparison of reported and replicated sample size ($n = 56$). Comparison of reproduced and reported sample sizes (as percentage of reported value). The red markers represent trials excluded from the figure where the difference was over 50% (five trials) and where the difference was below –50% (four trials). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 5. Accuracy of the standard deviation ($n = 29$)^a

Accuracy of standard deviation (SD)	n (%)
Follow-up SD >30% larger than assumed value	6 (21%)
Follow-up SD 10%–30% larger than assumed value	3 (10%)
Follow-up SD within 10% of assumed value	9 (31%)
Follow-up SD 10%–30% smaller than assumed value	5 (17%)
Follow-up SD >30% smaller than assumed value	6 (21%)

^a For 47 trials, the assumed and follow-up standard deviation could not be compared. For two trials, this was not applicable because the primary outcome was not continuous.

3.8. Example of good reporting of a sample size calculation

Beselga et al. reported all of the core components and justified their target difference and standard deviation with a citation [31]:

“The sample size was calculated using Ene 3.0 software ... detecting differences of 2.0 units in the NPRS, considered as the minimum clinical important difference (MCID) (Farrar et al., 2001), assuming a standard deviation of 1.7 (based on pilot data), an α of 0.05, β of 90%, and a 2-tailed t -test. The estimated sample size was calculated to be at least 17 subjects in each group ... increased to 20 subjects in each group to allow a drop out of 15%.”

4. Discussion

4.1. Summary of findings

This systematic review summarizes current practice in the methodology, reporting, and replicability of sample size calculations in randomized trials of hip and knee osteoarthritis. Two-thirds of the trials reported a sample size calculation. Most of the remaining one-third made no reference to their choice of sample size. Almost all sample size justifications were based on a conventional power calculation approach. The sample size calculation was fully described in very few studies. The sample size calculation could often not be replicated.

The studies most commonly omitted the anticipated attrition or standard deviation. Where reported, the justification for the target difference was based on findings of previous trials and/or a published estimate of the minimum clinically important difference. The standard deviation was commonly based on the results of previously published trials. However, for many trials, the standard deviation assumed in the power calculation was inaccurate (either too small or too large) when compared to the follow-up results of the trial. Underestimating the standard deviation

can seriously affect a trial's ability to detect a meaningful treatment difference.

4.2. Comparison with related literature

This review shows that poor reporting is a problem, specifically in osteoarthritis trials, despite it being a well-developed area of clinical research. Overall, the reporting of sample size calculations found in this study was similar to that found in other clinical and methodological areas, with previous reviews finding that 50%–70% of trials reported a sample size calculation and around 25% of sample size calculations included all core components [12,13,32–34]. Although one review across multiple clinical areas found a higher proportion of studies reporting a power calculation (95%) and reporting all core components, this may be because it only considered publications in journals with a high impact factor [7].

The level of reporting of specific components of the sample size calculation (e.g., power, significance level, target difference) was consistent with other reviews [7,15,16]. Although our review agreed with Rutterford et al. that the assumed standard deviation, level of attrition, and justification for the target treatment difference were poorly reported, Rutterford et al. found much lower levels of reporting for these components [33]. This could be due to differences in reporting practices over time or between clinical areas. The target difference was most commonly justified using previous trials, which aligns with the results of a survey suggesting high awareness and endorsement of this method among trialists [35]. However, the survey also reported that trialists commonly used pilot studies to justify the target difference, which was rare in the osteoarthritis trials in our review [35].

There are mixed findings in the literature on the reproducibility of sample size calculations. Some reviews have found a similar level of discrepancies between the replicated and reported sample sizes to our review [5,16]. Reviews that found a higher quality of reporting also found a much higher proportion of replicable calculations [7,15].

4.3. Strengths and limitations

The key strengths of this review are the systematic search strategy and restricted eligibility criteria. Restricting eligibility to trials of hip and knee osteoarthritis produced a more homogeneous sample in terms of the outcome measures used and population from which the trials recruited, compared to reviews considering trials in any clinical area. By including a contemporary sample of trials, this review should provide insight into the clinical trial methodology used in current practice.

The main limitation of this review is that, while the overall sample was substantial, some of the subgroups were small. The findings of these subgroup analyses should be interpreted cautiously. As the review was restricted to trials published in 2016, we cannot draw conclusions about changes in reporting and methodology over time. The

sample of included articles may be less representative of lower impact journals, as they usually take longer time to be indexed in databases [36].

Our assessment of sample size calculations relied on published information and thus was hindered by poor reporting. For example, a trial's sample size may have been calculated using an appropriate power calculation without this being reported in the results paper. It is also possible that a sample size calculation may have been modified after trial design but before publication [37]. Therefore, the a priori sample size calculation conducted during the study design stage may not have been described accurately in the reported results paper. A review of trial protocols or ethics applications may more accurately reflect sample size calculations done during the design phase [5]. The results of this review may not be applicable to other clinical areas, particularly where dichotomous or time-to-event primary outcomes are common.

4.4. Implications

Although there are examples of good practice in the literature, it is concerning that one-third of trials of hip and knee osteoarthritis made no reference to the choice of sample size. When a power calculation was reported, there was often insufficient information to reproduce the calculation. Sample size calculations thus often cannot be verified, making it difficult for readers to interpret the trial results in view of the assumptions made when the study was designed. There is potential for improvement in the reporting of the predicted level of attrition, standard deviation, use of a one-tailed or two-tailed test, and justification for values used in the calculation.

For some trials, the value produced by attempting to replicate the sample size calculation was very different to the sample size reported in the trial publication. The reported information may have been misleading or inaccurate. Alternatively, there may have been a fundamental error in the original calculation, as was the case for at least a few trials where the calculation was clearly inappropriate for the trial design. Our results highlight inaccuracies in the standard deviation assumed in sample size calculations. Underestimating the standard deviation can lead to underpowered trials. Trialists should prespecify the target difference in terms of the between-group difference in the original scale of the primary outcome measure and the corresponding standard deviation, rather than specifying only the standardized effect size [38]. Trialists should perform sensitivity analyses to explore how changes in the assumed standard deviation will affect study power.

The poor reporting and lack of reproducibility of sample size calculations found in this review contribute to research waste [39,40]. Conducting a power calculation when designing a study can prevent underpowered trials from being carried out if they are likely to be uninformative. Clear and complete reporting of a power calculation allows the reader to see the primary outcome and the treatment effect believed to be clinically meaningful [41,42]. This helps the reader to

interpret the trial results in terms of the likelihood of a false result and the clinical relevance of the findings. Although reporting guidelines have attempted to improve reporting, there is a clear need for statistically trained peer reviewers to ensure adequate reporting of sample size calculations in trial protocols, funding applications, ethical approval, and trial results publications [40]. Trial teams should be encouraged to involve members with formal training in statistics and research design early on in their trials [40].

4.5. Future research

Future research could explore the reasons for the lack of reproducibility of sample size calculations, for example, by contacting trial teams for additional information where the methods of calculation are unclear or deemed inaccurate. Future work could also explore whether other factors are associated with high-quality reporting of the sample size calculation, such as statistical peer review [43–45]. Future studies could examine the values of components used in sample size calculations in more detail, such as assessing the clinical relevance of the target differences or developing methods to more accurately predict the standard deviation used in the power calculation.

5. Conclusion

Sample size calculations in trials of hip and knee osteoarthritis are not consistently reported adequately. Even when reported in sufficient detail, the calculation cannot always be accurately reproduced. This raises concerns about whether the sample size calculation was performed correctly and whether the trial was appropriately designed to achieve its primary objective. It also makes it difficult to establish how likely it is that a meaningful difference between the treatments exists. Clear and accurate reporting of a sample size calculation (or justification) should be mandatory, with endorsement by journal editors and peer reviewers for grant applications, trial protocols, and results publications.

Acknowledgments

The authors would like to acknowledge Dr Jennifer A de Beyer for her assistance in editing the article.

Supplementary data

Supplementary data related to this article can be found at <https://doi.org/10.1016/j.jclinepi.2018.08.013>.

References

- [1] Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. *Designing clinical research*. Philadelphia, Pennsylvania, USA: Lippincott Williams & Wilkins; 2013.
- [2] Altman DG. Statistics and ethics in medical research: III How large a sample? *Br Med J* 1980;281:1336–8.
- [3] Halpern SD, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. *JAMA* 2002;288:358–62.
- [4] Altman DG, Moher D, Schulz KF. Peer review of statistics in medical research. Reporting power calculations is important. *BMJ* 2002;325:491. author reply 491.
- [5] Clark T, Berger U, Mansmann U. Sample size determinations in original research protocols for randomised clinical trials submitted to UK research ethics committees: review. *BMJ* 2013;346:f1135.
- [6] Fernandes-Taylor S, Hyun JK, Reeder RN, Harris AH. Common statistical and research design problems in manuscripts submitted to high-impact medical journals. *BMC Res Notes* 2011;4:304.
- [7] Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting of sample size calculation in randomised controlled trials: review. *BMJ* 2009;338:b1732.
- [8] Tavernier E, Giraudeau B. Sample size calculation: inaccurate a priori assumptions for nuisance parameters can greatly affect the power of a randomized controlled trial. *PLoS One* 2015;10(7):e0132578.
- [9] Vickers AJ. Underpowering in randomized trials reporting a sample size calculation. *J Clin Epidemiol* 2003;56:717–20.
- [10] Chen H, Zhang N, Lu X, Chen S. Caution regarding the choice of standard deviations to guide sample size calculations in clinical trials. *Clin Trials* 2013;10(4):522–9.
- [11] Keen HI, Pile K, Hill CL. The prevalence of underpowered randomized clinical trials in rheumatology. *J Rheumatol* 2005;32(11):2083–8.
- [12] Abdul Latif L, Daud Amadera JE, Pimentel D, Pimentel T, Fregni F. Sample size calculation in physical medicine and rehabilitation: a systematic review of reporting, characteristics, and results in randomized controlled trials. *Arch Phys Med Rehabil* 2011;92:306–15.
- [13] Froud R, Rajendran D, Patel S, Bright P, Bjorkli T, Eldridge 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(11):E680–6.
- [14] Bariani GM, de Celis Ferrari AC, Precivale M, Arai R, Saad ED, Riechelmann RP. Sample size calculation in oncology trials: quality of reporting and implications for clinical cancer research. *Am J Clin Oncol* 2015;38:570–4.
- [15] Abdulatif M, Mukhtar A, Obayah G. Pitfalls in reporting sample size calculation in randomized controlled trials published in leading anaesthesia journals: a systematic review. *Br J Anaesth* 2015;115:699–707.
- [16] Koletsis D, Fleming PS, Seehra J, Bagos PG, Pandis N. Are sample sizes clear and justified in RCTs published in dental journals? *PLoS One* 2014;9:e85949.
- [17] Copsey B, Dutton S, Fitzpatrick R, Lamb SE, Cook JA. Current practice in methodology and reporting of the sample size calculation in randomised trials of hip and knee osteoarthritis: a protocol for a systematic review. *Trials* 2017;18(1):466.
- [18] Hodges JL Jr, Lehmann EL. Estimates of location based on rank tests. *Ann Math Stat* 1963;34(2):598–611.
- [19] Newson R. somersd-Confidence intervals for nonparametric statistics and their differences. *Stata Tech Bull* 2001;10(55):47–54.
- [20] Altman R, Alarcon G, Appelrouth D, Bloch D, Borenstein D, Brandt K, et al. The American college of rheumatology criteria for the classification and reporting of osteoarthritis of the hip. *Arthritis Rheum* 1991;34:505–14.
- [21] Altman R, Asch E, Bloch D, Bole G, Borenstein D, Brandt K, et al. Development of criteria for the classification and reporting of osteoarthritis. Classification of osteoarthritis of the knee. Diagnostic and therapeutic criteria committee of the American rheumatism association. *Arthritis Rheum* 1986;29:1039–49.
- [22] Kellgren JH, Lawrence JS. Radiological assessment of osteoarthritis. *Ann Rheum Dis* 1957;16(4):494–502.
- [23] Neyman J, Pearson ES. The testing of statistical hypotheses in relation to probabilities a priori. In: *Mathematical Proceedings of the Cambridge*

- Philosophical Society. Cambridge, UK: Cambridge University Press; 1933:492–510.
- [24] Julious SA. Sample sizes for clinical trials. Florida, USA: CRC Press; 2009.
- [25] Stamey JD, Natanegara F, Seaman JW Jr. Bayesian sample size determination for a clinical trial with correlated continuous and binary outcomes. *J Biopharm Stat* 2013;23(4):790–803.
- [26] Ciarleglio MM, Arendt CD, Peduzzi PN. Selection of the effect size for sample size determination for a continuous response in a superiority clinical trial using a hybrid classical and Bayesian procedure. *Clin Trials* 2016;13(3):275–85.
- [27] Cao J, Lee JJ, Alber S. Comparison of Bayesian sample size criteria: ACC, ALC, and WOC. *J Stat Plann Inference* 2009;139(12):4111–22.
- [28] Dincer U, Ariba S, Saygin H, Incedayi M, Rodop O. The effects of closed kinetic chain exercise on articular cartilage morphology: myth or reality? A randomized controlled clinical trial. *Turk J Phys Med Rehab* 2016;62(1):28–36.
- [29] Notarnicola A, Maccagnano G, Moretti L, Pesce V, Tafuri S, Fiore A, et al. Methylsulfonylmethane and boswellic acids versus glucosamine sulfate in the treatment of knee arthritis: randomized trial. *Int J Immunopathol Pharmacol* 2016;29:140–6.
- [30] Banerjee M, Mondal S, Sarkar R, Mondal H, Bhattacharya K. Comparative study of efficacy and safety of tapentadol versus etoricoxib in mild to moderate grades of chronic osteoarthritis of knee. *Indian J Rheumatol* 2016;11:21–5.
- [31] Beselga C, Neto F, Albuquerque-Sendin F, Hall T, Oliveira-Campelo N. Immediate effects of hip mobilization with movement in patients with hip osteoarthritis: a randomised controlled trial. *Man Ther* 2016;22:80–5.
- [32] Arnup SJ, Forbes AB, Kahan BC, Morgan KE, McKenzie JE. The quality of reporting in cluster randomised crossover trials: proposal for reporting items and an assessment of reporting quality. *Trials* 2016;17(1):575.
- [33] Rutterford C, Taljaard M, Dixon S, Copas A, Eldridge S. Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review. *J Clin Epidemiol* 2015; 68:716–23.
- [34] Castellini G, Gianola S, Bonovas S, Moja L. Improving power and sample size calculation in rehabilitation trial reports: a methodological assessment. *Arch Phys Med Rehabil* 2016;97:1195–201.
- [35] Cook JA, Hislop J, Adewuyi TE, Harrild K, Altman DG, Ramsay CR, et al. Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review. *Health Technol Assess* 2014;18(28):v–vi. 1–175.
- [36] Irwin AN, Rackham D. Comparison of the time-to-indexing in PubMed between biomedical journals according to impact factor, discipline, and focus. *Res Social Adm Pharm* 2017;13(2):389–93.
- [37] Ramagopalan S, Skingsley AP, Handunnetthi L, Klingel M, Magnus D, Pakpoor J, et al. Prevalence of primary outcome changes in clinical trials registered on ClinicalTrials.gov: a cross-sectional study. *F1000Res* 2014;3:77.
- [38] Cook JA, Hislop J, Altman DG, Fayers P, Briggs AH, Ramsay CR, et al. Specifying the target difference in the primary outcome for a randomised controlled trial: guidance for researchers. *Trials* 2015;16:12.
- [39] Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* 2014;383(9913):267–76.
- [40] Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* 2014;383(9912):166–75.
- [41] Williamson P, Hutton J, Bliss J, Blunt J, Campbell M, Nicholson R. Statistical review by research ethics committees. *J R Stat Soc Ser A* 2000;163(1):5–13.
- [42] Campbell MJ. Doing clinical trials large enough to achieve adequate reductions in uncertainties about treatment effects. *J R Soc Med* 2013;106(2):68–71.
- [43] Black N, van Rooyen S, Godlee F, Smith R, Evans S. What makes a good reviewer and a good review for a general medical journal? *JAMA* 1998;280:231–3.
- [44] Cobo E, Selva-O'Callaghan A, Ribera JM, Cardellach F, Dominguez R, Vilardell M. Statistical reviewers improve reporting in biomedical articles: a randomized trial. *PLoS One* 2007;2:e332.
- [45] Costa ML, Griffin XL, Parsons N, Dritsaki M, Perry D. Efficacy versus effectiveness in clinical trials. *Bone Joint J* 2017;99-b(4):419–20.