Review

Assessment of the Methodological Quality of Medical and Surgical Clinical Trials in Patients with Arthroplasty

JASVINDER A. SINGH, STEPHEN MURPHY, and MOHIT BHANDARI

ABSTRACT. ***Objective***. To assess the methodological quality of randomized controlled trials (RCT) of medical and surgical therapy in patients with arthroplasty.

Methods. We conducted a Medline database search for all arthroplasty RCT from 1997 and 2006. The quality of the methods of all eligible RCT was assessed by a trained abstractor. We used a checklist of trial quality characteristics, and the overall trial quality was assessed by 3 scales: Jadad (range 0–5), Delphi list (range 0–9), and numeric rating scale (NRS; range 1–10), based on User's Guides to the Medical Literature.

Results. A total of 196 articles were included in the analysis; most included hip (n = 81) or knee (n = 80) or both hip/knee arthroplasty (n = 19); 66 (34%) assessed pharmacological treatments, 117 (60%) nonpharmacological treatments, and 13 (7%) both. Mean (SEM) overall quality scores of arthroplasty RCT were low: Jadad score 2.36 (1.4), Delphi list 5.33 (1.6), and NRS score 4.30 (2.6). Multivariable analyses revealed that nonpharmacological intervention RCT had lower odds (odds ratio 0.28–0.39; p = 0.008–0.033) and those with no funding had lower odds (OR 0.28–0.50; p = 0.014–0.119) of being in the highest quartiles of the 3 overall quality scores. In contrast, multicenter RCT had 1.8–4.7 times higher odds of being in highest tertiles of quality scores (p = 0.017–0.185).

Conclusion. Methodological deficiencies in reporting of hip/knee arthroplasty RCT offer an opportunity for improvement. Type of intervention, number of trial centers, and presence of funding were independently associated with overall trial quality. In future, multicenter RCT (rather than single-center) and modeling protocols of single-center RCT similar in rigor to multicenter RCT may improve the quality of arthroplasty RCT. (J Rheumatol First Release Nov 1 2009; doi:10.3899/jrheum.090333)

Key Indexing Terms:
ASSESSMENT QUALITY ARTHROPLASTY CLINICAL TRIALS

In the US, arthritis and other rheumatic conditions affected an estimated 70 million people in 2001[1], led to 744,000 hospitalizations and 44 million ambulatory care visits in 1997[2], and cost $149 billion in direct and indirect costs in 1992 (2.5% of the gross national product)[3]. Arthritis leads to significant physical and psychological morbidity[4,5] and is the leading cause of disability in adults in the US[6]. Joint arthroplasty is the most significant advance in treatment of patients with endstage arthritis; 202,500 primary total hip arthroplasty (THA) and 402,100 primary total knee arthroplasty (TKA) procedures were performed in the US in 2003[7]. Arthroplasty is associated with relief of pain and improvement of function and quality of life[7]. THA has been called "the operation of the century"[8].

Due to significant public health burden and cost associated with hip and knee arthroplasty, we need high quality evidence upon which physicians and patients can base their decisions. To our knowledge, there are no published reports assessing the quality of arthroplasty randomized controlled trials (RCT).

A recent systematic review of RCT in osteoarthritis found differences between pharmacological and nonpharmacological RCT[9]. We conducted a systematic review of the available literature to examine the quality of reporting across randomized trials in arthroplasty. Specifically, we aimed (1) to examine the methodological quality of arthro-

From the Department of Medicine, University of Minnesota, and Minneapolis VA Medical Center, Minneapolis, Minnesota; Mayo Clinic School of Medicine, Rochester, Minnesota; University of Notre Dame, Notre Dame, Indiana, USA; and Division of Orthopaedic Surgery, McMaster University, Hamilton, Ontario, Canada.

Supported by the NIH CTSA Award 1 KL2 RR024151-01 (Mayo Clinic Center for Clinical and Translational Research); and the Minneapolis VA Medical Center, Minneapolis, MN.

J.A. Singh, MBBS, MPH, Department of Medicine, University of Minnesota and Minneapolis VAMC, and Mayo Clinic School of Medicine; S. Murphy, BS, University of Notre Dame; M. Bhandari, MD, MSc, Division of Orthopaedic Surgery, McMaster University.

Address correspondence to Dr. J.A. Singh, Minneapolis VA Medical Center, Rheumatology Office (111 R), One Veterans Drive, Minneapolis, MN 55417. E-mail: jasvinder.md@gmail.com

Accepted for publication July 9, 2009.

plasty RCT; and (2) to study whether intervention (pharmacological vs nonpharmacological), trial (funding source, number of centers, number of patients per trial), or publication (year of publication, type of journal, journal impact factor) characteristics were associated with overall trial quality (Jadad score, etc.) and with specific quality standards (allocation concealment, use of placebo, etc.).

MATERIALS AND METHODS

Search strategy. Medline was searched by a librarian from the Cochrane Library Systematic Review Group (IR) using the following search terms for arthroplasty: "exp arthroplasty, Replacement, Knee/ or exp Joint Prosthesis/ or exp Arthroplasty, Replacement/ or joint arthroplasty.mp. or exp Arthroplasty, Replacement, Hip." This search was further limited to RCT published in the 2 calendar years 2006 (the most recent year at the time of review) and 1997 (a year about a decade earlier), to examine if the quality of RCT had changed over a decade. Upon review of the titles and abstracts by a senior author (JS), articles were excluded if they were letter/editorial, nonrandomized, or published in non-English language, were not arthroplasty-related, or did not include clinical outcomes (i.e., economic analyses, etc.). There were no restrictions by the journal name or specialty.

Detailed evaluation of study quality. Training of a single abstractor (SM) the senior epidemiologist (JS) consisted of: (1) review of the literature and key articles describing the quality assessments of trials; (2) detailed discussion of key assessment components, including allocation concealment, blinding etc.; (3) 3 rounds of independent abstraction of articles (14 articles) by both the senior author (JS) and the trained abstractor (SM), which led to > 95% agreement on all abstracted data. After the training period, SM, who was blinded to the study hypotheses, assessed and abstracted trial quality data from all included studies using a structured abstraction form, modified from that used by Boutron, *et al*[9]. Data were entered into forms created using Microsoft Access 2003 (MicroSoft, Redmond, WA, USA) (Appendix 1).

We obtained the following characteristics for each included study: (1) year of publication, journal, title; (2) body region involved — upper (shoulder, elbow, hand) or lower extremity (hip, knee, foot, long bones); (3) financial support — public, private, neither, both, or not clear; (4) number of centers involved — single center, multicenter, not clear; (5) number of patients/study: ≤ 50, 51–100, 101–200, 201–500, and > 500; (6) treatment classification: pharmacologic (oral, topical, intramuscular, intravenous, intraarticular, or other) versus nonpharmacological (surgery, arthroscopy, joint lavage, acupuncture, rehabilitation, behavioral, or other); (7) type of study — original versus followup/subgroup analyses; (8) type of journal — orthopedics/surgery, anesthesia, internal medicine/medical subspecialties, and rehabilitation/others; and (9) journal impact factor — classified as ≤ 0.5, > 0.5–1, > 1–2, > 2–5, > 5–10, and > 10. Impact factor and number of patients were categorized due to a skewed distribution.

We examined whether the CONSORT (Consolidated Standards of Reporting Trials) criteria[10] were reported in a flowchart or in the text and whether the loss to followup was < 20%. We assessed trial design, mode of randomization, blinding, and outcome assessment. The CONSORT checklist was not used, since this was described in 2001, after one of the years of included articles (1997). Generation of randomization sequence was considered (1) appropriate if selection bias was prevented by use of random numbers, computerized random number generation, pharmacy controlled, opaque sealed envelopes, numbered or coded bottles; (2) inappropriate if patients were allocated alternately, according to date of birth, date of admission, hospital number etc.; and (3) indeterminate. Allocation was considered concealed if both patients and investigators enrolling patients in the study could not foresee the assignments due to centralized randomization/pharmacy control/opaque envelopes, etc.

We assessed if blinding of patients, care providers, and outcome assessors was reported, if it was appropriate[11], whether it was theoretically efficient, and whether it was tested. Appropriateness of blinding was categorized as follows: (1) appropriate — stated that neither person doing assessments nor study participant could identify the intervention being tested or use of active placebos, identical placebos, or dummies; (2) inappropriate — comparison of tablet versus injection with no double-dummy; and (3) indeterminate.

The following details regarding the intervention were extracted: (1) Was the intervention individualized (i.e., treatment modification according to individual's profile)?; (2) Was the intervention described in enough detail to be reproducible?; (3) Was there a control intervention? If so, was this placebo, active control, usual care, waiting list, or other?; (4) Was the potential placebo effect of each treatment similar?; (5) Was the quality of intervention and control intervention assessed?; (6) Could care providers influence the treatment effect? If so, was this due to their experience, learning curve, or training of care providers at the beginning of the trial?; (7) Was there a contamination of the 2 groups (by providing intervention to the control group)?; (8) Were concomitant treatments reported?; and (9) Was treatment compliance tested?; (10) If tested, how was it assessed (pill counts, patient report, video, reporting diary, not reported)?

The statistical analysis section was examined to determine whether a trial reported a justification for sample size, whether the analyses were described as intention to treat analysis (ITT), i.e., all participants randomized were included in the analysis and kept in the original groups[12], or modified ITT, i.e., analysis excluded those who never received treatment or who were never evaluated while receiving treatment.

Study outcomes. Outcomes included reporting of each trial quality characteristic and the overall quality assessment. Trial quality was assessed in detail by examining the adequacy of reporting of allocation concealment, generation of allocation (i.e., randomization) sequence, use of placebo, CONSORT diagram, reproducibility of intervention, loss to followup, adverse events, sample size justification, use of intention to treat analysis, and blinding of patients, care providers and outcome assessors.

Overall trial quality was assessed using 3 validated measures: Jadad score[13,14], Delphi list's overall score[15], and overall subjective assessment of validity of the study as described in the Users' Guides to the Medical Literature[16]. Jadad scale assesses the appropriateness of randomization, blinding, and loss to followup, and ranges from 0 to 5. The Delphi list includes 9 items that assess trial characteristics on a 0–9 score, including randomization, similarity at baseline, eligibility criteria, allocation concealment, blinding of outcome assessor, patient and care provider, inclusion of ITT, and report of point estimates and variability. The overall subjective evaluation of the study's quality was assessed on a numerical rating scale (NRS) ranging from 1 to 10 by answering the question, "To what extent were systematic errors or bias avoided in this report?". We included multiple scales of overall quality for 2 reasons: (1) Jadad scale is heavily weighted to double-blinding, which is often not possible in surgical RCT; we therefore included the Delphi list, which has no points for use of placebo and awards only one point each for blinding of care providers, assessors, and patients; and (2) for robustness of analyses. For all 3 measures, a higher score indicates higher quality.

Statistical analysis. For continuous measures, we calculated mean and standard error of the mean and for categorical variables the frequencies and percentages. We used chi-square and independent sample Student's t tests to examine the univariate association of trial, intervention, and publication characteristics with trial quality — assessed by both individual quality characteristic (allocation concealment, etc.) and the overall trial quality (Jadad scale, Delphi list, and subjective overall score), respectively. We performed 3 separate multivariable-adjusted logistic regression analyses to assess which of the trial characteristics significant in the univariate analyses were independently associated with overall trial quality, outcome being the highest tertiles of Jadad, NRS, and Delphi list scores. The cutoffs for the highest tertiles were ≥ 3 on Jadad score, ≥ 6 on NRS, and ≥ 6 on the Delphi list score. Variables with a right-skewed distribution, i.e., the journal impact factor and the number of patients, were categorized into dichotomous and categorical variables, respectively, allowing enough numbers in each category.

Sensitivity analyses were done for the above-described multivariable analyses by considering 2 predictors, impact factor and number of patients, as continuous variables instead of categorical variables as in the previous models. All tests were 2-sided, and we considered $p < 0.05$ statistically significant.

RESULTS

Characteristics of included studies. After screening the abstracts and full text, excluding non-English language, 196 articles were eligible for abstraction, 67 from year 1997 and 129 from 2006 (Figure 1). Of these, 130 articles assessed nonpharmacological therapy and 79 assessed pharmacological therapies (13 articles assessed both). Eighty articles included surgical interventions, 17 rehabilitation therapy, 3 education intervention, 3 behavior therapy, 22 oral medications, 25 parenteral, 40 intraarticular, and 30 other interventions (Figure 1).

Of the 196 studies included for analyses, 81 included only THA, 80 only TKA, and 19 both THA and TKA (Table 1). Over one-third of studies (35%) included sample sizes of 50 patients or less. The studies were primarily published in orthopedics (64%) or anesthesia journals (17%), with a few in internal medicine and related subspecialty (14%) and rehabilitation/other journals (6%).

Methodological quality of included studies — univariate analyses for overall quality and individual quality characteristics. The overall quality of studies was low: Jadad score was 2.36 (range 0–5), Delphi list scale score 5.33 (range 0–9), and overall NRS score 4.3 (range 1–10); scores were at or below the mean of the range of each scale (Table 2). Univariate analyses showed that type of intervention, number of centers, number of patients, funding source, type of journal, and journal impact factor were significantly associated with overall quality (Table 2). The year of publication was not associated with overall RCT quality in univariate analyses.

Examination of individual study characteristics revealed that a low proportion of studies described the following: adequate generation of allocation sequence (43%); allocation concealment (39%); CONSORT diagram (10%); blinding of patients (31%), care providers (17%), and outcome assessors (45%); use of ITT or modified ITT for analyses (10%); and sample size justification (36%). Only 18% of the studies used placebo and 51% reported < 20% loss to followup. Since it may not be possible/ethical to blind patients/care providers or use placebo in surgical trials, when restricting this to pharmacological trials, numbers were still low at 61%, 42%, and 47%. On the other hand, some quality indicators were reasonably well described, including adverse event reporting (57%), potential similarity of placebo to treatment (64%), and reproducibility of intervention (97%) (Table 3).

RCT of pharmacological interventions or those that had both pharmacologic and nonpharmacological interventions had significantly better quality standards than nonpharmacological intervention RCT (Table 3). Specifically, trials of surgical or rehabilitation interventions had significantly lower use of placebo, blinding of patients, care providers or outcome assessors, or sample size justification (Appendix 1). Similar deficits were noted in individual quality characteristics in small sample size RCT, compared to larger sample size RCT (Table 4).

Studies published in 1997 were significantly less likely than those published in 2006 to describe allocation concealment or provide sample size justification, but were more likely to describe the blinding of care providers or outcome

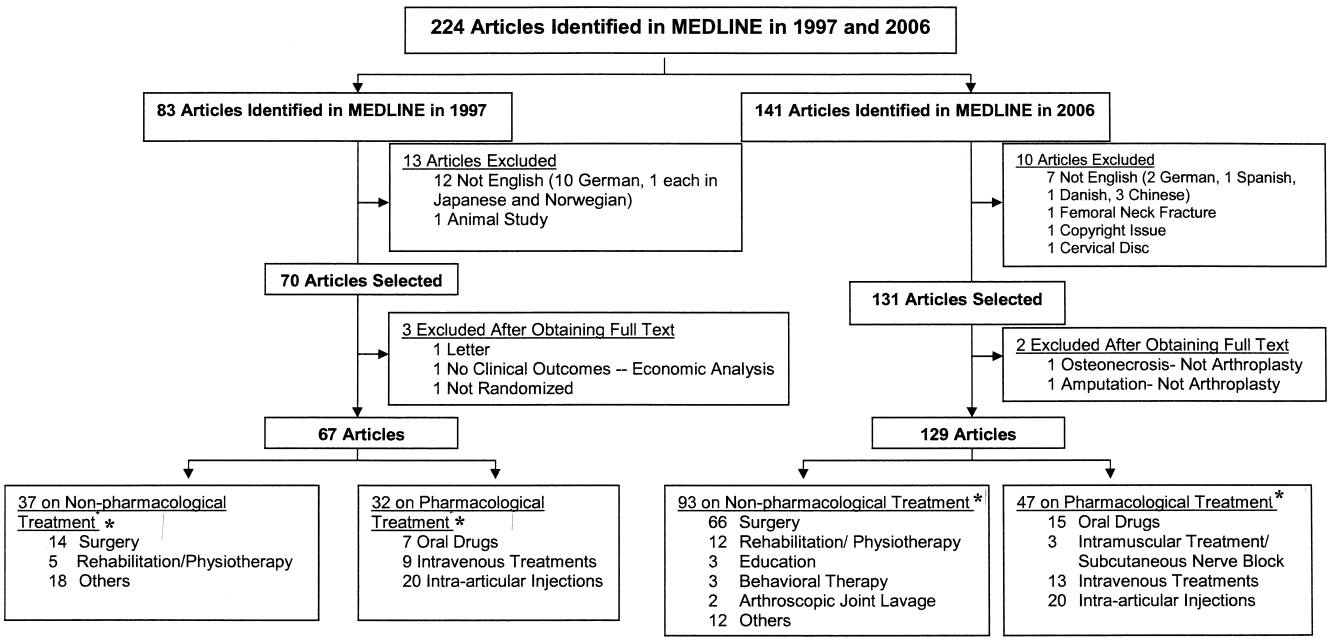

Figure 1. The process of articles selected for review. *Totals add up to more than a simple sum since many studies had multiple types of interventions.

Table 1. Characteristics of studies. Values are N (%). Numbers are rounded to the nearest digit; total may add up to > 100, since many trials had > 1 type of intervention.

	All, n = 196	1997, n = 7	2006, n = 129
Body part			
Upper extremity			
Shoulder	3 (2)	0	3 (2)
Elbow	1 (0.5)	1 (0.5)	0
Hand	3 (2)	1 (0.5)	2 (1)
Lower extremity			
Hip	100 (51)	43 (22)	57 (29)
Knee	99 (51)	32 (16)	67 (34)
Foot	1 (0.5)	0	1 (0.5)
Long bones	3 (2)	0	3 (2)
No. of centers			
Single center	70 (36)	27 (40)	43 (33)
Multicenter	37 (19)	15 (22)	22 (17)
Unclear	89 (45)	25 (37)	63 (49)
No. of patients			
≤ 50	69 (35)	25 (13)	44 (22)
51–100	66 (34)	18 (9)	48 (25)
201–500	16 (8)	7 (4)	9 (5)
> 500	14 (7)	5 (3)	9 (5)
Type of study			
Original	183 (94)	64 (96)	119 (92)
Followup/subgroup	13 (6)	3 (2)	10 (5)
Type of intervention			
Pharmacological	79 (40)	32 (48)	47 (36)
Oral medications	22 (11)	7 (10)	15 (8)
Intramuscular/subcutaneous/ nerve block	3 (1)	0	3 (2)
Intravenous	22 (11)	9 (5)	13 (7)
Intraarticular	40 (20)	20 (10)	20 (10)
Nonpharmacological	130 (66)	37 (19)	93 (47)
Surgery	80 (40)	14 (7)	66 (34)
Arthroscopy/joint lavage	2 (1)	0	2 (1)
Rehabilitation/physiotherapy	17 (9)	5 (3)	12 (6)
Behavioral intervention	3 (2)	0	3 (2)
Education	3 (2)	0	3 (2)
Other*	30 (15)	18 (9)	12 (6)
Funding support			
No information provided	100 (50)	46 (69)	54 (42)
None	21 (11)	3 (5)	18 (14)
Private	62 (31)	18 (27)	44 (34)
Private and public	4 (2)	0	4 (3)
Public	9 (5)	0	9 (7)
Type of journal			
Orthopedics/surgery	125 (64)	38 (57)	87 (67)
Other/rehabilitation	10 (6)	14 (20)	20 (16)
Anesthesia	34 (17)	13 (19)	14 (11)
Internal medicine	27 (14)	2 (3)	9 (6)

* Includes diet, ultrasound, electrical stimulation, drainage, irradiation, navigation system, implants, wound dressings, etc.

assessors (Appendix 2). Studies published in internal medicine journals (Appendix 2) or in journals with higher impact factor (Appendix 3) were significantly more likely to have better reported methodological quality.

Multivariable correlates of overall quality. Multivariable models of overall RCT quality included all variables significant in univariate analyses, namely, type of intervention, number of centers, number of patients, funding source, type of journal, and journal impact factor. We found that compared to pharmacological intervention RCT, nonpharmacological intervention RCT had lower odds of being in the highest tertiles of Jadad, Delphi list, and NRS scores [odds ratio (OR) 0.28–0.39, p = 0.033–0.008] (Table 5). Higher number of centers was significantly associated with higher Delphi list score (OR 4.7, p = 0.017) and lack of funding was significantly associated with lower NRS score (OR 0.28, p = 0.014). Number of patients, journal type, and jour-

Table 2. Association of trial characteristics with overall quality as assessed by Jadad (range, 0–5), Delphi list (0–9), and numeric rating scale (1–10) scores. Values are mean (SEM).

	Jadad Scale	Delphi Scale	NRS Score
All studies	2.36 (1.4)	5.33 (1.57)	4.30 (2.6)
Funding support	p = 0.05	p = 0.01	p = 0.006
No information	2.17 (0.132)	5.07 (0.15)	3.85 (0.237)
None	1.90 (0.308)	4.76 (0.275)	3.33 (0.509)
Private	2.73 (0.178)	5.84 (0.957)	5.10 (0.334)
Private and public	2.75 (0.629)	5.5 (0.957)	5.25 (1.493)
Public	2.78 (0.619)	5.89 (0.716)	5.44 (1.26)
No. of centers	p = 0.005	p < 0.001	p < 0.001
< 2	2.19 (0.153)	4.74 (0.150)	3.80 (0.282)
≥ 2	3.03 (0.264)	6.43 (0.289)	5.76 (0.467)
Not clear	2.21 (0.142)	5.33 (0.157)	4.07 (0.264)
No. of patients	p = 0.003	p < 0.001	p < 0.001
< 50	2.14 (0.164)	5.16 (0.177)	3.87 (0.304)
51–100	2.20 (0.153)	4.95 (0.155)	3.88 (0.277)
101–500	2.64 (0.225)	5.71 (0.267)	4.91 (0.417)
> 500	3.50 (0.344)	6.93 (0.412)	6.71 (0.641)
Year of publication	p = 0.742	p = 0.991	p = 0.708
1997	2.40 (0.166)	5.33 (0.207)	4.19 (0.311)
2006	2.33 (0.126)	5.32 (0.133)	4.34 (0.186)
Type of intervention	p < 0.001	p < 0.001	p < 0.001
Pharmacological	3.02 (0.172)	6.26 (0.218)	5.44 (0.351)
Nonpharmacological	1.89 (0.111)	4.75 (0.104)	3.52 (0.194)
Both	3.23 (0.1)	5.77 (0.469)	5.38 (0.828)
Type of intervention	p < 0.001	p < 0.001	p < 0.001
Oral	2.69 (0.414)	5.62 (0.538)	4.69 (0.95)
Intravenous	3.24 (0.338)	6.59 (0.403)	5.59 (0.67)
Intraarticular	3.13 (0.243)	6.40 (0.294)	5.83 (0.468)
Surgery	1.96 (0.144)	4.82 (0.135)	3.58 (0.249)
Rehabilitation	2.00 (0.378)	4.88 (0.398)	4.00 (0.681)
> 1 intervention	2.87 (0.401)	5.73 (0.527)	4.87 (0.810)
Others	1.90 (0.204)	4.76 (0.192)	3.60 (0.366)
Impact factor	p = 0.01	p = 0.004	p = 0.005
≤ 0.5	2.29 (0.522)	5.43 (0.571)	4.14 (1.01)
> 0.5–1	2.14 (0.257)	4.93 (0.263)	3.86 (0.457)
> 1–2	2.21 (0.145)	5.15 (0.161)	4.06 (0.262)
> 2–5	2.71 (0.197)	5.82 (0.228)	4.98 (0.367)
> 5–10	3.75 (0.479)	7.00 (1.080)	7.25 (1.380)
> 10	4.33 (0.333)	7.33 (0.333)	8.00 (0.577)
Type of journal	p < 0.001	p < 0.001	p < 0.001
Orthopedics/surgery	2.01 (0.117)	4.98 (0.115)	3.71 (0.202)
Other/rehabilitation	2.40 (0.521)	5.40 (0.581)	4.00 (0.989)
Anesthesia	3.06 (0.235)	5.97 (0.311)	5.44 (0.478)
Internal medicine	3.07 (0.238)	6.07 (0.366)	5.63 (0.542)

Table 3. Characteristics of randomized arthroplasty trials by the type of intervention. Values are n (%).

Characteristic	Pharmacological, n = 66	Nonpharmacological, n = 117	Both, n = 13	All Trials, n = 196
Randomization				
Generation of allocation sequence	p = 0.104			
Adequate	26 (39)	51 (44)	8 (62)	85 (43)
Inadequate	1 (2)	12 (10)	0 (0)	13 (7)
Not reported	39 (59)	54 (46)	5 (39)	98 (49)
Allocation concealment	p = 0.403			
Adequate	29 (44)	40 (34)	7 (54)	76 (39)
Inadequate	1 (2)	5 (4)	0 (0)	6 (3)
Not reported	36 (55)	72 (62)	6 (46)	114 (58)
Intervention reproducible	p = 0.430			
Adequate	64 (97)	114 (97)	12 (93)	190 (97)
Inadequate	1 (2)	2 (2)	0 (0)	3 (2)
Not reported	1 (2)	1 (1)	1 (8)	3 (2)
Influence of care provider skill	p = 0.850			
Yes	1 (2)	1 (1)	0 (0)	2 (1)
No	65 (99)	116 (99)	13 (100)	194 (99)
CONSORT diagram reported	p = 0.19			
Adequate	9 (14)	7 (7)	4 (30)	20 (10)
Not reported	57 (86)	110 (93)	9 (69)	176 (90)
Control intervention	p < 0.001			
Placebo	31 (47)	0 (0)	5 (39)	36 (18)
No placebo	35 (53)	107 (100)	8 (61)	150 (82)
Influence of care provider skill	p = 0.850			
Yes	1 (2)	1 (1)	0	2 (1)
No	65 (99)	116 (99)	13 (100)	194 (99)
Potential similar placebo effect	p < 0.001			
Adequate	57 (86)	55 (47)	13 (100)	125 (64)
Inadequate	2 (3)	10 (9)	0 (0)	12 (6)
Not reported*	7 (11)	52 (45)	0 (0)	59 (30)
Blinding				
Patients	p < 0.001			
Adequate	40 (61)	14 (12)	6 (46)	60 (31)
Inadequate	1 (2)	7 (4)	2 (15)	10 (5)
Not reported*	25 (38)	98 (84)	5 (39)	128 (65)
Care providers	p < 0.001			
Adequate	28 (42)	2 (2)	3 (23)	33 (17)
Inadequate	13 (20)	89 (76)	4 (31)	106 (54)
Not reported	25 (38)	26 (22)	6 (46)	57 (29)
Outcome assessors	p = 0.001			
Adequate	42 (64)	39 (33)	8 (62)	89 (45)
Inadequate	1 (2)	12 (10)	0 (0)	13 (7)
Not reported	23 (35)	66 (57)	5 (39)	94 (48)
Statistical analysis	p < 0.001			
Described as ITT or modified ITT	15 (23)	4 (3)	1 (8)	20 (10)
Not described as ITT or modified ITT	51 (77)	113 (97)	12 (92)	176 (90)
Sample size justification/power reported	p = 0.001			
Adequate	30 (46)	31 (27)	9 (69)	70 (36)
Not reported	36 (55)	86 (74)	4 (31)	126 (64)
Were adverse events reported	p = 0.116			
Adequate	46 (70)	58 (50)	7 (54)	111 (57)
Inadequate	20 (30)	1 (1)	0 (0)	21 (11)
Not reported	0 (0)	58 (50)	6 (46)	64 (33)
Lost to followup < 20	p = 0.524			
Adequate	35 (53)	56 (48)	8 (62)	99 (51)
Inadequate	8 (12)	10 (9)	2 (15)	20 (10)
Not reported	23 (35)	51 (44)	3 (23)	77 (39)

* Combination of not reported and not applicable. CONSORT: Consolidated Standards of Reporting Trials[10]; ITT: intention to treat. Percentages in parentheses are rounded; total may not sum to 100%.

Table 4. Characteristics of randomized trials of arthroplasty by number of patients and number of centers. Values are n (%).

Characteristic	No. of Patients				No. of Centers		
	≤ 50, n = 69	51–100, n = 66	101–500, n = 45	> 500, n = 14	≤ 2, n = 70	> 2, n = 37	Not Clear, n = 87
Randomization							
Generation of allocation sequence	p = 0.279				p = 0.873		
Adequate	24 (35)	31 (47)	22 (49)	8 (57)	33 (47)	18 (49)	34 (38)
Inadequate	3 (4)	5 (8)	5 (11)	0 (0)	5 (7)	2 (5)	6 (7)
Not reported	42 (60)	30 (46)	18 (40)	6 (43)	32 (46)	17 (46)	49 (54)
Allocation concealment	p = 0.634				p = 0.003		
Adequate	22 (32)	28 (43)	17 (38)	8 (57)	20 (29)	20 (54)	36 (40)
Inadequate	3 (4)	2 (3)	1 (2)	0 (0)	6 (9)	0 (0)	0 (0)
Not reported	44 (64)	36 (55)	27 (60)	6 (43)	44 (63)	17 (46)	53 (60)
Intervention reproducible	p = 0.449				p = 0.751		
Adequate	66 (96)	65 (99)	44 (98)	13 (93)	67 (96)	36 (97)	87 (98)
Inadequate	3 (4)	1 (2)	1 (2)	1 (7)	3 (4)	1 (3)	2 (2)
CONSORT diagram reported	p < 0.001				p = 0.058		
Adequate	1 (1)	7 (12)	6 (13)	6 (43)	5 (9)	8 (22)	7 (8)
Not reported	68 (99)	59 (89)	39 (87)	8 (57)	63 (91)	29 (78)	82 (92)
Control intervention	p < 0.001				p = 0.003		
Placebo	11 (16)	5 (8)	12 (27)	8 (57)	6 (9)	13 (35)	17 (19)
No placebo	58 (84)	61 (92)	33 (73)	6 (43)	64 (91)	24 (65)	72 (81)
Influence of care provider skill	p = 0.300				p = 0.162		
Yes	2 (3)	0 (0)	0 (0)	0 (0)	2 (3)	0 (0)	0 (0)
No	67 (97)	66 (100)	45 (100)	14 (100)	68 (97)	37 (100)	89 (100)
Potential similar placebo effect	p = 0.733				p = 0.087		
Adequate	42 (61)	38 (58)	32 (71)	12 (86)	36 (51)	30 (81)	59 (66)
Inadequate	5 (7)	5 (8)	2 (4)	0 (0)	7 (10)	0 (0)	5 (6)
Not reported	22 (32)	23 (35)	11 (24)	2 (14)	27 (39)	7 (19)	25 (28)
Blinding							
Patients	p = 0.021				p = 0.003		
Adequate	21 (30)	12 (18)	18 (40)	9 (64)	12 (17)	21 (57)	27 (30)
Inadequate	2 (3)	5 (8)	1 (2)	0 (0)	5 (7)	0 (0)	3 (3)
Not reported*	46 (67)	49 (74)	26 (58)	5 (36)	53 (76)	16 (43)	59 (66)
Care providers	p < 0.001				p = 0.008		
Adequate	11 (16)	4 (6)	11 (24)	7 (50)	6 (8)	13 (35)	14 (15)
Inadequate	43 (62)	41 (62)	19 (42)	1 (7)	45 (64)	15 (41)	46 (52)
Not reported	15 (22)	21 (32)	15 (33)	6 (43)	19 (27)	9 (24)	29 (33)
Outcome assessors	p = 0.031				p < 0.001		
Adequate	31 (45)	22 (33)	25 (56)	11 (79)	23 (33)	28 (76)	38 (43)
Inadequate	4 (6)	7 (11)	1 (2)	0 (0)	8 (11)	0 (0)	5 (6)
Not reported	34 (49)	37 (56)	19 (42)	3 (21)	39 (56)	9 (24)	46 (52)
Were adverse events reported	p = 0.013				p = 0.005		
Adequate	31 (45)	38 (58)	27 (60)	14 (100)	36 (51)	31 (84)	44 (49)
Inadequate	37 (54)	28 (42)	18 (40)	0 (0)	34 (49)	6 (16)	44 (49)
Not reported	1 (5)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (1)
Sample size justification/power reported	p = 0.020				p = 0.087		
Adequate	19 (28)	25 (38)	16 (36)	10 (71)	23 (33)	19 (51)	28 (32)
Not reported	50 (73)	41 (62)	29 (64)	4 (29)	47 (67)	18 (49)	61 (69)
Statistical analysis	p < 0.001				p = 0.001		
Described as ITT or modified ITT	1 (1)	4 (6)	8 (18)	7 (50)	4 (6)	10 (27)	6 (7)
Not described as ITT or modified ITT	68 (99)	62 (94)	37 (82)	7 (50)	66 (94)	27 (73)	83 (93)
Lost to followup < 20	p = 0.039				p = 0.435		
Adequate	36 (52)	33 (50)	24 (53)	6 (43)	33 (47)	20 (54)	46 (52)
Inadequate	2 (3)	6 (9)	8 (18)	4 (29)	8 (11)	6 (16)	6 (7)
Not reported	31 (45)	27 (41)	3 (29)	4 (29)	29 (41)	11 (30)	37 (42)

* Combination of not reported and not applicable. CONSORT: Consolidated Standards of Reporting Trials[10]; ITT: intention to treat. Percentages in parentheses are rounded; total may not sum to 100%

nal impact factor were no longer significantly associated with overall RCT quality in multivariable-adjusted analyses. Sensitivity analyses that adjusted the described multivariable analyses for journal impact factor and number of patients as continuous variables (instead of categorical in the main analyses) did not change these findings.

Table 5. Multivariable adjusted predictors of overall trial quality.

	Jadad Score[a]			Numeric Rating Scale[b]			Delphi List[c]		
	p	OR	95% CI	p	OR	95% CI	p	OR	95% CI
No. of centers (ref: single center)	0.185			0.125			0.017		
Multicenter	0.240	1.84	0.66, 5.12	0.041	3.03	1.04, 8.78	0.004	4.69	1.61, 13.61
Unclear	0.401	0.70	0.30, 1.61	0.354	1.54	0.62, 3.83	0.118	2.04	0.83, 5.01
Type of intervention (ref: pharmacological)	0.033			0.034			0.008		
Nonpharmacological	0.025	0.39	0.17, 0.89	0.011	0.33	0.14, 0.78	0.03	0.28	0.12, 0.64
Both	0.617	1.46	0.33, 6.48	0.743	0.78	0.18, 3.41	0.813	0.84	0.20, 3.59
Type of journal (ref: orthopedics/surgery)	0.069			0.206			0.811		
Anesthesia	0.066	2.69	0.93, 7.72	0.217	1.93	0.68, 5.49	0.483	1.45	0.51, 4.09
Internal medicine[d]	0.045	3.25	1.02, 10.29	0.053	3.11	0.98, 9.79	0.432	1.61	0.49, 5.24
Rehabilitation/other	0.332	0.31	0.03, 3.26	0.387	2.43	0.32, 18.12	0.666	1.62	0.18, 14.39
Source of funding (ref: private)	0.119			0.014			0.082		
Public	0.304	0.39	0.07, 2.33	0.452	0.51	0.09, 2.90	0.568	0.60	0.11, 3.43
Private and public	0.230	4.77	0.37, 61.11	0.714	1.55	0.15, 16.23	0.890	1.18	0.11, 12.29
None/not mentioned	0.071	0.50	0.24, 1.06	0.002	0.28	0.13, 0.62	0.011	0.37	0.17, 0.80
Total no. of patients (ref: ≤ 50)	0.253			0.458			0.337		
51–100	0.762	1.14	0.48, 2.70	0.431	1.46	0.57, 3.71	0.491	0.73	0.29, 1.80
101–500	0.132	2.12	0.80, 5.64	0.111	2.40	0.82, 7.04	0.297	1.72	0.62, 4.78
> 500	0.123	4.36	0.67, 28.32	0.430	1.92	0.38, 9.72	0.442	1.92	0.36, 10.18
Impact factor (ref: ≤ 2)									
> 2	0.808	0.90	0.40, 2.05	0.570	1.27	0.55, 2.93	0.755	1.14	0.45, 2.63
Constant	0.665	1.28		0.423	0.62		0.958	0.97	

[a] Highest tertile, score = 3–5; R^2= 0.297; [b] Highest tertile, score = 6–10; R^2 = 0.327; [c] Highest tertile, score = 6–10; R^2 = 0.334; [d] Internal medicine and subspecialties.

DISCUSSION

What does this report add to the literature? In this first systematic review of a large number of arthroplasty RCT, we found many methodological deficiencies in allocation concealment, blinding, use of placebo, ITT/modified ITT and sample size calculations, with most scores ranging from 20% to 50%, resulting in low overall trial quality. Our multivariable-adjusted analyses suggested that nonpharmacological intervention, lack of funding support, or single-center location were independent predictors of lower overall trial quality.

Limitations of our study. Our study has several limitations. We examined the quality of RCT reporting, not the quality of RCT; it is possible that reporting of methods was inadequate for some RCT that were conducted with more rigor[17]. However, readers have access only to published reports and we suggest that due attention be paid to the reporting of the RCT. Second, certain methodological aspects such as blinding and use of placebo may not be easily amenable to improvement in RCT of nonpharmacological interventions[18]. Many potential areas of improvement exist in conducting and reporting arthroplasty trials of both pharmacological and nonpharmacological interventions, including adequate use of ITT, sample size calculation, allocation sequence generation/concealment, outcome assessor blinding, etc. Third, Jadad score focuses primarily on double-blinding, which may not fairly evaluate the quality of nonpharmacological interventions, as discussed above[18]. We included Delphi list score to avoid this bias since Delphi list awards only 2 of the 9 points for blinding of patients and providers, but scores were low on Delphi list for both pharmacological and nonpharmacological trials. In addition, individual quality standards were still met in < 50% cases, even for pharmacological trials, confirming that the surgical nature of 60% of the arthroplasty trials does not completely explain these deficits in trial quality reporting. Fourth, limiting to English language may limit generalizability; however, < 10% of articles were non-English, so inclusion of these articles is unlikely to have substantially changed our findings or conclusions. Last, due to multiple comparisons, at least 8 statistically significant differences in our study may have been due to chance (total comparisons, about 150). We acknowledge this as a limitation, and thus our findings should be interpreted with some caution until confirmatory studies are available. However, we are fairly confident that differences with p values < 0.001 are unlikely to be due to chance. We found consistent patterns for most of the differences, examining individual quality and overall quality stan-

dards and with sensitivity analyses, and we had stated our hypotheses *a priori*.

Predictors of overall trial quality in multivariable models. An important finding in our study is the observation of significant independent association of nonpharmacological intervention with lower trial quality in multivariable analyses, which confirms and extends the previous similar findings from univariate analyses of osteoarthritis and general RCT[9,19]. Nonpharmacological arthroplasty RCT scored low on most quality standards including use of placebo and blinding of patients (which are less amenable to improvement due to ethical/practical issues). However, these trials also scored low on potential similarity of interventions, sample size justification, and ITT/modified ITT analyses, which are amenable to improvement, as much in surgical as in nonsurgical RCT. Thus, opportunities for improving arthroplasty trial reporting exist, especially for nonpharmacological trials. These improvements should be made in conjunction with reporting additional CONSORT criteria specifically focused on nonpharmacological RCT, as described by Boutron, *et al*[18]. Such improvements in study design and reporting will not only result in better study design for surgical RCT, but will allow for replication of results of RCT in different study populations, making interventions more generalizable.

Neither journal type nor journal impact factor was significantly associated with overall study quality in the multivariable analyses. This implies that RCT quality cannot be inferred by journal type or journal impact factors. Readers should be aware that high impact journals may not necessarily publish high quality studies. Simply relying on impact factor may, in fact, provide a false sense of assurance of study validity. Critically reviewing the methodology of individual trials, regardless of journal or impact factor, remains the most important safeguard.

Our findings further support inclusion of more centers as a significant predictor of RCT quality. This is likely correlated with increasing sample size and reflects that studies with larger sample sizes were also more likely to be multicenter studies. In principle, larger studies are more likely to be coordinated carefully due to increased complexity of their conduct (i.e., multiple investigators). Thus, multicenter trials may, in fact, be a surrogate for study quality. Smaller trials are often single-center, single-investigator studies with limited funding and resultant methodological pitfalls such as insufficient sample size and limited study power (Type II errors or beta errors). Based on our findings, we recommend investigators carefully consider patient-important outcomes and adequately power their studies to have high probability of success. These resultant larger sample sizes will inevitably require multicenter rather than single-center trials. Alternatively, when arthroplasty RCT are being done as single-center RCT, authors should consider examining methods/protocol from multicenter RCT to improve the RCT quality.

Our finding of independent association of presence of funding with better overall trial quality confirms similar univariate associations noted for industry-funded trials[20,21]. In our study this was noted in univariate and confirmed in multivariable analyses. On further analysis, we found that the difference noted was primarily due to better reporting for RCT with private funding compared to those with no/unclear funding. No significant differences were noted between privately and publicly funded RCT. Due to limited funding resources, obtaining funding may be beyond the control of investigators in many circumstances. However, presence of financial support seems to correlate with better quality RCT, likely due to availability of better resources to plan, conduct, analyze, and report RCT.

Variation in study quality in univariate analyses. Lack of significant association of year of publication with RCT quality in univariate analysis disproved one of our hypotheses, that study quality would have improved over time. This observation is similar to that reported for RCT of antibacterial agents[22] and of low back pain[23] over time, but is in contrast to studies of RCT in sepsis[24] and colorectal/laparoscopic resections[25], which showed improvement in quality over time. Arthroplasty RCT published in 2006 reported < 60% for most quality standards, identifying several areas for improvement.

One published study reported weak correlation of 0.21 between impact factor and trial quality of oncology RCT[26]. We found a significant increase in overall trial quality for journals with higher impact factor in univariate, but not in multivariable adjusted analyses. This was most notable for journals with impact factor > 2 and may have been due to more methodological rigor in higher impact journals. Our study confirmed a previous report of a better overall quality score in RCT published in internal medicine/rheumatology journals versus orthopedics/rehabilitation/surgery journals in univariate analysis[9,27,28].

Comparison with previous similar studies. Compared to the earlier study of osteoarthritis RCT[9], we report even lower use of placebo (18% vs 52%, respectively); blinding of patients (31% vs 65%), care providers (17% vs 47%), and outcome assessors (45% vs 85%); use of ITT/modified ITT (20% vs 56%); and sample size justification (36% vs 52%) in arthroplasty RCT. Allocation concealment (39% vs 21%, respectively) was higher, and reproducibility of intervention (97% vs 91%) and allocation sequence generation were similar to osteoarthritis RCT (43% vs 49%). These differences seem to be attributable primarily to higher proportion of RCT of nonpharmacological interventions among arthroplasty RCT (60%) versus osteoarthritis RCT (45%), which had lower quality than pharmacological RCT, both in this and in a previous study[9], and difference in study populations of arthroplasty versus osteoarthritis.

Methodological reviews of RCT from various fields of medicine and surgery have found many deficiencies in their

reporting[9,19,22,29-32]. The low overall quality score we found for arthroplasty RCT is in agreement with previous studies that included RCT in surgical specialties[33,34], as well as reviews in other fields, including headache[35], physical therapy[19], and infectious disease[22]. Our findings of quality deficits (< 50% reporting) in arthroplasty trials are similar to studies of RCT in surgical specialties — review of ophthalmology RCT found < 50% of RCT reported sequence gen-

Appendix 1. Characteristics of arthroplasty trials by the specific intervention type. Values are n (%).

Characteristic	Others, n = 42	Oral, n = 13	Intravenous, n = 17	Intraarticular, n = 30	Surgery, n = 71	Rehabilitation, n = 8	≥ 1 Intervention, n = 15
Randomization							
Generation of allocation sequence	p = 0.786						
Adequate	19 (45)	7 (54)	7 (41)	14 (47)	28 (39)	5 (63)	5 (33)
Inadequate	3 (7)	1 (8)	0 (0)	0 (0)	8 (11)	1 (13)	0 (0)
Not reported	20 (45)	5 (39)	10 (59)	16 (53)	35 (48)	2 (25)	10 (67)
Allocation concealment	p = 0.058						
Adequate	14 (33)	6 (46)	8 (47)	14 (47)	24 (34)	4 (50)	6 (40)
Inadequate	2 (5)	0 (0)	0 (0)	0 (0)	1 (1)	2 (25)	1 (7)
Not reported	26 (62)	7 (54)	9 (53)	16 (53)	46 (65)	2 (25)	8 (53)
Intervention reproducible	p = 0.114						
Adequate	41 (98)	12 (92)	16 (94)	30 (100)	70 (99)	7 (88)	14 (93)
Inadequate	1 (2)	1 (8)	1 (6)	0 (00	1 (1)	1 (13)	1 (7)
CONSORT diagram reported	p = 0.001						
Adequate	3 (7)	3 (23)	3 (18)	2 (7)	2 (3)	3 (38)	4 (33)
Not reported	39 (93)	10 (77)	14 (82)	28 (93)	69 (97)	5 (63)	11 (67)
Control intervention	p < 0.001						
Placebo	1 (2)	6 (46)	8 (47)	14 (47)	0 (0)	0 (0)	7 (47)
No placebo	41 (98)	7 (54)	9 (53)	16 (53)	71 (100)	8 (100)	8 (53)
Influence of care provider skill	p = 0.726						
Yes	1 (2)	0 (0)	0 (0)	1 (3)	0 (0)	0 (0)	0 (0)
No	41 (98)	13 (100)	17 (100)	29 (97)	71 (100)	8 (100)	15 (100)
Potential similar placebo effect	p < 0.001						
Adequate	11 (26)	10 (77)	15 (88)	27 (90)	46 (65)	1 (13)	15 (100)
Inadequate	6 (14)	2 (15)	0 (0)	0 (0)	1 (1)	3 (38)	0 (0)
Not reported*	25 (60)	1 (8)	2 (12)	3 (10)	24 (34)	4 (50)	0 (0)
Blinding							
Patients	p < 0.001						
Adequate	6 (14)	6 (46)	12 (71)	18 (60)	11 (16)	0 (0)	7 (47)
Inadequate	2 (5)	0 (0)	0 (0)	2 (7)	1 (1)	2 (25)	1 (7)
Not reported*	34 (81)	7 (54)	5 (29)	10 (33)	59 (80)	6 (75)	7 (47)
Care providers	p < 0.001						
Adequate	3 (7)	4 (31)	9 (53)	13 (43)	0 (0)	0 (0)	4 (27)
Inadequate	23 (55)	0 (0)	3 (18)	7 (23)	67 (94)	2 (25)	4 (27)
Not reported	16 (38)	9 (69)	5 (29)	10 (33)	4 (6)	6 (75)	7 (47)
Outcome assessors	p < 0.001						
Adequate	11 (26)	4 (31)	12 (71)	22 (73)	27 (38)	4 (50)	9 (60)
Inadequate	8 (19)	0 (0)	0 (0)	1 (3)	2 (3)	2 (25)	0 (0)
Not reported	23 (55)	9 (69)	5 (29)	7 (23)	42 (59)	2 (25)	6 (40)
Were Adverse Events Reported	p = 0.820						
Adequate	22 (52)	10 (77)	10 (59)	20 (67)	35 (49)	4 (50)	10 (67)
Inadequate	20 (48)	3 (23)	7 (41)	10 (33)	35 (49)	4 (50)	5 (33)
Not reported	0 90)	0 (0)	0 (0)	0 (0)	1 (1)	0 (0)	0 (0)
Sample size justification/power reported	p = 0.001						
Adequate	11 (26)	4 (31)	7 (42)	16 (53)	20 (28)	2 (25)	10 (67)
Not reported	31 (74)	9 (69)	10 (59)	14 (47)	51 (72)	6 (75)	5 (33)
Statistical Analysis							
Described as ITT or modified ITT	1 (2)	2 (15)	3 (18)	7 (23)	1 (1)	2 (25)	4 (20)
Not described as ITT or modified ITT	41 (98)	11 (85)	14 (82)	23 (77)	70 (99)	6 (75)	11 (73)
Lost to followup < 20	p = 0.176						
Adequate	17 (41)	7 (54)	9 (53)	17 (57)	37 (52)	5 (63)	7 (47)
Inadequate	1 (5)	2 (15)	1 (6)	3 (10)	7 (10)	2 (25)	4 (27)
Not reported	24 (57)	4 (31)	7 (41)	10 (33)	27 (38)	1 (13)	4 (27)

* Combination of not reported and not applicable. CONSORT: Consolidated Standards of Reporting Trials[10]; ITT: intention to treat.

Appendix 2. Characteristics of arthroplasty trials by journal type and year of publication. Values are n (%).

Characteristic	Journal Type				Publication Year	
	Orthopedics/surgery Journal, n = 125	Anesthesia, n = 34	Internal Medicine, n = 27	Other/Rehab, n = 10	1997, n = 67	2006, n = 129
Randomization						
Generation of allocation sequence	p = 0.146				p = 0.155	
Adequate	49 (39)	21 (62)	11 (41)	4 (40)	23 (34)	62 (48)
Inadequate	13 (10)	0 (0)	0 (0)	0 (0)	3 (5)	10 (8)
Not reported	61 (49)	13 (38)	16 (59)	6 (60)	40 (60)	56 (43)
Allocation concealment	p = 0.586				p = 0.008	
Adequate	42 (34)	17 (50)	13 (48)	4 (40)	16 (24)	60 (47)
Inadequate	4 (3)	1 (3)	1 (4)	0 (0)	2 (3)	4 (3)
Not reported	79 (62)	16 (47)	13 (48)	6 (60)	49 (73)	65 (50)
Intervention reproducible	p = 0.774				p = 0.999	
Adequate	120 (96)	34 (100)	26 (96)	10 (100)	65 (97)	125 (97)
Inadequate	5 (4)	0 (0)	1 (4)	0 (0)	2 (3)	4 (3)
CONSORT diagram reported	p < 0.001				p = 0.008	
Adequate	3 (3)	7 (21)	8 (30)	2 (20)	1 (3)	19 (15)
Not reported	122 (97)	27 (79)	19 (70)	8 (80)	66 (97)	110 (85)
Control intervention	p < 0.001				p = 0.196	
Placebo	11 (9)	12 (35)	10 (37)	3 (30)	15 (22)	21 (16)
No placebo	114 (91)	22 (65)	17 (63)	7 (70)	52 (78)	108 (84)
Influence of care provider skill	p = 0.766				p = 0.432	
Yes	2 (2)	0 (0)	0 (0)	0 (0)	0 (0)	2 (2)
No	123 (98)	34 (100)	27 (100)	10 (100)	67 (100)	127 (98)
Potential similar placebo effect	p = 0.054				p < 0.001	
Adequate	75 (60)	28 (82)	18 (67)	4 (40)	30 (45)	95 (74)
Inadequate	8 (6)	2 (6)	1 (4)	1 (10)	6 (9)	6 (5)
Not reported*	42 (34)	4 (12)	8 (30)	5 (50)	31 (46)	28 (22)
Blinding						
Patients	p = 0.008				p = 0.769	
Adequate	26 (21)	17 (50)	14 (52)	3 (30)	23 (34)	37 (29)
Inadequate	4 (3)	3 (9)	1 (4)	0 (0)	2 (3)	6 (5)
Not reported*	95 (76)	14 (41)	12 (44)	7 (70)	42 (63)	86 (67)
Care providers	p < 0.001				p = 0.017	
Adequate	9 (7)	13 (38)	9 (33)	2 (20)	16 (24)	17 (13)
Inadequate	84 (67)	15 (44)	5 (19)	2 (20)	27 (40)	79 (61)
Not reported	32 (26)	6 (18)	13 (48)	6 (60)	24 (36)	33 (26)
Outcome assessors	p = 0.168				p < 0.001	
Adequate	50 (40)	18 (53)	17 (63)	4 (40)	32 (48)	57 (44)
Inadequate	8 (6)	4 (12)	0 (0)	1 (10)	4 (6)	9 (7)
Not reported	67 (54)	12 (35)	10 (37)	5 (50)	31 (46)	63 (49)
Were Adverse Events Reported	p = 0.980				p = 0.487	
Adequate	70 (56)	21 (62)	15 (56)	5 (50)	35 (52)	76 (59)
Inadequate	54 (43)	13 (38)	12 (44)	5 (50)	32 (48)	52 (40)
Not reported	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	1 (1)
Sample size justification/power reported	p = 0.110				p = 0.013	
Adequate	37 (30)	17 (50)	12 (44)	4 (40)	16 (24)	54 (42)
Not reported	88 (70)	17 (50)	15 (56)	6 (60)	51 (76)	75 (58)
Statistical Analysis	p = 0.061				p = 0.935	
Described as ITT or modified ITT	8 (6)	4 (12)	6 (22)	2 (20)	7 (10)	13 (10)
Not described as ITT or modified ITT	117 (94)	30 (88)	21 (78)	8 (80)	60 (90)	116 (90)
Lost to followup < 20	p = 0.129				p = 0.435	
Adequate	63 (50)	19 (56)	13 (48)	4 (40)	30 (45)	69 (54)
Inadequate	10 (8)	2 (6)	7 (26)	1 (10)	4 (6)	16 (12)
Not reported	52 (42)	13 (38)	7 (26)	5 (50)	33 (49)	44 (34)

* Combination of not reported and not applicable. CONSORT: Consolidated Standards of Reporting Trials[10]; ITT: intention to treat.

eration, randomization restriction, allocation concealment, allocation implementation, patient flow diagrams, and sample size calculation[33]. Less than one-third of obstetrics and gynecology RCT reported allocation sequence generation or allocation concealment[34].

On the other hand, quality standards such as description

Appendix 3. Characteristics of arthroplasty randomized trials by journal impact factor. Values are n (%).

Characteristic	Impact Factor ≤ 0.5, n = 7	> 0.5–1, n = 29	> 1–2, n = 81	> 2–5, n = 55	> 5–10, n = 4	> 10, n = 3
Randomization						
Generation of allocation sequence	p = 0.906					
Adequate	1 (14)	13 (45)	38 (47)	25 (46)	3 (75)	1 (33)
Inadequate	1 (14)	2 (7)	6 (7)	2 (4)	0 (0)	0 (0)
Not reported	5 (71)	14 (47)	37 (46)	28 (51)	1 (25)	2 (67)
Allocation concealment	p = 0.010					
Adequate	0 (0)	6 (21)	35 (43)	27 (49)	2 (50)	2 (67)
Inadequate	0 (0)	1 (3)	1 (1)	1 (2)	1 (25)	0 (0)
Not reported	7 (100)	22 (76)	45 (56)	27 (49)	1 (25)	1 (33)
Intervention reproducible	p = 0.014					
Adequate	7 (100)	28 (97)	77 (95)	55 (100)	4 (100)	2 (67)
Inadequate	0 (0)	1 (3)	4 (5)	0 (0)	0 (0)	1 (33)
CONSORT diagram reported	p < 0.001					
Adequate	0 (0)	1 (3)	5 (9)	7 (13)	4 (100)	1 (33)
Not reported	7 (100)	28 (97)	76 (91)	48 (87)	0 (0)	2 (67)
Control intervention	p = 0.008					
Placebo	2 (29)	5 (17)	10 (12)	13 (24)	3 (75)	2 (67)
No placebo	5 (71)	24 (83)	71 (88)	42 (76)	1 (25)	1 (33)
Influence of care provider skill	p = 0.811					
Yes	0 (0)	0 (0)	0 (0)	1 (2)	0 (0)	0 (0)
No	7 (100)	29 (100)	81 (100)	54 (98)	4 (100)	3 (100)
Potential similar placebo effect	p = 0.048					
Adequate	3 (43)	11 (38)	57 (70)	41 (75)	3 (75)	2 (67)
Inadequate	1 (14)	1 (3)	4 (5)	3 (6)	1 (25)	0 (0)
Not reported*	3 (43)	17 (59)	20 (25)	11 (20)	0 (0)	1 (33)
Blinding						
Patients	p = 0.227					
Adequate	3 (43)	6 (21)	20 (25)	23 (42)	3 (75)	3 (100)
Inadequate	0 (0)	1 (3)	4 (5)	2 (4)	0 (0)	0 (0)
Not reported*	4 (57)	22 (76)	57 (70)	30 (55)	1 (25)	0 (0)
Care providers	p = 0.070					
Adequate	2 (29)	5 (17)	8 (10)	13 (24)	2 (50)	2 (67)
Inadequate	3 (43)	18 (62)	48 (59)	29 (53)	0 (0)	0 (0)
Not reported	2 (29)	6 (21)	25 (31)	13 (24)	2 (50)	1 (33)
Outcome assessors	p = 0.277					
Adequate	4 (57)	11 (38)	32 (40)	31 (56)	3 (75)	3 (100)
Inadequate	0 (0)	4 (14)	5 (6)	3 (6)	0 (0)	0 (0)
Not reported	3 (43)	14 (48)	44 (54)	21 (38)	1 (25)	0 (0)
Were Adverse Events Reported	p = 0.200					
Adequate	6 (86)	10 (35)	46 (57)	35 (64)	3 (75)	3 (100)
Inadequate	0 (0)	0 (0)	1 (1)	0 (0)	0 (0)	0 (0)
Not reported	1 (14)	19 (66)	34 (42)	20 (36)	1 (25)	0 (0)
Sample size justification/power reported	p = 0.007					
Adequate	2 (29)	6 (21)	25 (31)	27 (49)	3 (75)	3 (100)
Not reported	5 (71)	23 (79)	56 (69)	28 (51)	1 (25)	0 (0)
Statistical Analysis	p = 0.007					
Described as ITT or modified ITT	2 (29)	0 (0)	6 (7)	9 (16)	2 (50)	1 (33)
Not described as ITT or modified ITT	5 (71)	29 (100)	75 (93)	46 (84)	2 (50)	2 (67)
Lost to followup < 20	p = 0.326					
Adequate	4 (57)	15 (52)	42 (52)	31 (56)	1 (25)	1 (33)
Inadequate	0 (0)	1 (3)	9 (11)	7 (13)	2 (50)	1 (33)
Not reported	3 (43)	13 (45)	30 (37)	17 (31)	1 (25)	1 (33)

* Combination of not reported and not applicable. CONSORT: Consolidated Standards of Reporting Trials[10]; ITT: intention to treat.

of intervention to be reproducible and use of placebo with similar potential effect as the intervention were described in the majority of most arthroplasty trials (64%–97%). Learning curve, standardization and reproducibility of the

procedure, center's volume of care, and care provider expertise[36] are specific important methodological issues for surgical trials[18], and therefore this finding is reassuring.

Blinding of patients and/or surgeons is a challenge in surgical RCT[37,38]. One study found that only 33% of surgery RCT were blinded[39]. Another study found that it was impossible to blind in 72% of the orthopedics RCT[40]. The same study also reported that in 16%, 50%, and 50% of RCT where blinding was possible for providers, patients, and assessors, respectively, the RCT did not blind or did not describe blinding[40]. This implies that even for orthopedics RCT that have challenges with regards to blinding of patients and in some cases providers, blinding is still possible in the majority and should be done when possible. The most room for improvement exists in blinding outcome assessors, which is possible in most instances. It is also important to ensure assessors are independent of the surgeons/providers. This alone has a huge potential in reducing observer bias and improving the RCT quality.

In summary, we found methodological deficiencies in several areas of reporting of arthroplasty RCT. Overall trial quality is associated with trial and intervention characteristics. We have identified many areas of improvement for conduct and reporting of arthroplasty RCT.

ACKNOWLEDGMENT

We thank Indy Rutks from the Cochrane Library for performing the literature search, and Ruth Brady (research associate), Pearlita Ochoa (administrative assistant), and Amy Anderson for their administrative help.

REFERENCES

1. Centers for Disease Control and Prevention. Prevalence of self-reported arthritis or chronic joint symptoms among adults — United States, 2001. MMWR Morb Mortal Wkly Rep 2002;51:948-50.
2. Centers for Disease Control and Prevention. Impact of arthritis and other rheumatic conditions on the health-care system — United States, 1997. MMWR Morb Mortal Wkly Rep 1999;48:349-53.
3. Yelin E, Callahan LF. The economic cost and social and psychological impact of musculoskeletal conditions. National Arthritis Data Work Groups. Arthritis Rheum 1995;38:1351-62.
4. Deyo RA, Inui TS, Leininger J, Overman S. Physical and psychosocial function in rheumatoid arthritis. Clinical use of a self-administered health status instrument. Arch Intern Med 1982;142:879-82.
5. Achterberg-Lawlis J. The psychological dimensions of arthritis. J Consult Clin Psychol 1982;50:984-92.
6. Centers for Disease Control and Prevention. Prevalence of disabilities and associated health conditions among adults in United States, 1999. MMWR Morb Mortal Wkly Rep 2001;50:120-5.
7. Kurtz S, Ong K, Lau E, Mowat F, Halpern M. Projections of primary and revision hip and knee arthroplasty in the United States from 2005 to 2030. J Bone Joint Surg Am 2007;89:780-5.
8. Learmonth ID, Young C, Rorabeck C. The operation of the century: Total hip replacement. Lancet 2007;370:1508-19.
9. Boutron I, Tubach F, Giraudeau B, Ravaud P. Methodological differences in clinical trials evaluating nonpharmacological and pharmacological treatments of hip and knee osteoarthritis. JAMA 2003;290:1062-70.
10. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. Ann Intern Med 2001;134:663-94.
11. Schulz KF, Grimes DA. Blinding in randomised trials: Hiding who got what. Lancet 2002;359:696-700.
12. Hill CL, Buchbinder R, Osborne R. Quality of reporting of randomized clinical trials in abstracts of the 2005 Annual Meeting of the American College of Rheumatology. J Rheumatol 2007;34:2476-80.
13. Jadad AR, Cook DJ, Jones A, Klassen TP, Tugwell P, Moher M, et al. Methodology and reports of systematic reviews and meta-analyses: A comparison of Cochrane Reviews with articles published in paper-based journals. JAMA 1998;280:278-80.
14. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Control Clin Trials 1996;17:1-12.
15. Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, Bouter LM, et al. The delphi list: A criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by delphi consensus. J Clin Epidemiol 1998;51:1235-41.
16. Guyatt GH, Veldhuyzen Van Zanten SJ, Feeny DH, Patrick DL. Measuring quality of life in clinical trials: A taxonomy and review. CMAJ 1989;140:1441-8.
17. Hill CL, LaValley MP, Felson DT. Discrepancy between published report and actual conduct of randomized clinical trials. J Clin Epidemiol 2002;55:783-6.
18. Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P. Extending the consort statement to randomized trials of nonpharmacologic treatment: Explanation and elaboration. Ann Intern Med 2008;148:295-309.
19. Foley NC, Bhogal SK, Teasell RW, Bureau Y, Speechley MR. Estimates of quality and reliability with the physiotherapy evidence-based database scale to assess the methodology of randomized controlled trials of pharmacological and nonpharmacological interventions. Phys Ther 2006;86:817-24.
20. Kjaergard LL, Gluud C. Funding, disease area, and internal validity of hepatobiliary randomized clinical trials. Am J Gastroenterol 2002;97:2708-13.
21. Kjaergard LL, Nikolova D, Gluud C. Randomized clinical trials in hepatology: Predictors of quality. Hepatology 1999;30:1134-8.
22. Falagas ME, Pitsouni EI, Bliziotis IA. Trends in the methodological quality of published randomized controlled trials on antibacterial agents. Br J Clin Pharmacol 2008;65:942-54.
23. Koes BW, Bouter LM, van der Heijden GJ. Methodological quality of randomized clinical trials on treatment efficacy in low back pain. Spine 1995;20:228-35.
24. Graf J, Doig GS, Cook DJ, Vincent JL, Sibbald WJ. Randomized, controlled clinical trials in sepsis: Has methodological quality improved over time? Crit Care Med 2002;30:461-72.
25. Schwenk W, Haase O, Gunther N, Neudecker J. Methodological quality of randomised controlled trials comparing short-term results of laparoscopic and conventional colorectal resection. Int J Colorectal Dis 2007;22:1369-76.
26. Berghmans T, Meert AP, Mascaux C, Paesmans M, Lafitte JJ, Sculier JP. Citation indexes do not reflect methodological quality in lung cancer randomised trials. Ann Oncol 2003;14:715-21.
27. Schulz KF, Grimes DA, Altman DG, Hayes RJ. Blinding and exclusions after allocation in randomised controlled trials: Survey of published parallel group trials in obstetrics and gynaecology. BMJ 1996;312:742-4.
28. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ 1999;319:670-4.

29. Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002;287:2973-82.
30. Boutron I, Estellat C, Guittet L, Dechartres A, Sackett DL, Hróbjartsson A, et al. Methods of blinding in reports of randomized controlled trials assessing pharmacologic treatments: A systematic review. PLoS Med 2006;3:e425.
31. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med 2001;135:982-9.
32. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998;352:609-13.
33. Lai TY, Wong VW, Lam RF, Cheng AC, Lam DS, Leung GM. Quality of reporting of key methodological items of randomized controlled trials in clinical ophthalmic journals. Ophthal Epidemiol 2007;14:390-8.
34. Schulz KF, Chalmers I, Altman DG, Grimes DA, Dore CJ. The methodologic quality of randomization as assessed from reports of trials in specialist and general medical journals. Online J Curr Clin Trials 1995;Doc. No. 197:[81 paragraphs].
35. Fernandez-de-las-Penas C, Alonso-Blanco C, San-Roman J, Miangolarra-Page JC. Methodological quality of randomized controlled trials of spinal manipulation and mobilization in tension-type headache, migraine, and cervicogenic headache. J Orthop Sports Phys Ther 2006;36:160-9.
36. Halm EA, Lee C, Chassin MR. Is volume related to outcome in health care? A systematic review and methodologic critique of the literature. Ann Intern Med 2002;137:511-20.
37. Boutron I, Guittet L, Estellat C, Moher D, Hrobjartsson A, Ravaud P. Reporting methods of blinding in randomized trials assessing nonpharmacological treatments. PLoS Med 2007;4:e61.
38. McCulloch P, Taylor I, Sasako M, Lovett B, Griffin D. Randomised trials in surgery: Problems and possible solutions. BMJ 2002;324:1448-51.
39. Solomon MJ, Laxamana A, Devore L, McLeod RS. Randomized controlled trials in surgery. Surgery 1994;115:707-12.
40. Poolman RW, Struijs PA, Krips R, Sierevelt IN, Marti RK, Farrokhyar F, et al. Reporting of outcomes in orthopaedic randomized trials: Does blinding of outcome assessors matter? J Bone Joint Surg Am 2007;89:550-8.