The Prevalence of Underpowered Randomized Clinical Trials in Rheumatology

HELEN I. KEEN, KEVIN PILE, and CATHERINE L. HILL

ABSTRACT. ***Objective.*** The conduct of underpowered randomized controlled trials (RCT) has recently been criticized in medical journals. We investigated the current prevalence of underpowered RCT in rheumatology.

Methods. We searched to identify randomized, prospective RCT assessing clinical efficacy of treatments for adult rheumatic diseases published in English in 2001 and 2002. RCT were assessed as positive or negative based on the result of the primary outcome measure. For phase III RCT with negative results without power analysis, we calculated adequate sample size using beta = 0.20 and alpha = 0.05. We also examined trial quality by assessing the adequacy of reported random sequence generation, allocation concealment, and analysis, and compared the quality of reporting of RCT with adequate and inadequate sample size.

Results. A total of 228 RCT met inclusion criteria; of the 205 phase III trials, 119 were positive, 81 were negative. The remaining 5 trials made no statistical comparison between interventions, and did not supply enough information for a result to be calculated. Of the 86 negative or indeterminate RCT, 37 reported sample size calculations (all but 4 had adequate power). Of the 49 remaining phase III trials that did not report power calculations, we conducted sample size calculations; only 10 were adequately powered. Few of the underpowered RCT studied rare rheumatic diseases. Negative RCT with inadequate sample size were less likely to describe adequate random sequence generation or allocation concealment than positive RCT or negative RCT with adequate sample size.

Conclusion. The conduct of underpowered trials is not an infrequent occurrence in rheumatology, with only 50% of negative or indeterminate phase III rheumatology RCT in 2001-2002 having adequate sample size. (J Rheumatol 2005;32:2083-8)

Key Indexing Terms:
SAMPLE SIZE CALCULATION RANDOMIZED CLINICAL TRIAL RHEUMATOLOGY

The conduct of underpowered randomized controlled trials (RCT) has been widely debated in medical literature over many years[1-4]. Power analyses are done to determine the study sample size required to ensure a reasonable probability that a study will reject a null hypothesis when it is false (i.e., avoid a type II or beta error). Studies are arbitrarily accepted to be adequately powered when there is an 80% probability the study would show a treatment effect if it were present. Underpowered RCT have been criticized as unethical because such studies may not adequately test the underlying hypothesis they were designed to test[1]. In order for these underpowered studies to be conducted, resources are utilized and participants exposed to potential risks of research. In addition, these studies may wrongly conclude that the studied treatment is inefficacious because of the potential for type II error.

We investigated the prevalence of underpowered RCT in rheumatology in the period 2001-2002, and assessed if there were any differences in the methodological quality of underpowered RCT compared to adequately powered RCT in rheumatology in this period.

MATERIALS AND METHODS

Search strategy. We sought RCT assessing clinical efficacy of treatments for adult rheumatic diseases published in English in 2001 and 2002. These included RCT on osteoarthritis (OA), rheumatoid arthritis (RA), fibromyalgia (FM), systemic lupus erythematosus (SLE), systemic sclerosis, Raynaud's phenomenon, Sjögren's syndrome, vasculitis, Behçet's disease, gout, pseudogout, ankylosing spondylitis, spondyloarthropathy, psoriatic arthritis, myositis, adhesive capsulitis, reactive arthritis, and arthritis. We excluded RCT that evaluated back pain, reflex sympathetic dystrophy, soft tissue rheumatism, tendinitis, and bursitis. We also excluded RCT that evaluated only adverse effects.

The RCT were found using a Medline search that incorporated MeSH terms for the diseases outlined, the English language, and the specified years of publication. In addition, Pre-Medline was also searched. A trial was deemed a RCT if the terms "randomized," "randomization," or "randomly" appeared in the title, abstract, or Methods section. Only trials that were described as being prospective clinical trials, with a parallel or crossover design in their Methods section, were included for analysis.

One reviewer (HIK) reviewed trial abstracts, and RCT were identified as those that did not meet inclusion criteria, and those that required reviewing of the entire report. Two hundred ninety-three reports were obtained

From the Rheumatology Unit, The Queen Elizabeth Hospital, Adelaide, Australia.

H.I. Keen, MBBS; K. Pile, MB, BCh, MD; C.L. Hill, MBBS, MSc.

Address reprint requests to C.L. Hill, Rheumatology Unit, The Queen Elizabeth Hospital, 28 Woodville Road, Woodville 5011, Australia. E-mail: Catherine.Hill@nwahs.sa.gov.au

Accepted for publication June 6, 2005.

and reviewed for eligibility. A second reviewer (CLH) assessed all 293 trials to ensure inclusion and exclusion criteria were appropriately addressed.

Evaluation for RCT quality. Each RCT that met inclusion criteria was assessed for quality using a modified version of the Jadad scale[5], modified to include more detailed information regarding the methods of allocation concealment and analysis (Table 1). In addition, the authors and acknowledgments sections were screened for information regarding input from a center of statistical or biostatistical expertise. All trials were reviewed by one reviewer (HIK) with respect to reporting of power calculations, the timing of the calculation, allocation concealment, blinding, randomization, and type of analysis. It was calculated that a random sample of 74 RCT would be required, aiming at an anticipated kappa of 0.6, with 95% confidence intervals of 0.4–0.8 with alpha of 0.05[6]. A sample of 74 RCT from both time periods was evaluated by a second reviewer (CLH) to determine interobserver reliability for random sequence generation, allocation concealment, double blinding, power analysis, and intention-to-treat analysis (kappa 0.80 for all features combined; 95% CI 0.73, 0.87). Almost all the discrepancies related to analysis type actually used. This is often difficult to interpret due to poor reporting by authors.

Data extraction. Demographic data were extracted, including disease, country of first author, type of intervention, design of prospective trial (crossover vs parallel), number of participants, and length of followup. The level of industry support was also recorded according to categories defined in a study by Rochon, *et al*[7].

Analysis. We divided reports into pilot, phase I, phase II, and phase III studies based on the description in the published RCT report. As only phase III RCT need to be powered for efficacy, we included these only in subsequent power analyses. Phase III RCT were deemed to be positive or negative according to their primary outcome measure. Where the primary outcome measure was not stated, the first measure of efficacy mentioned in the Methods section was deemed to be the primary outcome measure for the purpose of our analysis. For RCT involving more than one dose of drug therapy, we assumed the highest dose was the variable around which the trial was designed for the purpose of study outcome and sample size calculations.

Trials were deemed to be positive or negative based on the statistical difference between the intervention groups of the designated outcome measure. If no statistical comparison was made between the interventions in the trial report, we calculated the result based on information given. If this could not be determined (for example, if no variability data were given for the designated outcome), trials were labelled "indeterminate" (n = 5).

For all negative phase III trials where no power calculation had been described in the report, sample size calculations were performed using beta = 0.20 and alpha = 0.05. For trials in which inadequate variability data were provided, we were unable to perform sample size calculations due to missing data (n = 12). We assumed a clinically meaningful treatment effect to be 20% for the purposes of calculating adequate sample size, unless we were able to find a published, validated, clinically meaningful effect size for the primary outcome measure. This was only the case where a visual analog scale was used to assess pain in OA, where a difference of 20 mm was used[8].

Categorical data were analyzed using chi-square tests (or Fisher's test when numbers were small). Continuous measures were analyzed using t tests or Wilcoxon tests for nonparametric data. P values reported are 2-sided. Analyses were undertaken comparing the positive RCT, negative RCT with adequate sample size, and negative RCT with inadequate sample size. We also examined the relationship between adequacy of sample size and journal impact factor. The citation index for each journal in which a RCT in the study was published was determined from the 2001 Science Citation Index. RCT from journals without a citation index were excluded from this analysis. RCT were considered to be from "high impact" journals if the citation index was above the median of the journals included in the study. The remainder were considered low impact.

RESULTS

Of the 294 reports reviewed, 228 were included in the study, 111 published in 2001 and 117 published in 2002. The trial characteristics are presented in Table 2. The number of prospective RCT evaluating treatment efficacy in rheumatology remain relatively constant over the 2 years, with a similar proportion being devoted to each disease type. The most commonly studied disease was OA, accounting for 39.6% of trials.

Of the 205 phase III trials, 180 (87.8%) made direct sta-

Table 1. Evaluation of trials, modified from Jadad, *et al*[5].

Characteristic/Quality	Description
Randomization sequence generation	
Adequate	Random number table, computer random number generation, coin tossing, shuffling cards, adaptive randomization
Inadequate	Case record number, alternation, date of admission, date of birth, even/odd, minimization
Unclear/unreported	
Allocation concealment	
Adequate	Central allocation, local pharmacy allocation, numbered or coded bottles, sealed, opaque envelopes
Unclear/unreported	
Double blinding	
Adequate	Use of placebo, dummies, sham treatment
Inadequate	No use of placebo, double dummies, or sham treatment
Analysis	
Intention-to-treat (ITT)	All participants randomized included in analysis
Modified ITT	Analysis excludes those never treated or evaluated on therapy
Modified modified ITT	Analysis excludes some that drop out after treatment or analysis has begun, but includes not only those that complete the trial protocol
Completers	Includes only those who complete trial protocol
Unclear/not done	

Table 2. Characteristics of trials.

Type of Trial	2001, n = 111	2002, n = 117	Combined, n = 228 (%)
Pilot	5	3	8 (3.5)
Phase I	0	2	2 (0.9)
Phase II	4	9	13 (5.7)
Phase III	102	103	205 (89.9)
Disease type			
RA	27	31	58 (25.4)
OA	46	44	90 (39.6)
FM	6	12	18 (7.9)
CTD	16	18	34 (14.9)
Other	16	12	28 (12.3)
Drug RCT	78	74	152 (66.7)
NSAID	17	15	32/152 (21.1)
Published in			
Rheumatology journals	63	60	123 (54.4)
Manufacturer support*			
No support	69	69	158 (60.5)
Grant	22	28	50 (22.0)
Industry employee listed as author	18	16	34 (15.0)
Drug supplied	12	9	21 (9.2)
Journal supplement sponsored by a company	0	0	0 (0)

* Categories defined by Rochon, *et al*[7]. RA: rheumatoid arthritis, OA: osteoarthritis, FM: fibromyalgia, CTD: connective tissue disease, NSAID: nonsteroidal antiinflammatory drug.

tistical comparison between treatment groups; 108 (60.0%) were positive and 72 (40.0%) negative according to our criteria. For the remaining 25 phase III RCT, no statistical comparison was made between the 2 interventions studied. In most of these trials, a within-group statistical comparison was made, but not a between-group comparison. A result could be calculated for all but 5 of the remaining 25 RCT; 11 were positive and 9 negative. Therefore we determined that of the 205 phase III RCT, 119 (58.0%) were positive, 81 (39.5%) were negative, and 5 (2.4%) were indeterminate (Figure 1). Of 119 positive trials, only 50 (42.0%) reported power calculations. Of the 86 negative or indeterminate studies, 37 (43.0%) reported power calculations (6 of which were calculated *post hoc*). Only 4 RCT reported having an inadequate sample size. There was no significant difference between the number of positive and negative trials that reported power calculations (p = 0.96).

For the remaining 49 negative or indeterminate trials that did not report power calculations, we determined that 10 had adequate sample size, 27 had inadequate sample size, and 12 did not provide variability data required to enable trial power to be calculated. Therefore, of the 86 negative or indeterminate trials only 43 (50%) could be found to have adequate sample size for the trial to have 80% probability of showing a treatment effect if one was truly present.

The diseases studied in the 43 negative, underpowered studies were OA (16 RCT), RA (10), FM (3), systemic sclerosis/Raynaud's phenomenon (3), SLE (5), frozen shoulder (3), and one each of ankylosing spondylitis, seronegative arthropathy and myositis. Therefore, few trials studied rare diseases.

The 205 phase III RCT were sourced from 88 journals, of which 59 were recorded in the 2001 Science Citation Index. There was no difference between RCT published in high and low impact journals with regard to adequacy of sample size (Table 3). Even in high impact journals, 45.8% of negative RCT were underpowered. However, in journals without an impact factor, 61.5% of negative RCT were underpowered.

Table 4 describes the methodological quality of positive and negative RCT. Positive RCT and negative RCT with adequate sample size were not more likely to report adequate random sequence generation than negative RCT with inadequate sample size or indeterminate RCT. Only 40.3% of positive RCT gave adequate descriptions of random sequence generation. Reporting of adequate allocation concealment was poor among all 3 groups of RCT, ranging from 22.9% to 50.6% (Table 3). Positive RCT were also more likely to analyze results according to intention-to-treat or modified intention-to-treat methods, but only around half of the positive RCT used these optimal statistical techniques. There was no methodological difference between negative and indeterminate trials that had adequate or inadequate sample size; however, numbers were small. When we compared the quality of RCT with adequate power (i.e., positive RCT and negative RCT with adequate sample size) and those with inadequate sample size or indeterminate, the only significant difference was that adequately powered RCT

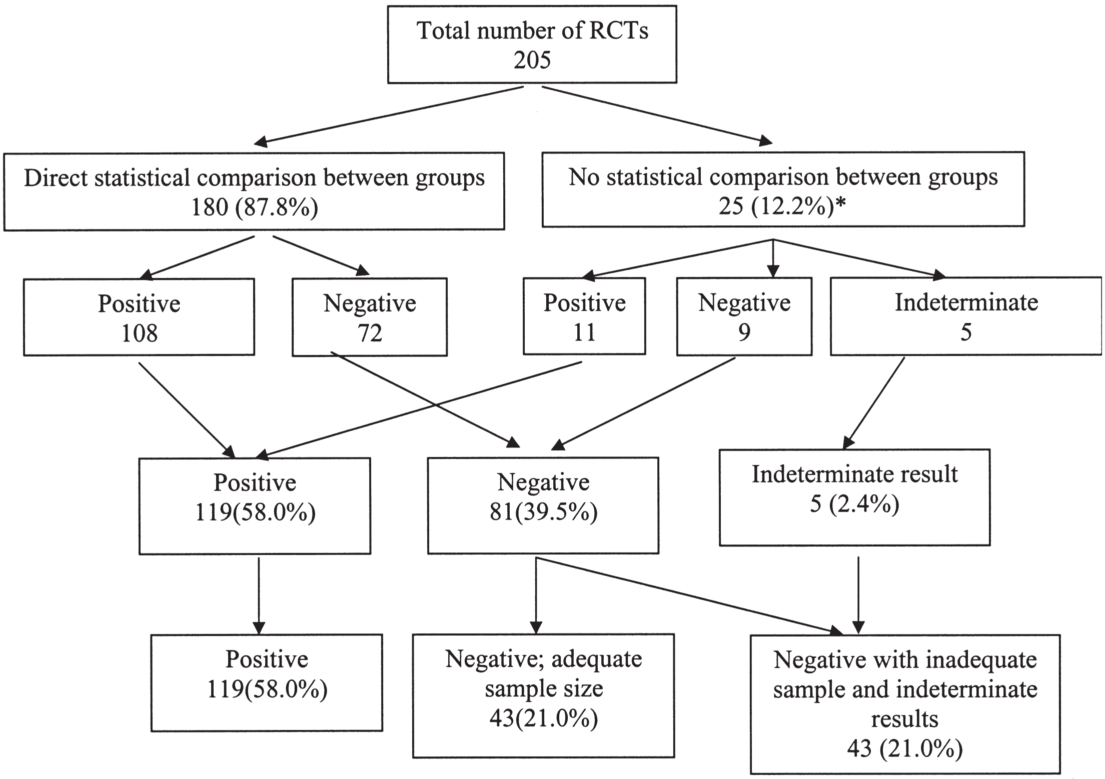

Figure 1. Randomized controlled trials (RCT) reviewed in this study. *Statistical result calculated by authors from data given in published RCT.

were more likely to use adequate random sequence generation (p = 0.02). There was no difference in use of adequate allocation concealment (p = 0.10) or use of intention-to-treat (p = 0.06).

DISCUSSION

We found that 20.9% of phase III randomized controlled trials of adult rheumatic diseases published in English in 2001 and 2002 were underpowered, accounting for 50% of RCT

Table 3. Relationship between journal impact factor and trial power for phase III trials.

	High Impact Journal, n = 120	Low Impact Journal, n = 47	No Impact Rating, n = 38
Positive RCT	72	22	25
Negative RCT: adequate sample size	26	12	5
Negative and indeterminate RCT: inadequate sample size	22	13	8

3 × 2 chi-square comparing high versus low impact journals (p = 0.26).

Table 4. Relationship between trial quality and power for phase III trials.

	Positive RCT, n = 119 (%)	Negative RCT, Adequate Sample Size, n = 43 (%)	Negative and Indeterminate RCT, Inadequate Sample Size, n = 43 (%)
Adequate random sequence generation*	48 (40)	17 (39.5)	9 (20.9)
Adequate concealment of allocation schedule**	40 (33.6)	11 (25.6)	8 (18.6)
Intention-to-treat/modified intention-to-treat analysis***	66 (55.5)	17 (39.5)	15 (34.9)

3 × 2 chi-square: * p = 0.06, ** p = 0.15, *** p = 0.03.

with negative results. These trials may not have been of sufficient sample size to detect clinically meaningful treatment effects.

Arguments to support the continuing conduct of underpowered trials include the premise that small or underpowered clinical trials can provide information that may be pooled in metaanalyses to provide more scientifically meaningful information[1,4,9]. This is pertinent to uncommon diseases, where recruitment of large numbers of participants may be unrealistic. To justify this argument, these trials clearly need to be designed with a view to compilation of the information gained into a metaanalysis[4,9]. There is also the additional problem of publication bias, whereby negative RCT are less likely to be published and therefore are usually unavailable for use in metaanalyses[1]. The majority of the negative underpowered studies we assessed were of RA and OA. No authors reported that their study was planned with a view to collaboration for a metaanalysis. The other situation in which underpowered RCT may be justified is in early-phase trials. However, even in this situation, Halpern and colleagues argue that these should still be powered for other purposes, such as adverse effects[1]. We included only phase III RCT.

When determining the sample size of the negative RCT, we assumed that the RCT were not aiming for equivalence. Indeed, none of the RCT explicitly stated that its aim was to establish equivalence of the 2 interventions, although in some cases this was likely to be the case, for example in an RCT comparing 2 nonsteroidal antiinflammatory drugs. In such RCT, we are likely to have underestimated the sample size required. In addition, we categorized whether an RCT was adequately powered based on the recruited number of participants. However, in some RCT the study was underpowered at the conclusion of the trial due to large numbers of dropouts. This emphasizes the need to take into account the potential number of dropouts at study commencement and the importance of intention-to-treat analysis.

In analyzing these studies, we had to make assumptions that may have affected our results. In studies examining the efficacy of multiple drug dosages, we assumed the highest dose was the variable around which the trial was designed. Hence we used the effect size at the highest dose to calculate sample size, ignoring the effect size of lower doses. We accept that this is a limitation of our study; however, the complexity of determining power for dose-ranging studies was beyond the scope of our investigation. We also assumed a clinically meaningful difference to be 20%. Although this was an arbitrary choice for the purpose of calculating sample size for our analysis, the available composite response criteria for RA[10] and OA[11] use 20% as an indicator of improvement. Therefore, our study does highlight the lack of knowledge about clinically meaningful effect sizes in many outcome measures in rheumatology, and the difficulty this can present to calculating sample size.

Unfortunately, recognition of the presence of underpowered trials is not a new phenomenon. In an audit of RCT published in *Journal of the American Medical Association*, *Lancet*, and *New England Journal of Medicine* in 1975, 1980, 1985, and 1990, Moher, *et al*[12] found that only 16% of RCT had adequate power to detect a 25% relative difference and 36% power to detect a 50% relative difference. In addition, Dickinson, *et al*[13] reviewed RCT of the management of head injury published prior to 1998. Of 208 separate trials reported, they found no trial was large enough to detect a 5% absolute reduction in the risk of death or disability, and only 4% were large enough to detect a difference of 10%.

We have assumed that positive RCT are adequately powered. We acknowledge that positive RCT may not have been adequately powered *a priori*; rather, that they may have found a statistically significant effect by chance, because either the outcome effect was larger than clinically or logically expected, or because the sample population may have been skewed.

Overpowered RCT have also been criticized as unethical. If the sample size studied is larger than required to have an 80% chance of detecting a treatment effect, then more participants than need be are exposed to the potential risks of trial intervention or are missing out on the benefits of a proven intervention[3]. Of the positive trials we identified, a minority reported sample size calculations.

Problems also exist with respect to reporting of trial methodology and performance of statistical analysis. We found RCT often lacked adequate descriptions of randomization sequence generation, allocation concealment, and blinding. We hypothesized that negative underpowered RCT would be of lower quality than adequately powered negative RCT. However, we found no difference in reporting of trial methods between negative RCT with and those without adequate sample size, suggesting that there are not drastic differences in quality of trial methodology. In contrast, positive RCT appeared to have better reporting of randomization and allocation concealment than negative RCT, irrespective of the adequacy of their sample size. Use of appropriate analysis did not differ between the groups. In addition, some studies fail to make any statistical comparison between interventions. Previously, Hill, *et al*[14] found similar problems, with 89.9% and 79.3% of RCT not describing the method of randomization, and a further 87.4% and 80.2% failed to describe adequate allocation concealment in rheumatology RCT in 2 time periods, 1987-88 and 1997-98. In addition, there has been little change in the reporting of power calculations over the past 5 years. In 1997-98, 35.1% of negative RCT reported power calculations, compared to 42.9% in the present study. These shortcomings are not unique to rheumatology literature[12,13,15].

Trials with inadequate power, inadequate reporting of methods, or poorly designed protocols have limited scientific merit with respect to interpretation of the results they report. We found that the conduct of underpowered trials is not an infrequent occurrence in rheumatology. As few underpowered trials concerned rare diseases, or were designed with a view to compilation in a planned meta-analysis, the conduct of these trials is difficult to justify.

In conclusion, a significant number of RCT in rheumatic diseases are of limited clinical value, with concomitant ethical issues. Both investigators and institutional review boards have a responsibility to prospective study participants to determine the adequate sample size that will answer their research question, and to be realistic about their ability to recruit and retain enough participants in a given timeframe, prior to study commencement. In addition, journal editors have a responsibility to ensure that authors adequately address power issues in their reports so the wider rheumatology community can assess the likelihood of a type II error in an RCT with negative results.

ACKNOWLEDGMENT

We acknowledge the helpful advice of Associate Professor Michael La Valley, Biostatistician, Clinical Epidemiology Research and Training Unit, Boston University, Boston, USA.

REFERENCES

1. Halpern S, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA 2002;288:358-62.
2. Newell DJ. Type II errors and ethics. BMJ 1978;4:1789.
3. Altman DG. Statistics and ethics in medical research: III. How large a sample? BMJ 1980;281:1336-8.
4. Edwards SJL, Lilford RJ, Braunholtz D, Jackson J. Why "underpowered" trials are not necessarily unethical. Lancet 1997;350:804-7.
5. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomised controlled trials: is blinding necessary? Control Clin Trials 1996;17:1-12.
6. Donner A, Eliasziw M. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. Stat Med 1992;11:1511-9.
7. Rochon PA, Gurwitz JH, Simms RW, et al. A study of manufacturer-supported trials of nonsteroidal antiinflammatory drugs in the treatment of arthritis. Arch Intern Med 1994;154:157-63.
8. Bellamy N, Carette S, Ford PM, et al. Osteoarthritis antirheumatic drug trials. III. Setting the delta for clinical trials — results of consensus development (delphi) exercise. J Rheumatol 1992;19:451-7.
9. Chalmers TC, Lau J. Meta-analytic stimulus for changes in clinical trials. Stat Methods Med Res 1993;2:161-72.
10. Felson DT, Anderson JJ, Boers M, et al. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727-35.
11. Pham T, van der Heijde D, Lassere M, et al. Outcome variables for osteoarthritis clinical trials: The OMERACT-OARSI set of responder criteria. J Rheumatol 2003;30:1648-54.
12. Moher D, Dulberg C, Wells G. Statistical power, sample size and their reporting in randomised clinical trials. JAMA 1994;272:122-4.
13. Dickinson K, Bunn F, Wentz R, Edwards P, Roberts I. Size and quality of randomised controlled trials in head injury: review of published studies. BMJ 2000;320:1308-11.
14. Hill CL, La Valley M, Felson DT. Secular changes in the quality of published randomised clinical trials in rheumatology. Arthritis Rheum 2002;46:779-84.
15. Bhandari M, Richards RR, Sprague S, Schemitsch EH. The quality of reporting of randomized trials in the Journal of Bone and Joint Surgery from 1988 through 2000. J Bone Joint Surg Am 2002;84:388-96.