Acta Orthop. Belg., 2015, **81**, 115-122

ORIGINAL STUDY

Sample size calculations in orthopaedics randomised controlled trials : revisiting research practices

Sanjeeve SABHARWAL, Nirav K. PATEL, Ian HOLLOWAY, Thanos ATHANASIOU

From Department of Surgery and Cancer, Imperial College, London, UK

The purpose of this study was to identify how often sample size calculations were reported in recent orthopaedic randomized controlled trials (RCTs) and to determine what proportion of studies that failed to find a significant treatment effect were at risk of type II error.

A pre-defined computerized search was performed in MEDLINE to identify RCTs published in 2012 in the 20 highest ranked orthopaedic journals based on impact factor. Data from these studies was used to perform post hoc analysis to determine whether each study was sufficiently powered to detect a small (0.2), medium (0.5) and large (0.8) effect size as defined by Cohen. Sufficient power (1-β) was considered to be 80% and a two-tailed test was performed with an alpha value of 0.05.

120 RCTs were identified using our stated search protocol and just 73 studies (60.80%) described an appropriate sample size calculation. Examination of studies with negative primary outcome revealed that 68 (93.15%) were at risk of type II error for a small treatment effect and only 4 (5.48%) were at risk of type II error for a medium sized treatment effect.

Although comparison of the results with existing data from over 10 years ago infers improved practice in sample size calculations within orthopaedic surgery, there remains an ongoing need for improvement of practice. Orthopaedic researchers, as well as journal reviewers and editors have a responsibility to ensure that RCTs conform to standardized methodological guidelines and perform appropriate sample size calculations.

Keywords : Sample size ; power calculation ; research quality.

INTRODUCTION

Appropriately designed, conducted and reported randomized controlled trials (RCTs) represent the gold standard for evaluating healthcare interventions. Nevertheless, such studies can produce bias results if they lack methodological rigor (*19*). Inadequacies in the quality of RCTs resulted in the original Consolidated Standards of Reporting Trials (CONSORT) statement in 1996 (*1*), with its most recent revision published in 2010 (*19*). Within its guidance, the determination of sample size is a mandatory component for conducting an RCT, yet

■ Sanjeeve Sabharwal, MBBS MRCS MSc.
■ Nirav Patel, MBBS MRCS.
■ Ian Holloway, FRCS (T&O).
■ Thanos Athanasiou MD PhD FRCS FETCS.
■ *Department of Surgery and Cancer, Imperial College, London, UK.*

Correspondence : Mr Sanjeeve Sabharwal, Department of Surgery and Cancer, Imperial College, 10th Floor QEQM building, St Mary's Hospital, London W2 1NY, UK. E-mail : sanjeeve.sabharwal@ic.ac.uk

No benefits or funds were received in support of this study. The authors report no conflict of interests.

across all specialties in clinical medicine many RCTs still fail to report their sample size calculation or report them erroneously (*3*). Such failings were identified amongst orthopaedic RCTs in a study published in 2001 (*8*), however a more up to date examination of power and sample size calculations amongst orthopaedic RCTs has yet to be reported.

The purpose of a sample size calculation is to determine the number of patients required to detect a clinically significant treatment effect and to minimize the risk of type I and type II error occuring (*20*). A type I error occurs when a treatment effect has been found and such an effect does not actually exist (a false positive result). This probability (p) is denoted α and is typically set at 0.05. This suggests that there is a 5% chance of a significant effect occurring as a result of chance. A type II error occurs when no treatment effect is found, when in fact it such an effect does exist (a false negative result). The probability of type II error occurring is known as β. The probability of avoiding a type II error is derived by the equation $1-\beta$, and is known as the power of a study (*9*). Adequate power of a study has been defined at 80% ($\beta \leq 0.20$) (*4*). The parameters typically required for a priori sample size calculations are type I error (0.05), power (80%), assumptions in the control group including response rate along with standard deviation, and expected treatment effect or effect size (*13*). Effect size is often poorly reported (*9*) and post hoc or retrospective power calculations may be performed using proxy mathematical values such as those defined by Cohen (4). When a treatment effect is statistically significant ($p < 0.05$) the result infers sufficient sample size. However studies that fail to show a difference may suffer from type II error because the sample is not large enough for smaller treatments effects to be detected (*17*).

This study was performed to revisit deficiencies in sample size calculations amongst orthopaedic RCTs described over 10 years ago (*8*) and determine whether improvements have been made. Our two specific aims were to (1) identify how often appropriate sample size calculations were reported in recent orthopaedic RCTs and (2) to determine whether sample sizes used in studies with negative primary outcomes were sufficient.

MATERIALS AND METHODS

Information sources

A previous search protocol identified 3 orthopaedic journals that had high citation indexes and were viewed to be 'prestigious' to sample the RCTs used for their analysis (*8*). Selection of orthopaedic journals within this study was based on their impact factor as this remains the most widely accepted tool for benchmarking journals (*18*). The ISI Web of Knowledge journal citation report (JCR®) is widely used to provide information on a journal's bibliometric data (*7*). After selection of the orthopaedic journals within the JCR®, the top 20 journals with respect to impact factor were selected for sampling published RCTs (Table I).

Eligibility and Search Strategy

RCTs published in the 20 orthopaedic journals in the year 2012 were identified using a pre-defined computerized search on MEDLINE. Studies labeled "Randomized Controlled Trials" between January 2012 and December 2012 were selected in MEDLINE. This method of study selection has been found to capture approximately 92% of published RCTs (*14*). The authorship of our study deemed it to be an acceptable method to examine a large and representative sample of RCTs published in 2012 by the selected journals. The search terms used were "orthopaedic" and "orthopedic" and all search fields were included. Identified studies within the 20 specified journals were examined and included within the study after screening the articles to ensure they were clinical RCTs performed in humans. Animal and laboratory studies were excluded from the analysis.

Data extraction

The following information was extracted from each RCT : the study region, the orthopaedic sub-specialty, the journal impact factor, the sample size, whether an appropriate sample size calculation was described within the paper and the number of primary outcomes for each study. A sample size calculation was defined as a statistical test used to determine sample size which included a statement on the value of type I error, the power value, assumptions in the control group and the expected treatment effect. A primary outcome was defined as an outcome stated to be the main focus of the study or if this was not explicitly stated, then the outcome used in a

Table I. — The 20 journals identified by ISI Web of Knowledge journal citation report (JCR®) within their orthopaedic category and with selection based on their impact factor rank

Rank	Journal Title	Impact Factor (to 2 decimal places)
1	American Journal of Sports Medicine	4.40
2	Osteoarthritis and Cartilage	4.26
3	Spine Journal	3.36
4	Journal of Bone and Joint Surgery (American)	3.23
5	Arthroscopy	3.10
6	Journal of Orthopaedics and Sports Physical Therapy	2.95
7	Journal of Orthopaedic Research	2.88
8	Clinical Orthopaedics and Related Research	2.79
9	Physical Therapy	2.78
10	Acta Orthopaedica	2.74
11	The Bone and Joint Journal	2.69
12	Knee Surgery Sports Traumatology Arthroscopy	2.68
13	Journal of the American Academy of Orthopaedic Surgeons	2.46
14	International Orthopaedics	2.32
14	Journal of Elbow and Shoulder Surgery	2.32
16	Journal of Physiotherapy	2.26
17	Spine	2.16
18	European Spine Journal	2.13
19	Journal Of Arthroplasty	2.10
20	Knee	2.01

sample size calculation. When a significant difference related to treatment effect was identified this was labeled a positive outcome and if no significant difference was identified this was labeled a negative outcome. Furthermore studies were separated into two groups ; studies that reported positive outcomes for all their primary outcomes and studies that reported at least one negative outcome.

Statistical analysis

Retrospective analysis of sample size calculations used values of effect size as defined by Cohen *(4)*. This method has previously been used in studies examining sample size calculations in clinical trials *(8,17)*. Large, medium and small effect sizes correspond to a difference of 0.8, 0.5 and 0.2 respectively. G*Power 3.1 was used to perform the retrospective sample size analysis. This computer software is widely used as a power analysis program for social, behavioral and biomedical sciences *(6)*. The type of power analysis performed was a post hoc compute achieved power. A two-tailed t-test was performed with α set at 0.05 and the sample size of the selected study. For each RCT, the power (1-β) was determined for a small, medium and large effect size. A power of 80% was considered the lowest possible value for each effect size to protect against type II error. The number of studies that were significantly powered to detect small, medium and large effect size were identified and further examination of the data was performed to identify studies with negative outcomes and whether they were sufficiently powered to protect against type II error for the various effect sizes used. Descriptive statistics were performed using SPSS 20.0 (SPSS Inc, Chicago, IL).

Example of analysis

For the purpose of understanding the methodology and analysis performed in this study, an example of a calculation carried out on a fictitious study is presented.

A randomized controlled trial is performed to compare surgical and non-surgical treatment of acute achilles tendon rupture. The primary outcome was a validated functional outcome measure. Within the methodology the authors state that their power calculation is based on results from a previous study performed at their institution and for an 80 % power (or a power value of 0.8) they require a sample of 60. The size of their defined treatment effect is not stated. Their level of significance for statistical analysis comparing the two groups is set at $P < 0.05$. The study found no significant difference in functional outcome between the two treatment strategies.

A post hoc 2-tailed t-test is performed in G*Power 3.1 to evaluate the sample size calculation in this study. α is set at 0.05. For a large effect size (0.8), the power of this study is 1. For a medium effect size (0.5) the power of this study is 1 and for a small effect size (0.2) the power of this study is 0.34. This study appears well powered to detect large and medium treatment effects. However, in view of its results which stated there was no statistical difference between the two treatment strategies based on functional outcome measures, the results are vulnerable to Type II error (a false-negative) as there is not sufficient power to detect a small treatment effect. The sample size calculation software (G*Power 3.1) states that 188 is the smallest patient population that would have provided an 80% power for a small effect size. The conclusion of the study which states that there is no significant difference in functional outcomes between the treatment strategies is based on a sample size calculation at risk of Type II error and the authors have failed to state this within their limitations.

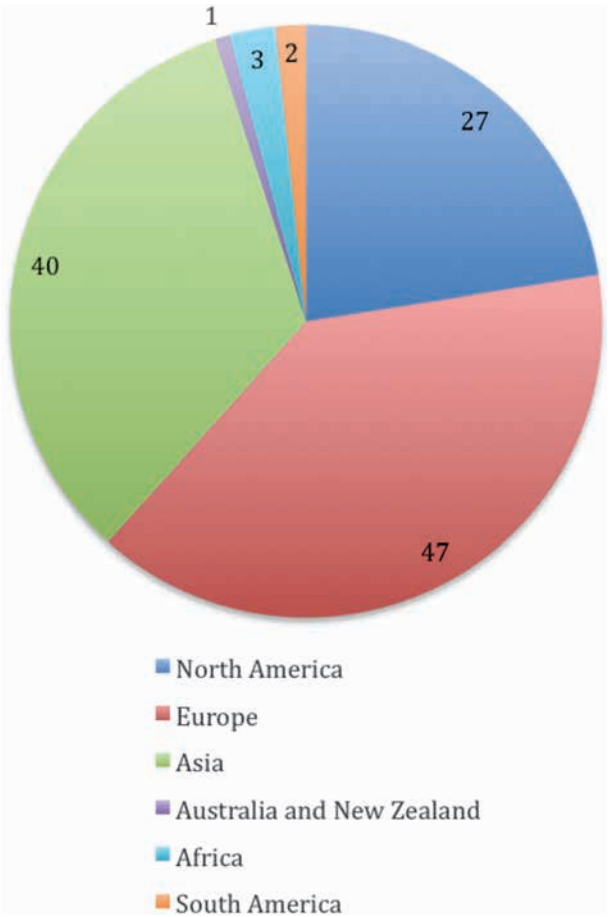

Fig. 1. — Origin of the RCTs based on region of development.

RESULTS

MEDLINE search and RCT characteristics

Application of the algorithm described within the 'Methods and materials' section identified 419 orthopaedic RCTs published in 2012. Within this group 142 were published within the 20 orthopaedic journals of interest, however 18 studies were available online in the year 2012 but appeared in print in 2013 and these were excluded. Furthermore, 4 studies represented laboratory or animal studies and were also excluded. In total 120 RCTs met the inclusion criteria and were subject to the analysis of this study. The studies identified were identical when the search was performed independently by two of the study's authors (SS and NKP).

The 19 RCTs identified within the Journal of Bone and Joint Surgery (American Volume) provided the largest proportion of RCTs from all 20 journals (15.80%).

The volume of studies from Europe surpassed those from Asia or North America (Fig. 1). There were 69 lower limb surgery (hips and knees) RCTs and this sub-specialty provided the vast majority of studies that were examined (Fig. 2). A Kolmogorov-Smirnov test indicated that neither journal impact factor nor the studies' sample size were not normally distributed. The median impact factor for the 20 journals examined was 2.69 (range 2.01-4.44) and the median sample size as 65.50 (range 10-598).

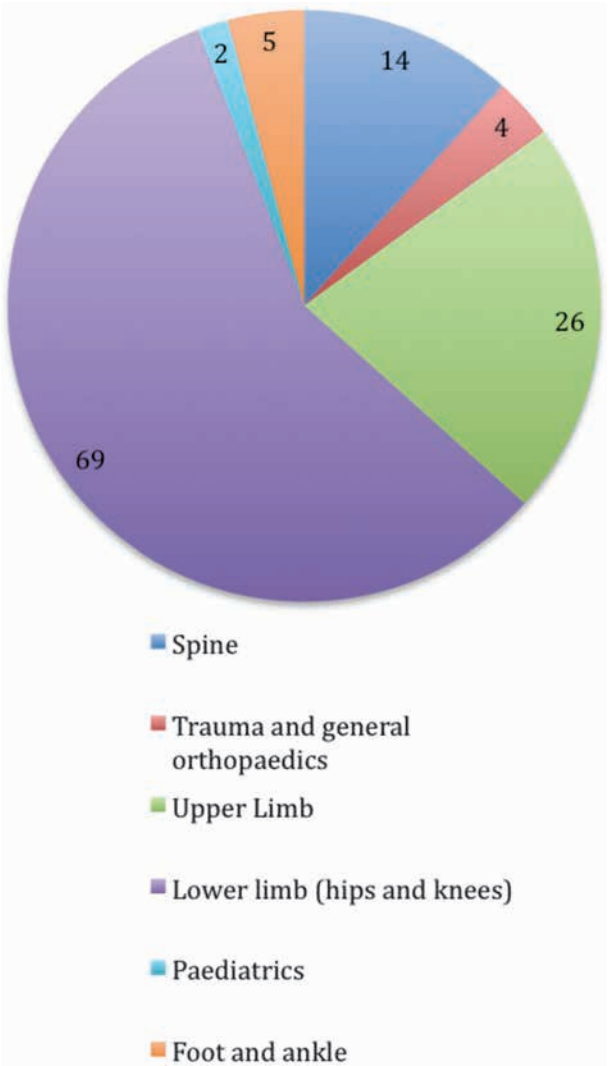

Fig. 2. — Focus of the RCTs based on their orthopaedic subspecialty.

Evaluation of the primary outcomes of the RCTs

Two authors independently examined all 120 RCTs to determine the number of primary outcomes based on the stated criteria. Inter-rater agreement on what constituted a primary outcome was strong with a Spearman's correlation of 0.922. Outcomes that the authors disagreed on (n = 4) were excluded from the analysis. In total 213 primary outcomes were identified. Amongst these outcomes there was no significant difference found in 98 (46%) and these were labeled 'negative' outcomes. A significant difference was found in 115 (54%) outcomes and these were labeled 'positive outcomes'. In 47 (39.20%) of the RCTs all the primary outcomes were 'positive' and in 73 (60.80%) at least one primary outcome was 'negative'.

Sample size calculations

Appropriate sample size calculations were reported in only 73 (60.80%) of the RCTs. Within this group of RCTs 25 (34.24%) had positive results for all their primary outcomes and 48 (65.76%) studies had at least one negative result.

In total, only 11 (9.2%) of studies had sufficient power to detect a small treatment effect. 111 (92.5%) of the studies were sufficiently powered to detect a medium treatment effect and all 120 (100%) of the RCTs had enough power to detect a small treatment effect. The approximate power values for all 120 RCTs for small, medium and large treatment effects are provided in Table II.

Sub group analysis of the RCTs that reported at least one negative outcome (n = 73) was performed to identify studies at risk of type II error. The approximate power values of these studies for small, medium and large treatment effects are provided in Table III. Only 5 (6.85%) studies were sufficiently powered to detect a small treatment effect. There were 69 (94.52%) studies that had enough power to detect a medium treatment effect and all the studies had enough to detect a large treatment effect. Therefore amongst the 73 RCTs that had negative primary outcomes, there were 68 (93.15%) at risk of type II error for a small treatment effect and 4 (5.48%) at risk of type II error for a medium treatment effect.

DISCUSSION

Principle findings and implications

In this study of 120 RCTs published during 2012 in the 20 highest ranking orthopaedic journals based on their impact factor, just 73 studies (60.80%) described a sample size calculation. Although all the RCTs had sufficient power to detect a large treatment effect, only 69 (94.52%) and 5 (6.85%) would

Table II. — The power values of all 120 RCTs required to detect a small, medium and large effect size

Power	Small	Medium	Large
> 0.8	11	111	120
0.60 ≤ x < 0.80	9	8	0
0.40 ≤ x < 0.60	39	0	0
0.20 ≤ x < 0.40	44	1	0
< 0.20	17	0	0

Table III. — Sub group analysis of the 73 studies with negative primary outcomes and their power values required to detect a small, medium and large effect size

Power	Small	Medium	Large
> 0.8	5	69	73
0.60 ≤ x < 0.80	7	3	0
0.40 ≤ x < 0.60	23	0	0
0.20 ≤ x < 0.40	31	1	0
< 0.20	7	0	0

have detected medium and small treatment effects respectively. Examination of studies with negative outcome revealed that 68 (93.15%) were at risk of type II error for a small treatment effect and only 4 (5.48%) were at risk of type II error for a medium sized treatment effect. All of these RCTs had sufficient power for a large effect size.

Although more than 15 years has elapsed since the first CONSORT statement recommended the reporting of appropriate sample size calculations in clinical trials (*1*), this study has shown that approximately a third of orthopaedic RCTs recently published in well recognized and high ranking orthopaedic journals fail to do so. Unfortunately, examination of the quality of orthopaedic research repeatedly identifies major shortcomings (*2,16*) despite the availability of guidance and methodological assessment tools (*12,19*). It has been suggested that this deficiency in research quality occurs because researchers, reviewers and editors fail to recognize the importance of sample size calculations (*3*). Researchers in orthopaedics must recognize the importance of sample size calculations in improving the methodological quality of their study. Furthermore, journal editors should ensure that all future published RCTs include this statistical analysis within the methods of a study. Encouragingly, the proportion of orthopaedic RCTs that perform these calculations has increased almost three fold since 2001 when Freedman et al. scrutinized sample size calculations in orthopaedics (*8*). It is possible that growing consensus within orthopedic research that all trials should be sufficiently powered for a given effect size (*8,17*) has contributed to this positive effect on research quality. Questions on the ethics of underpowered studies (*11*) may also be a driving force in improving reporting sample size calculations, however this study demonstrates that there is still room for improvement . Potential limitations to conducting sufficiently powered studies include the practical and financial feasibility of performing the large and multicentre studies that would be required (*15*). The argument for combining smaller sized studies to perform a meta-analysis with sufficient numbers to draw significant conclusions is countered by the belief that a single sufficiently powered randomized controlled trial is superior to a meta-analysis of numerous underpowered RCTs (*15*). Further controversy is drawn from some researchers who advocate that underpowered trials can be acceptable if the remainder of their methodology remains robust (*20*).

Despite the differences in opinion the statistical conclusion remains that RCTs that fail to determine

a significant effect of treatment are vulnerable to type II error if they are insufficiently powered for the appropriate effect size. Within this study this applied to 93.15% of such studies for a small effect size and 5.48% for a medium effect size using proxy values defined by Cohen (*4*). These results appeared to compare favorably to Freedman et al.'s results in 2001 (*8*) which demonstrated only 12.12% of such RCTs were sufficiently powered to detect a medium effect size, whilst none had sufficient power to detect a small effect size. Although there are some methodological differences between our studies, there is a vast difference in the reported results which infers that the risk of type II error may have improved 12 years after their recommendations were made (*8*).

Interestingly, only 1 study amongst the 120 that were examined underlined its own risk of type II error when considering small treatment effects, and this statement was made in an editorial note attached to the article (*23*). Although the authors of our study find this to be a commendable action by the journal's editorial team, we agree with other researchers in this field and believe that studies at risk of type II error should clearly report this as part of the limitations of their study (*14*). A key implication to be considered regarding undersized RCTs is that given that the volume of meta-analysis in orthopaedic surgery has surged over the last decade (*5*) and that there is concern that meta-analysis may be devalued by underpowered studies (*10*), there is a growing potential for this to occur within our field. Further evaluation of such an effect is probably warranted to determine whether this deficiency truly exists, because this is relevant to various areas in orthopaedic practice which are supported by clinical practice guidelines that may draw their recommendations from meta-analysis studies (*21*).

Study Limitations

There were 4 limitations to this study. Firstly, the use of post hoc analysis as a means of retrospectively determining sample size may be subject to criticism as it is commonly viewed to be an inferior method of sample size calculation compared to a priori power test (*22*). This is certainly true for prospective studies where the gold standard in methodology is to perform a sample size calculation prior to commencing the study (*19*), however post hoc analysis may be performed using values of effect size as those defined by Cohen (4) for retrospective analysis and is commonly adopted in studies such as ours that examine whether a sample of RCTs are sufficiently powered (*8,14,17*). Secondly our post hoc analysis used the assumption that all the primary outcomes of interest were normally distributed. It is likely that in some RCTs non-parametric statistics should have been employed however only 22 studies reported the distribution of their data and therefore in keeping with prior studies within this field we assumed normality of the data in all the studies when conducting our post hoc analysis (*8,14*). Thirdly, in determining what constitute an appropriate sample size calculation we did not scrutinize the assumptions made for each study. This was not done because it is not possible to determine whether assumptions have been manipulated to obtain a feasible sample size by just examining published data (*3*). So called "sample size samba" involves retrofitting assumption estimates to the available participants (*20*) and it is not possible to determine whether this has occurred without attending the planning meeting for the study (*3*). Finally, our selection of 20 orthopaedic journals based on their high rank with regards to impact factor may suggest that these journals are not representative of the entire body of available orthopaedic literature. Although this could be viewed as a limitation, the objective of this study was to examine RCTs from the most prestigious orthopaedic journals in order to strengthen the assumption that the research methodology and reporting from our study sample represented the highest quality in orthopaedic surgery.

CONCLUSIONS

Almost a third of RCTs published in the 2012 in the 20 highest ranking orthopaedic journals based on impact factor failed to describe an adequate sample size calculation. The vast majority of RCTs were susceptible to type II error for a small treatment effect however most were sufficiently powered for a medium sized treatment effect. Although

comparison of the results of this study with existing data from over 10 years ago infers improved practice in sample size calculations within orthopaedic surgery, there remains an ongoing need for improvement of practice. Orthopaedic researchers, as well as journal reviewers and editors have a responsibility to ensure that RCTs conform to standardized methodological guidelines and perform appropriate sample size calculations.

REFERENCES

1. **Begg C, Cho M, Eastwood S *et al.*** Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Jama* 1996 ; 276 : 637-639.
2. **Bhandari M, Richards RR, Sprague S *et al.*** 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-A : 388-396.
3. **Charles P, Giraudeau B, Dechartres A *et al.*** Reporting of sample size calculation in randomised controlled trials : review. *BMJ* 2009 ; 338 : b1732.
4. **Cohen J.** *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J. : L. Erlbaum Associates ; 1988.
5. **Dijkman BG, Abouali JA, Kooistra BW *et al.*** Twenty years of meta-analyses in orthopaedic surgery : has quality kept up with quantity ? *J Bone Joint Surg Am* 2010 ; 92 : 48-57.
6. **Faul F, Erdfelder E, Lang AG *et al.*** G*Power 3 : a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods* 2007 ; 39 : 175-191.
7. **Fitzpatrick RB.** ISI's Journal Citation Reports on the Web. *Med Ref Serv Q* 2003 ; 22 : 45-56.
8. **Freedman KB, Back S, Bernstein J.** Sample size and statistical power of randomised, controlled trials in orthopaedics. *J Bone Joint Surg Br* 2001 ; 83 : 397-402.
9. **Freedman KB, Bernstein J.** Sample size and statistical power in clinical orthopaedic research. *J Bone Joint Surg Am* 1999 ; 81 : 1454-1460.
10. **Guyatt GH, Mills EJ, Elbourne D.** In the era of systematic reviews, does the size of an individual trial still matter. *PLoS Med* 2008 ; 5 :e4.
11. **Halpern SD, Karlawish JH, Berlin JA.** The continuing unethical conduct of underpowered clinical trials. *Jama* 2002 ; 288 : 358-362.
12. **Jadad AR, Moore RA, Carroll D *et al.*** Assessing the quality of reports of randomized clinical trials : is blinding necessary ? *Control Clin Trials* 1996 ; 17 : 1-12.
13. **Machin D.** *Sample size tables for clinical studies*. Oxford : Wiley-Blackwell ; 2009.
14. **Maggard MA, O'Connell JB, Liu JH *et al.*** Sample size calculations in surgery : are they done correctly ? *Surgery* 2003 ; 134 : 275-279.
15. **Parker M.** Sample size and statistical power of randomised, controlled trials in orthopaedics. *J Bone Joint Surg Br* 2001 ; 83 : 1210.
16. **Parsons NR, Hiskens R, Price CL *et al.*** A systematic survey of the quality of research reporting in general orthopaedic journals. *J Bone Joint Surg Br* 2011 ; 93 : 1154-1159.
17. **Pike J, Leith J.** Type II error in the shoulder and elbow literature. *J Shoulder Elbow Surg* 2009 ; 18 : 44-51.
18. **Rieder S, Bruse CS, Michalski CW *et al.*** The impact factor ranking--a challenge for scientists and publishers. *Langenbecks Arch Surg* 2010 ; 395 Suppl 1 : 69-73.
19. **Schulz KF, Altman DG, Moher D *et al.*** CONSORT 2010 statement : updated guidelines for reporting parallel group randomised trials. *BMJ (Clinical research ed)* 2010 ; 340 : c332.
20. **Schulz KF, Grimes DA.** Sample size calculations in randomised trials : mandatory and mystical. *Lancet* 2005 ; 365 : 1348-1353.
21. **Shekelle PG, Woolf SH, Eccles M *et al.*** Clinical guidelines : developing guidelines. *BMJ* 1999 ; 318 : 593-596.
22. **Walters SJ.** Consultants' forum : should post hoc sample size calculations be done ? *Pharm Stat* 2009 ; 8 : 163-169.
23. **Xue H, Tu Y, Cai M.** Comparison of unilateral versus bilateral instrumented transforaminal lumbar interbody fusion in degenerative lumbar diseases. *Spine J* 2012 ; 12 : 209-215.